

Article

Seed and Oil Yield Prediction of Safflower (*Carthamus tinctorius* L.) Using UAV-Based Multispectral Imaging and Machine Learning Algorithms

İzzet Bozdemir ^{1,*} , Fatma Azizoglu ^{2,3,4} , Gokhan Azizoglu ^{2,3,5,*} , Aziz Şatana ⁶ , Ahmet Nusret Toprak ² 
and Ali Ünlükara ⁷ 

¹ Department of Field Crops, Graduate School of Natural and Applied Sciences, Erciyes University, Kayseri 38039, Türkiye

² Department of Computer Engineering, Faculty of Engineering, Erciyes University, Kayseri 38030, Türkiye; fatmaazizoglu@erciyes.edu.tr (F.A.); antoprak@erciyes.edu.tr (A.N.T.)

³ Graduate School of Natural and Applied Sciences, Computer Engineering, Erciyes University, Kayseri 38039, Türkiye

⁴ Department of Computer Engineering, Faculty of Engineering, Sivas Cumhuriyet University, Sivas 58140, Türkiye

⁵ Department of Software Engineering, Faculty of Engineering, Sivas Cumhuriyet University, Sivas 58140, Türkiye

⁶ Department of Field Crops, Faculty of Agriculture, Erciyes University, Kayseri 38030, Türkiye; azizsatana@erciyes.edu.tr

⁷ Department of Biosystems Engineering, Faculty of Agriculture, Erciyes University, Kayseri 38030, Türkiye; unlukara@gmail.com

* Correspondence: izzetbozdemir@gmail.com (İ.B.); gazizoglu@erciyes.edu.tr (G.A.)

Abstract

The objective of this study was to predict seed and oil yields in safflower (*Carthamus tinctorius* L.) using UAV-based multispectral imagery and machine learning algorithms. The study was conducted during the 2024 growing season under varying irrigation levels, fertilization practices, and applications of plant growth-promoting rhizobacteria. The experiment included four irrigation levels, fertilized and unfertilized conditions, and bacterial treatments consisting of *Bacillus pumilus*, *Bacillus albus*, their mixture, and a non-bacterial control. Sixty-two vegetation indices were calculated from multispectral images acquired during the harvest maturity period and used to predict seed and oil yields. To identify the most informative features, Mutual Information, Recursive Feature Elimination, and LASSO feature selection methods were applied; subsequently, Linear Regression, Decision Tree, Random Forest, Support Vector Regression, K-Nearest Neighbors, and XGBoost regression algorithms were compared. The results showed that Linear Regression combined with Mutual Information-based feature selection was the most successful approach for predicting both seed and oil yields. According to the 5-fold cross-validation results, average values of $R = 0.8600$, $MAE = 0.2870$, and $RMSE = 0.3765$ were obtained for seed yield prediction, while average values of $R = 0.8710$, $MAE = 0.0876$, and $RMSE = 0.1101$ were obtained for oil yield prediction.

Keywords: remote sensing; vegetation indices; feature selection; precision agriculture; seed yield estimation; oil yield estimation; mutual information; linear regression; plant growth-promoting rhizobacteria



Academic Editor: Jiangbo Li

Received: 26 May 2026

Revised: 19 July 2026

Accepted: 20 July 2026

Published: 22 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Global climate change poses a major threat to sustainable agriculture. Drought is a leading cause of yield loss, and, together with other abiotic stressors such as temperature extremes and salinity, directly limits plant growth and yield formation, resulting in significant agricultural losses [1,2]. With the global population projected to reach 9.7 billion by 2050, it is essential to restructure agricultural systems for higher productivity and more efficient resource use [3]. Understanding the effects of drought stress on plants and developing innovative management strategies are critical for global food security [4].

Water scarcity is one of the main factors limiting agricultural sustainability, and a significant portion of the world's arable land is exposed to varying degrees of drought [5]. Agriculture accounts for about 70% of global freshwater use, underscoring the strategic importance of water management for both agricultural and environmental sustainability. Nutrient management is also crucial: insufficient fertilization reduces yields, while excessive or unbalanced use harms the environment. Therefore, integrated water and nutrient management is essential for sustainable agriculture [6].

Plant growth-promoting rhizobacteria (PGPR) are gaining importance as alternatives or complements to chemical inputs in sustainable agriculture. *Bacillus* species enhance plant nutrition through nitrogen fixation and phosphorus solubilization, and increase tolerance to stress [7]. They also improve growth and yield by boosting nutrient uptake [8]. However, integrated assessments of these biological effects using remote sensing data are limited. In particular, correlating the effects of bacterial applications on plant pigments with seed and oil yields through spectral reflectance offers significant potential for sustainable and precision agriculture [9].

Safflower (*Carthamus tinctorius* L.) is an important oilseed crop grown in arid and semi-arid regions, known for its capacity to access water through a deep root system [10]. However, drought stress can significantly reduce seed yield, oil content, and physiological development. These challenges highlight the need for advanced analytical methods to better understand the factors affecting safflower yield and quality [11].

Crop yield represents a multidimensional trait shaped by the complex interactions among genotype, environment, and management practices. Traditional statistical methods frequently fail to adequately model these relationships, especially when analyzing nonlinear interactions [12]. Recently, remote sensing technologies and machine learning algorithms have become effective tools for capturing these complexities [13]. Multispectral imagery acquired via unmanned aerial vehicles (UAVs) enables rapid, non-destructive assessment of parameters such as plant health, chlorophyll content, biomass, and water status [13–15]. Vegetation indices (VI) from these images are widely used to monitor plant growth and stress. Combined with machine learning, this data provides high accuracy and generalizability for yield prediction [14].

However, few studies holistically evaluate the effects of rhizobacteria and fertilizer applications on safflower plants under drought stress using multispectral imaging and machine learning. There is also a notable lack of integrated approaches for simultaneously predicting seed and oil yields.

The primary objective of this study is to develop and evaluate the performance of a model for predicting seed and oil yields in safflower grown under different irrigation levels, fertilization regimes, and rhizobacterial treatments using UAV-based multispectral imaging and machine learning methods. To this end, VI derived from UAV multispectral imagery were integrated with machine learning algorithms to accurately predict seed and oil yields under varying cultivation practices and drought-stress conditions. By reflecting the combined effects of irrigation, fertilization, and rhizobacterial treatments through the spectral responses of the crop canopy, this approach is expected to support precision

and sustainable agriculture by enabling rapid, reliable, and cost-effective assessment of crop performance.

This study introduces a framework for estimating safflower seed and oil yield using machine learning and UAV-based multispectral imagery. The following are the primary contributions of this work:

- To the best of our knowledge, no previous study has specifically investigated the separate prediction of safflower seed yield and oil yield using an integrated framework that combines a large set of UAV-derived multispectral vegetation indices, systematic feature selection strategies, and multiple machine-learning regression algorithms.
- We evaluated sixty-two spectral VI derived from UAV-based multispectral images using three feature selection strategies to identify the most effective indices for predicting seed and oil yield in safflower.
- We propose a feature selection approach that integrates the Mutual Information (MI) algorithm with Linear Regression (LR) to identify the most informative index subsets for predicting safflower seed and oil yields.

The rest of this paper is organized as follows: Related works are presented in Section 2. Section 3 details the proposed method for predicting seed and oil yield in safflower plants. The experimental results and evaluation of the proposed method are presented in Section 4. Section 5 discusses the prediction results and the effects of VI. Finally, Section 6 concludes the paper.

2. Related Work

Remote sensing and machine learning techniques have been widely used in recent years for crop yield estimation, plant stress monitoring, and field phenotyping. In particular, UAV-based multispectral and hyperspectral imaging provides high-spatial- and temporal-resolution data, enabling more precise monitoring of crop development under field conditions [16,17]. These imaging systems also enable the extraction of VI, which are strongly correlated with canopy structure, chlorophyll content, biomass accumulation, plant health, and yield formation [18]. Therefore, VI derived from UAV imagery have been widely used as input variables in machine learning models for crop yield estimation [19].

Several studies have demonstrated that integration of UAV-based hyperspectral or multispectral data with feature selection methods can improve yield estimation performance by identifying the most informative spectral variables. For instance, Li et al. [20] used UAV-based hyperspectral imagery and ensemble machine learning models to predict winter wheat yield. In their study, feature selection methods such as recursive feature elimination (RFE), Boruta, and Pearson correlation coefficient (R) were applied to select the most relevant spectral indices and improve model performance. Similarly, Zhu et al. [21] employed UAV-based hyperspectral imaging and machine learning algorithms to predict early yield in oilseed rape, including successive projections algorithm, competitive adaptive reweighted sampling, uninformative variable elimination, and random forest, to identify effective wavelengths and develop accurate prediction models.

UAV-based multispectral imagery has also been widely used for yield estimation in different crops. Kumar et al. [22] used high-resolution UAV-based multispectral imagery and machine learning models to predict maize yield. Their study demonstrates that VI derived from multispectral images provides useful spectral information for accurate yield estimation. Bian et al. [14] estimated field-scale wheat yield using UAV-based multispectral data and machine learning models, while Lukas et al. [23] highlighted the usefulness of UAV-derived VI for oilseed rape yield prediction. More recently, Jiang et al. [24] integrated UAV multispectral imagery with meteorological data within a multimodal deep learning framework to predict cotton yield. Overall, these studies demonstrate that UAV-based

spectral data can effectively capture crop growth characteristics and that machine learning models can transform these spectral patterns into reliable yield estimations.

Despite these advancements, most existing studies have focused on major crops such as wheat, maize, oilseed rape/canola, soybean, and cotton. In contrast, safflower has received limited attention, particularly regarding UAV-based multispectral prediction of seed and oil yields. Furthermore, while feature selection has been shown to enhance prediction performance in hyperspectral studies, systematic evaluation of multiple feature selection strategies using UAV-derived multispectral VI remains limited for safflower. This gap is significant because safflower is an oilseed crop in which both seed and oil yields are critical agronomic and economic traits.

To address these gaps, this study focuses on predicting safflower seed and oil yields using UAV-based multispectral imagery, VI, feature selection methods, and machine learning regression models. A total of 62 VI derived from UAV multispectral imagery were evaluated using three distinct feature selection strategies to identify the most informative indices for accurate yield prediction. Additionally, a feature selection approach that integrates the MI algorithm with LR is proposed to identify the most informative subsets of VI. Based on our literature review, no previous study has specifically reported a UAV-based multispectral modeling framework that combines a large set of VI, systematic feature selection strategies, and machine-learning regression models for the separate prediction of safflower seed and oil yields. By focusing on safflower, systematically comparing VI and feature selection strategies, and developing a practical UAV-based prediction framework, this study contributes to precision agriculture and advances efficient yield estimation tools for oilseed crops.

3. The Proposed Seed and Oil Yield Prediction of Safflower

This section discusses the feature selection and machine learning methods used to estimate safflower seed and oil yields. Safflower plants were planted in the study area at the same time, as detailed in Section 3.1. UAV images of the plants were captured and processed to generate the dataset described in Section 3.3. To estimate seed and oil yields separately, the most informative spectral indices were identified from 62 candidates using the feature selection methods in Section 3.6. These selected indices served as input for the machine learning algorithms described in Section 3.8. A 5-fold nested cross-validation (nested CV) approach was used to ensure robust performance evaluation and to develop a more generalizable machine learning model.

3.1. Study Area, Image Acquisition, and Experimental Design

This study was conducted during the spring and summer of 2024 in Kayseri Province (38°43' N, 35°29' E), Central Anatolia, Türkiye. Fieldwork was carried out on plots at the Erciyes University Agricultural Research and Application Center (Erciyes Üniversitesi Tarımsal Araştırma ve Uygulama Merkezi, ERUTAM). The *Dinçer* safflower variety was used as the plant material and was sown on 26 April 2024.

Reference evapotranspiration (ET_0), which indicates atmospheric evaporative demand, was calculated using the meteorological parameters presented in Table 1. It reflects the combined influence of climatic factors, including temperature, solar radiation, wind speed, and relative humidity. Therefore, higher ET_0 values indicate greater potential crop water demand. During the experimental period from April to August 2024, monthly mean daily ET_0 values ranged from 3.2 to 5.5 mm per day⁻¹.

Table 1. Weather parameters and reference evapotranspiration ET_0 in 2024.

Month	$T_{\min}$ (°C)	$T_{\max}$ (°C)	RH (%)	U_2 (m s ⁻¹)	Precipitation (mm)	n (h month ⁻¹)	ET_0 (mm month ⁻¹)
April	4.2	17.3	68.9	1.71	32.4	294.0	96.0
May	7.1	22.1	66.1	1.69	34.7	303.8	124.0
June	11.2	26.4	44.5	1.53	22.6	357.0	159.0
July	13.3	30.1	49.9	1.73	8.1	347.2	170.5
August	13.0	29.8	42.8	1.53	4.3	325.5	155.0

Note: $T_{\min}$ and $T_{\max}$ denote the average daily minimum and maximum temperatures for each month, respectively; RH refers to the monthly mean relative humidity; U_2 represents wind speed measured at a height of 2 m; precipitation indicates monthly rainfall; n denotes the monthly total sunshine duration; and ET_0 is the monthly total reference evapotranspiration.

Before sowing, soil samples were collected from the experimental area at depths of 0.0–0.3 and 0.3–0.6 m to characterize the initial soil conditions. The soil had a loamy sand texture in both soil layers. The samples were analyzed to determine bulk density, lime content, organic matter, pH, electrical conductivity, volumetric water content at field capacity and permanent wilting point, and available nitrogen, phosphorus, and potassium contents. The initial physical and chemical properties of the experimental soil are presented in Table 2.

Table 2. Initial physical and chemical properties of the experimental soil.

Soil Depth (m)	Bulk Density (g cm ⁻³)	Lime (%)	Organic Matter (%)	pH	EC (dS m ⁻¹)	P_{v-fc} (%)	P_{v-wp} (%)	N (kg ha ⁻¹)	P ₂ O ₅ (kg ha ⁻¹)	K ₂ SO ₄ (kg ha ⁻¹)
0.0–0.3	1.40	2.60	1.22	7.49	1.18	27.8	13.0	21.8	20.8	272
0.3–0.6	1.38	4.50	0.89	7.55	1.11	28.5	14.1	10.9	11.7	375

Note: EC denotes the electrical conductivity of the soil saturation paste extract. P_{v-fc} and P_{v-wp} denote the volumetric water contents at field capacity and permanent wilting point, respectively. N denotes soil nitrogen content, P₂O₅ denotes soil phosphorus content, and K₂SO₄ denotes soil potassium content.

The experimental soil was slightly alkaline and had low organic matter and lime contents. The available volumetric water content, calculated as the difference between field capacity and permanent wilting point, was 14.8 and 14.4 percentage points for the 0.0–0.3 and 0.3–0.6 m soil layers, respectively. The soil had relatively high potassium content, whereas nitrogen and phosphorus contents were comparatively low. These initial soil conditions provide the basis for interpreting the responses of safflower plants to the fertilizer treatments applied during the experiment.

The experiment was conducted with four replicates in a split-split plot arrangement within a randomized complete block design. Four irrigation levels (0%, 35%, 65%, and 100%) were assigned to the main plots. Each main plot was subdivided into eight subplots to assess rhizobacterial applications under both fertilized and unfertilized conditions, as illustrated in Figure 1.

Each plot contained 300 plants arranged in six rows, each 5 m in length. The row spacing and intra-row spacing were 20 cm and 10 cm, respectively. The distance between subplots was 1 m, and the distance between blocks was 2.5 m. The total experimental area encompassed 1908.5 m². Figure 1 presents the experimental layout, based on the RGB orthomosaic generated from UAV images acquired with the DJI P4 Multispectral and processed in Pix4Dfields.

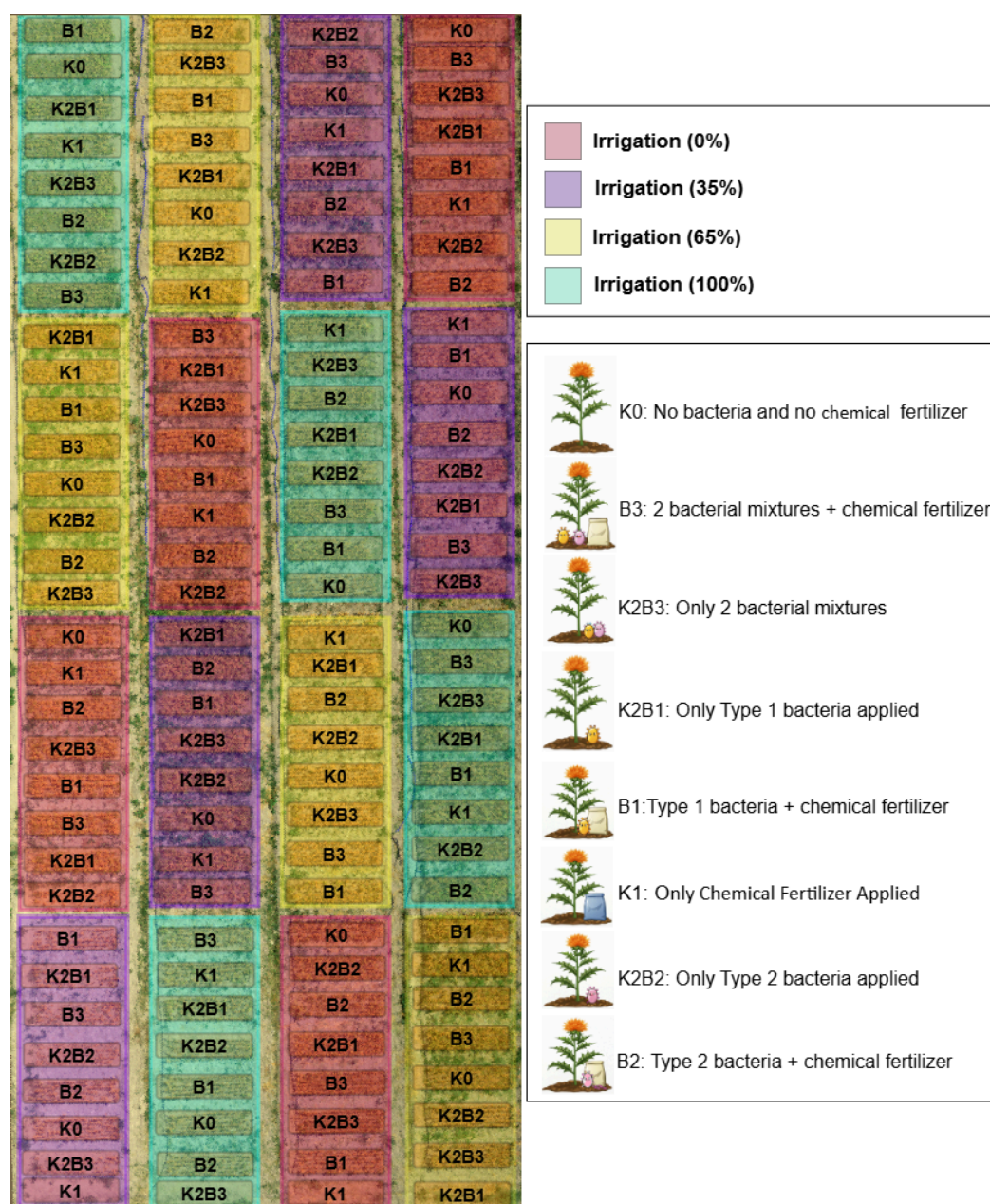


Figure 1. Experimental layout of safflower field plots cultivated in 2024, shown on the RGB orthomosaic generated from UAV imagery.

The effects of rhizobacterial treatments on safflower were evaluated under both fertilized and unfertilized conditions across four irrigation levels. The rhizobacterial treatments included a non-bacterial control (B0), *Bacillus pumilus* (B1), *Bacillus albus* (B2) and a combination of *Bacillus pumilus* and *Bacillus albus* (B3).

Bacterial suspensions were standardized to a concentration of 1×10^8 CFU mL⁻¹ before application. The isolation and identification of the bacteria used in this study are described in Section 3.2.

In the full irrigation treatment (100%), irrigation was applied before 55% of the total available water in the 0.60 m root zone depth of the plant had been depleted [25]. In the 35% and 65% irrigation treatments, 35% and 65% of the irrigation water applied to the full irrigation treatment were applied, respectively. In the rainfed treatment (0%), no irrigation was applied.

Chemical fertilizer was applied to the fertilized plots as urea and diammonium phosphate at rates of 120 kg N ha^{-1} and $100 \text{ kg P}_2\text{O}_5 \text{ ha}^{-1}$, respectively. No chemical fertilizer was applied to the unfertilized plots.

3.2. Bacterial Isolation and Identification

The bacteria used in this study were isolated from the roots of sugar beet plants cultivated in the study region (Kayseri). The isolation protocol was performed similarly to the method later described in detail by Ünlü et al. [26]. Within the scope of the isolation procedure, plant roots were washed with sterile Brucella Broth medium, and the obtained suspension was used as the stock suspension. Serial dilutions of 10^{-1} , 10^{-2} , 10^{-3} , and 10^{-4} were prepared from the stock suspension. Samples taken from the prepared dilutions were inoculated onto 7% sheep blood agar medium and incubated at 37°C for 24 h under aerobic, anaerobic, and microaerophilic conditions. Following incubation, the bacterial colonies were evaluated for morphological characteristics, and each colony was re-inoculated onto 7% sheep blood agar for purification and incubated again for 24 h.

Genomic DNA was extracted from pure bacterial colonies using a commercial kit (GeneJET Genomic DNA Purification Kit, Thermo Fisher Scientific, K0721, Waltham, MA, USA) according to the manufacturer's protocol. The 16S rRNA gene region of the obtained DNA samples was amplified by PCR based on the method reported by Edwards et al. [27]. PCR products were sent to a private company for DNA purification and Sanger sequencing analyses. Species identification was performed by comparing reference sequences in the NCBI GenBank database. The identified *Bacillus pumilus* (B_1) and *Bacillus albus* (B_2) strains were selected to evaluate their potential effects on safflower (*Carthamus tinctorius* L.) yield and quality under water-stress conditions.

3.3. Safflower Image Dataset

The safflower plant images used in this study were acquired at the harvest maturity stage using a DJI P4 Multispectral UAV (SZ DJI Technology Co., Ltd., Shenzhen, China), which simultaneously captures data in five spectral bands: blue, green, red, red-edge, and near-infrared (NIR). The UAV flight was conducted at the harvest maturity stage on 24 August 2024, between 09:00 and 10:00 to capture spectral responses near the final stages of seed and oil yield formation. The flight altitude was set to 5 m to obtain high-spatial-resolution multispectral images for plot-level analysis. Each experimental plot had a relatively small area of 6 m^2 . Therefore, the minimum operational flight altitude was preferred to better capture the spectral response of safflower plants and to reduce mixed-pixel effects from adjacent plots, soil background, and plot borders. With a fixed camera resolution of 1600×1300 pixels, a lower flight altitude reduces the ground area represented by each pixel and improves the spatial detail of the derived vegetation indices. In large-scale studies, higher flight altitudes may be suitable, provided sufficient spatial detail is preserved for plot-level analysis. The flight was controlled using DJI GS Pro software (version 2.0.18; SZ DJI Technology Co., Ltd., Shenzhen, China), and the main UAV flight and camera parameters are summarized in Table 3.

The UAV operated in automatic flight mode, capturing images from five spectral channels simultaneously at each imaging point. Figure 2 presents sample images from these five multispectral bands collected during the study.

Images acquired during the UAV flight were processed using Pix4Dfields (version 2.7.10; Pix4D SA, Prilly, Switzerland) to generate orthomosaic reflectance maps for five spectral bands (blue, green, red, red-edge, and NIR). During this process, the images were radiometrically calibrated and stitched using the photogrammetric processing workflow of Pix4Dfields. After the orthomosaic maps were generated, each experimental plot was man-

ually delineated on the reflectance maps. The delineated plot regions were then cropped from the orthomosaic maps and exported as individual plot-level image samples. The spatial resolution of the generated orthomosaic maps was approximately $0.263 \text{ cm pixel}^{-1}$, which allowed plot-level extraction of VI from the 6 m^2 experimental plots. Thus, each image sample corresponded to a specific experimental plot and was associated with the agronomic measurements collected from that plot.

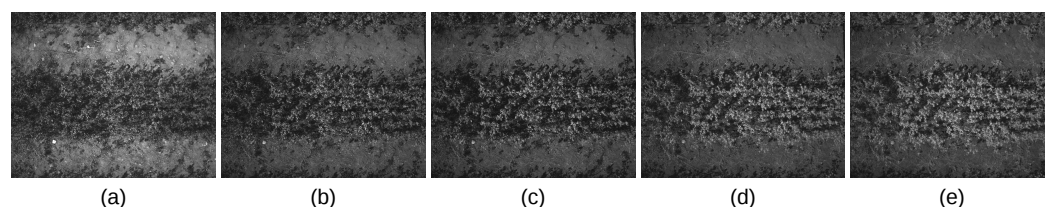


Figure 2. Representative multispectral images of safflower plants captured by a DJI P4 Multispectral UAV at 5 m altitude, showing the (a) blue, (b) green, (c) red, (d) red-edge, and (e) near-infrared bands.

Table 3. Flight parameters of the UAV and P4 multispectral camera.

Flight Parameter	Value
Time of day	09:00–10:00
Flight time	$\approx 11 \text{ min } 20 \text{ s}$
Flight length	$\approx 328.5 \text{ m}$
Flight speed	5 m s^{-1}
Flight altitude	5 m
Course angle	30°
Front overlap ratio	70%
Side overlap ratio	10%
Margin	0.0 m
Cover area	0.26 ha (2600 m^2)
Camera resolution	$1600 \times 1300 \text{ pixels}$
Image type	Multispectral
Image format	TIFF
Blue band central wavelength	$450 \pm 16 \text{ nm}$
Green band central wavelength	$560 \pm 16 \text{ nm}$
Red band central wavelength	$650 \pm 16 \text{ nm}$
Red-edge band central wavelength	$730 \pm 16 \text{ nm}$
Near-infrared band central wavelength	$840 \pm 26 \text{ nm}$

3.4. Calculation of Seed Yield and Oil Yield

For each multispectral image, the seed yield and oil yield values of the plot from which the image was cropped were linked. These values were used as target variables in the subsequent prediction models.

Seed yield was calculated using the harvested seed weight of each plot and the corresponding plot area. The plot seed weight was first measured in kilograms (kg), and then converted to seed yield in kg ha^{-1} according to Equation (1):

$$SY = \frac{W_p}{A_p} \times 10,000, \quad (1)$$

where SY is the seed yield (kg ha^{-1}), W_p is the plot seed weight (kg), and A_p is the plot area (6 m^2). The multiplication factor of 10,000 was used to convert the yield value from kg m^{-2} to kg ha^{-1} .

Oil content (%) was determined using a Soxhlet apparatus with petroleum ether extraction. Oil yield was then calculated by combining seed yield with the measured oil content value of the same plot, as given in Equation (2):

$$OY = SY \times \frac{OC}{100}, \quad (2)$$

where OY is the oil yield (kg ha^{-1}), SY is the seed yield (kg ha^{-1}), and OC is the oil content (%). Thus, the oil yield value assigned to each image represents the calculated oil yield for the corresponding plot.

3.5. Vegetation Indices

A total of 62 VI were calculated from multispectral images of safflower plants collected during the harvest season. The complete list of 62 vegetation indices, including their abbreviations, mathematical formulas, spectral bands, and references, is provided in Table S1 of the Supplementary Document. These indices were used to predict safflower seed and oil yields. The multispectral images used for index calculation were obtained from the experimental plots described in detail in Section 3.3.

The VI used in this study were primarily categorized into two main groups. The first group includes multispectral-based indices that incorporate NIR and red-edge bands in addition to the RGB bands [28]. The second group consists of RGB-based indices derived solely from the red, green, and blue bands [29,30].

3.6. Feature Selection Method

Feature selection is the process of identifying and selecting a subset of the most informative, relevant, and non-redundant features from a larger feature space in order to improve the performance, efficiency, and interpretability of a machine learning model. By retaining only the features that contribute meaningfully to prediction, this process helps reduce overfitting, lower computational cost, and improve predictive performance on new data [31].

We applied three feature selection methods (MI, LASSO, and RFE) to enhance the regression performance of the machine learning algorithm. We then compared the prediction accuracy of six algorithms for seed and oil yield using the selected features. Parameter settings and implementation details for the feature selection methods are listed in Table 4.

3.7. Machine Learning Based Regression Technique

Regression is a supervised learning method for quantitatively modeling the relationship between a dependent variable and one or more independent variables. In regression problems, the purpose is to predict continuous, numerical values of the dependent variable using the independent variables. In this study, seed and oil yields were estimated from multispectral plot images using VI as predictors in a regression model. In this context, MI [32], LASSO [33], and RFE feature selection methods [34] were used to determine the most defining VI. Following the feature selection process, machine learning-based regression models, including Linear Regression [35], Random Forest [36], Support Vector Regression [37], K-Nearest Neighbors [38], Decision Tree [39], and XGBoost [40], were used for prediction. The performance of the models was assessed using R , mean absolute error (MAE), and root mean square error (RMSE). The hyperparameters used to structure each machine learning algorithm are summarized in Table 5.

Table 4. Parameter settings of the feature selection methods used in this study.

Method	Parameter	Value
MI	Number of neighbors for estimation	3
	Discrete feature handling	Auto
	Optimal feature selection criterion	Highest R in inner CV
	Selection algorithm	SelectKBest
	Number of features (k)	Each value from 1 to 62 was evaluated in an independent run
RFE	Base estimator	Random Forest Regressor
	Split criterion	squared_error
	Maximum tree depth	None
	Minimum samples to split	2
	Number of estimators ($n_estimators$)	50
	Step size	1
	Number of selected features	Each value from 1 to 62 was evaluated in an independent run
	Optimal feature selection criterion	Highest R in inner CV
LASSO	Regularization parameter (α)	0.01
	Optimization method	Coordinate descent
	Tolerance for optimization	10^{-4}
	Fit intercept	True
	Maximum iterations	2000
	Selection threshold	$-np.inf$
	Optimal feature selection criterion	Highest R in inner CV
	Number of selected features	Each value from 1 to 62 was evaluated in an independent run

Table 5. Hyperparameters of the machine learning algorithms used in this study.

Algorithm	Hyperparameter	Value
Linear Regression	fit_intercept	True
Decision Tree	Splitting criterion	squared_error
	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
Random Forest	n_estimators	100
	max_depth	None
	Splitting criterion	squared_error
	min_samples_split	2
	min_samples_leaf	1
SVR	Kernel	Radial basis function (RBF)
	Regularization parameter (C)	1.0
	Epsilon (ϵ)	0.1
	Gamma (γ)	scale
KNN	n_neighbors	5
	Weights	uniform
	Distance metric	minkowski
XGBoost	n_estimators	100
	max_depth	6
	learning_rate	0.3
	Booster	gbtree

3.8. The Proposed Approach

In this study, seed and oil yields were estimated using UAV-derived multispectral vegetation indices with the proposed MI+LR framework. The irrigation levels, fertilization status, and rhizobacterial treatments were not directly included in the machine learning models as categorical input variables. Instead, these field treatments were used to generate agronomic and physiological variability among the experimental plots. Their effects on plant growth, canopy condition, pigment status, biomass accumulation, and yield formation were indirectly represented by the UAV-derived vegetation indices. Accordingly, the vegetation indices were used as predictor variables in the regression models, whereas seed yield and oil yield were treated as separate continuous target variables.

First, we calculated 62 spectral VI values for plots cropped from orthomosaic maps. We then applied min–max normalization, scaling each column so its maximum is 1 and its minimum is 0, with all other values adjusted linearly between them. The min–max normalization equation is as follows:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \quad (3)$$

where X is the original value of each vegetation index, $X_{\min}$ is the minimum value of the corresponding feature, $X_{\max}$ is the maximum value of the feature, and X' is the normalized value.

In this study, the MI feature selection method was applied in conjunction with LR as the base learning algorithm. MI enables selecting the most discriminative features by quantifying the information shared between each vegetation index and the target variable. In this context, the most informative variables were identified from among 62 VI, and different feature subsets were constructed. Seed and oil yields for the plots were estimated using an LR model based on subsets of traits defined separately for each trait. A schematic overview of the proposed methodology is presented in Figure 3.

These feature subsets were then used to predict seed yield and oil yield. During the modeling process, Nested Cross Validation was employed instead of the conventional train–test split. The Nested Cross Validation framework implemented in this study is presented in Figure 4. This approach performs model selection and performance evaluation within a two-stage framework. In the inner loop, the optimal feature selection and model combination were determined, whereas in the outer loop, the selected combination was tested on independent folds. Finally, the feature selection and model combination that yielded the best results in the outer loop were identified, and this combination was then applied to the dataset using the Single CV method. Thus, the model's generalization performance was assessed more robustly.

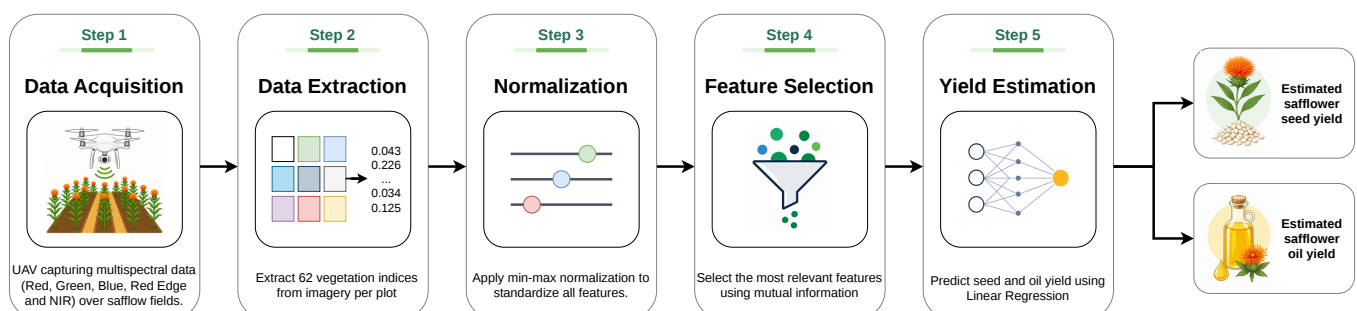


Figure 3. General workflow of the proposed model for safflower seed and oil yield prediction using UAV-based multispectral imagery, VI, feature selection, and machine learning.

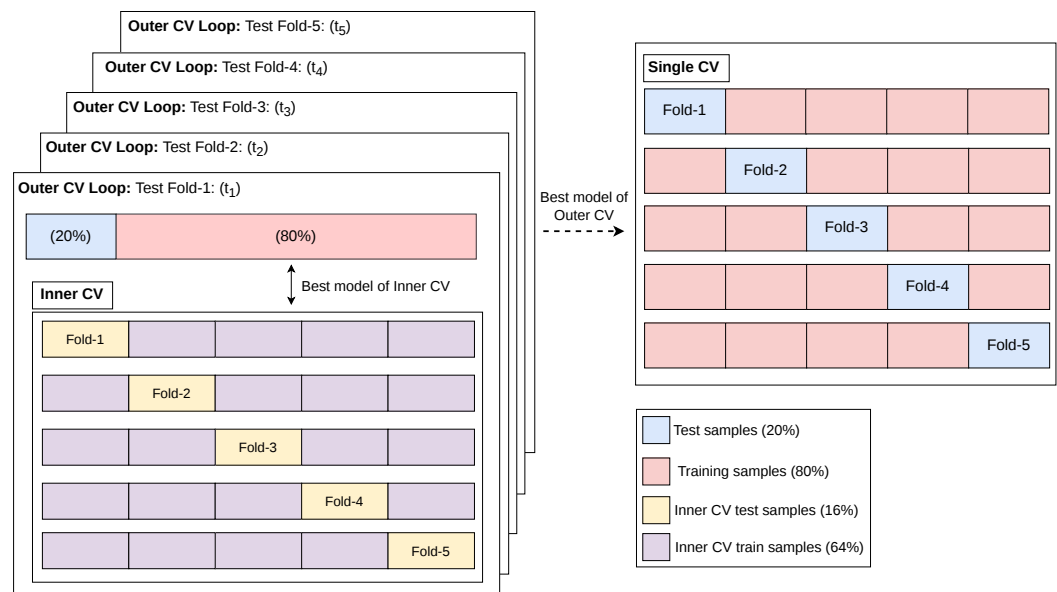


Figure 4. General block diagram of the Nested CV.

3.9. Evaluation Metrics

This study evaluates regression model performance using the R , MAE , and $RMSE$. These metrics quantify the agreement between measured and predicted safflower seed and oil yields.

The R is used to quantify the strength and direction of the linear relationship between measured and predicted values, as follows:

$$R = \frac{\sum_{i=1}^n (x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^n (x_i - m_x)^2 \sum_{i=1}^n (y_i - m_y)^2}} \quad (4)$$

MAE is used to evaluate the average magnitude of the prediction error and is calculated as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - y_i| \quad (5)$$

The $RMSE$ is calculated as the square root of the mean squared difference between measured and predicted values:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (6)$$

In Equations (4)–(6), x_i and y_i denote the measured and predicted target values, respectively; m_x and m_y denote the mean values of the measured and predicted targets, respectively; and n denotes the total number of samples. Higher R values and lower MAE and $RMSE$ values indicate better predictive performance.

3.10. Experimental Setup

Experiments on safflower seed and oil yield estimation were conducted on a computer system equipped with a 14th-generation Intel Core i9-14900K processor, 64 GB of RAM, and a NVIDIA GeForce RTX 4090 GPU with 24 GB of memory.

4. Experimental Results

This section presents detailed experimental results on the performance of proposed methods for estimating seed and oil yields in safflower plants subjected to different irri-

gation, fertilization, and bacterial treatment regimens. A comparative evaluation of six machine learning regression models, each combined with three feature selection strategies, was conducted to estimate seed and oil yields from multispectral images of the safflower plant. Specifically, the MI, RFE, and LASSO feature selection methods were tested in combination with LR, Decision Tree, Random Forest, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and XGBoost. The performance of the models was assessed using three standard regression metrics: the R , MAE , and $RMSE$.

4.1. Use of AI-Assisted Tools

ChatGPT 5.5 was used to generate safflower-related icon elements for the block diagrams shown in Figures 1 and 3. The tool was not involved in study design, data collection, data analysis, statistical evaluation, model development, interpretation of results, or the generation of scientific conclusions. All AI-assisted outputs were reviewed, edited, and approved by the authors.

4.2. Measured Seed and Oil Yield Responses to Agronomic Treatments

Before performing the machine learning analysis, the measured seed and oil yield responses were examined to characterize the agronomic variability generated by the experimental treatments. This step is important because the prediction models were developed using UAV-derived VI rather than the treatment labels themselves. Therefore, the irrigation, fertilization, and rhizobacterial treatments provided the biological and agronomic variability underlying the target variables, whereas the VI were used as numerical predictors of this variability.

The mean seed yield values under 0%, 35%, 65%, and 100% irrigation were 1.59, 2.45, 2.55, and 2.69 t ha⁻¹, respectively. Similarly, the corresponding mean oil yield values were 0.44, 0.69, 0.71, and 0.79 t ha⁻¹, respectively. These descriptive results show that both seed and oil yields generally increased with increasing irrigation level, whereas lower yield values were observed under reduced irrigation, particularly in the rainfed treatment.

As shown in Figure 5, seed yield also varied within each irrigation level depending on fertilization status and rhizobacterial treatment. The largest descriptive differences were observed among the irrigation levels, while fertilization and rhizobacterial applications further modified the yield patterns within each irrigation condition.

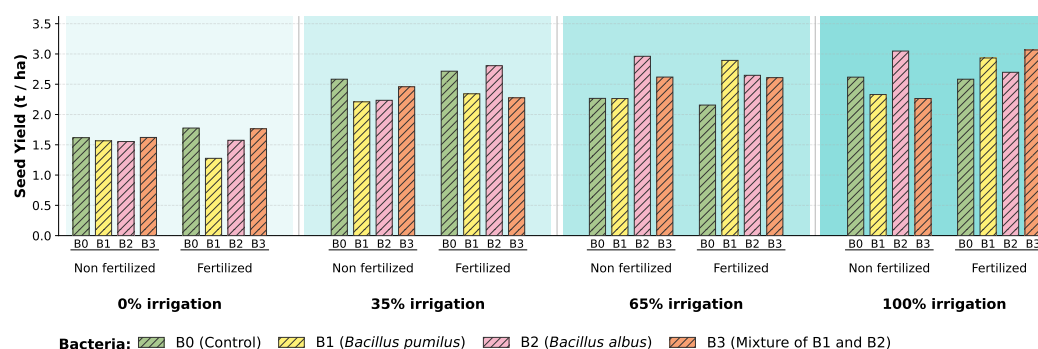


Figure 5. Measured seed yield responses of safflower under different irrigation levels, fertilization conditions, and rhizobacterial treatments.

A similar pattern was observed for oil yield (Figure 6), where the differences among fertilization and rhizobacterial treatments were also reflected within each irrigation level. This treatment-induced variability was important for the machine learning analysis because it created a broad range of seed and oil yield values to be predicted from UAV-derived VI. In this study, the treatment labels themselves were not used as direct categorical inputs

in the regression models. Instead, their effects on crop condition, canopy development, pigment status, biomass accumulation, and yield formation were indirectly represented by the spectral responses captured through UAV-derived VI. Therefore, the following machine learning analyses focused on modeling the relationship between UAV-derived VI and the measured seed and oil yield values.

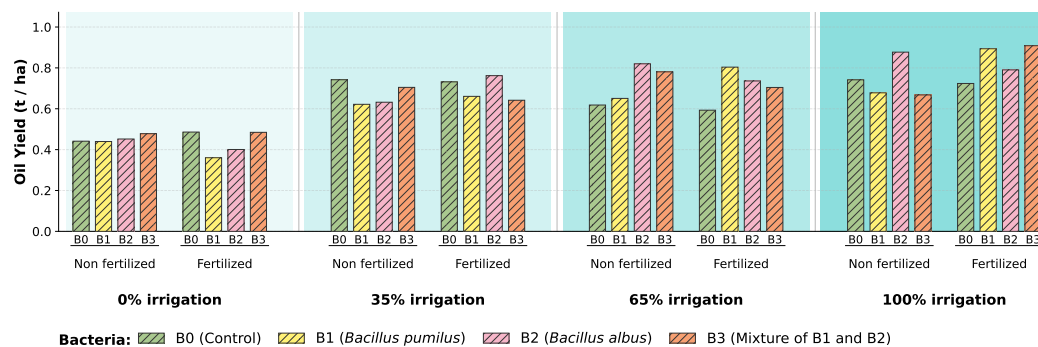


Figure 6. Measured oil yield responses of safflower under different irrigation levels, fertilization conditions, and rhizobacterial treatments.

Although the treatment labels were not used as predictor variables, they were retained as visual annotations in Figures 7 and 8 to show how the measured and predicted yield values were distributed across irrigation levels, rhizobacterial treatments, and fertilization conditions.

4.3. Comparison of Feature Selection and Regression Methods in Predicting Safflower Seed Yields

This section presents the results of a comparative analysis of feature selection methods and machine learning regression algorithms for estimating seed yield from VI in safflower plants. Table 6 presents the experimental results for safflower seed yield prediction, including the R , MAE, and RMSE values obtained across the five outer CV layers. The feature selection methods and regressors listed were selected from the model components that achieved the highest performance in each outer layer's inner CV process.

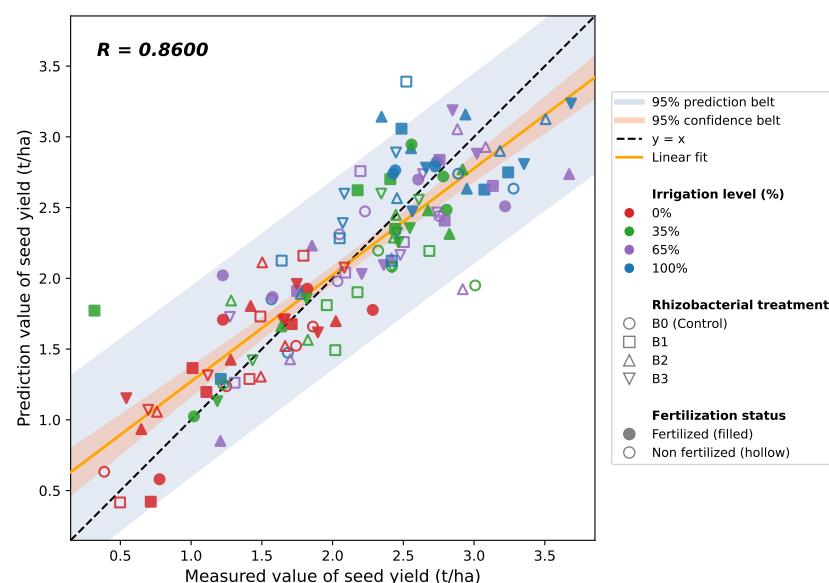


Figure 7. Prediction performance of the MI+LR model for safflower seed yield under 5-fold Single CV. Colors represent irrigation levels, marker shapes represent rhizobacterial treatments, and filled and hollow markers represent fertilized and non-fertilized plots, respectively.

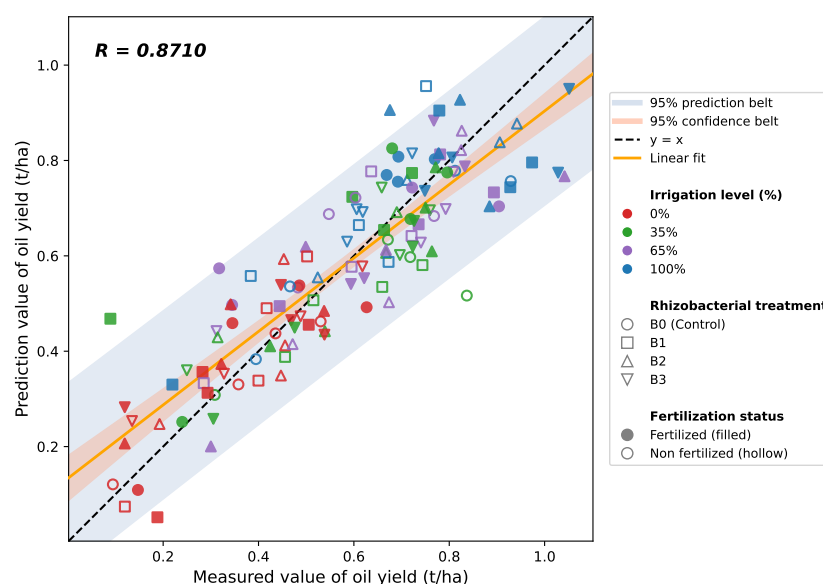


Figure 8. Prediction performance of the MI+LR model for safflower oil yield under 5-fold Single CV. Colors represent irrigation levels, marker shapes represent rhizobacterial treatments, and filled and hollow markers represent fertilized and non-fertilized plots, respectively.

The results in Table 6 show that the model achieved high performance in predicting safflower seed yield across all outer CV folds. The R coefficient, ranging from 0.7971 to 0.9082, indicates that the model can establish a strong relationship between observed and predicted values. Furthermore, the MI+LR combination yields the best results in 4 out of 5 folds, demonstrating that this approach offers a more successful framework than other methods. The highest performance was achieved in the fifth fold.

Table 6. Best outer CV results for seed yield estimation. The bold row indicates the fold that achieved the highest predictive performance.

Fold	Feature Selection	Regressor	R	MAE ($t\ ha^{-1}$)	$RMSE$ ($t\ ha^{-1}$)	Selected VI
1	MI	LR	0.8038	0.3722	0.4619	12
2	RFE	LR	0.7971	0.3412	0.4166	19
3	MI	LR	0.8123	0.3264	0.4589	16
4	MI	LR	0.8660	0.2930	0.3737	11
5	MI	LR	0.9082	0.2654	0.3145	17

Figure 9 shows the feature indices determined by the MI-based feature selection method, sorted by importance, for the 5th fold, where the highest performance was achieved in the Outer CV process.

Based on the MI+LR model, which achieved the highest prediction accuracy in the outer CV process, a 5-fold single cross-validation procedure was performed on the dataset. In this evaluation, the model was trained and tested independently in each fold to assess its predictive consistency. The corresponding results for safflower seed yield estimation are presented in Table 7. As shown in Table 7, the MI+LR model maintained strong predictive performance in the Single CV setting, achieving average R , MAE , and $RMSE$ values of 0.8600, 0.2870, and 0.3765, respectively. The agreement between measured and predicted seed yield values is further illustrated in Figure 7.

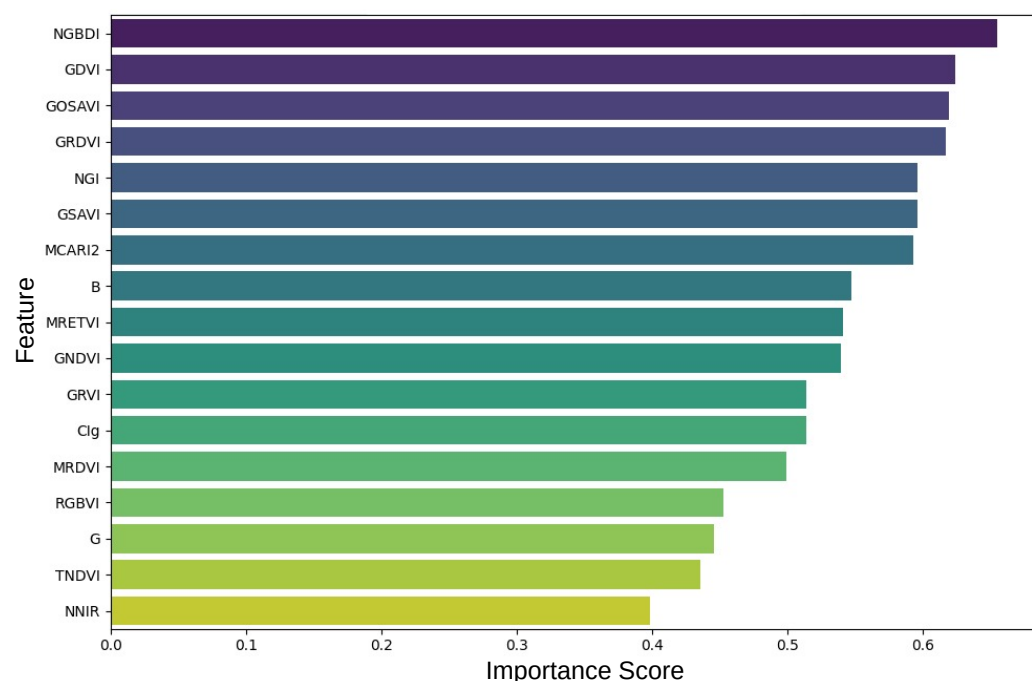


Figure 9. The 17 most important VI for seed yield prediction, determined by MI feature selection in the best outer CV fold (Fold 5).

Table 7. Single CV results for seed yield estimation. The bold row indicates the mean values of the metrics across all folds.

Fold	R	MAE (t ha ^{−1})	RMSE (t ha ^{−1})
1	0.8783	0.2772	0.3614
2	0.8388	0.2928	0.3783
3	0.8067	0.3306	0.4627
4	0.8682	0.2688	0.3654
5	0.9082	0.2654	0.3145
Mean	0.8600	0.2870	0.3765

Figure 7 presents the relationship between measured and predicted safflower seed yield values obtained using the MI+LR model under the 5-fold Single CV scheme. The predicted values were generally distributed close to the reference line, indicating that the model successfully captured the main variation in seed yield across the experimental plots. The fitted regression line showed a similar trend, suggesting that the selected UAV-derived VI captured important spectral information related to seed yield. This result is also supported by the mean cross-validation performance values of $R = 0.8600$, $MAE = 0.2870$, and $RMSE = 0.3765$. The confidence band around the regression line indicates the stability of the fitted trend, while the prediction band reflects the expected uncertainty for individual plot-level predictions. The treatment-based encoding showed that irrigation level primarily structured the observed seed-yield range. Rainfed plots were concentrated mainly at lower measured yields, whereas plots under 65% and 100% irrigation occurred more frequently at higher yield levels. Filled and hollow markers, as well as the different bacterial marker shapes, were distributed on both sides of the 1:1 line, indicating no visually apparent treatment-specific systematic bias associated with fertilization or rhizobacterial treatment. The fitted line was slightly above the 1:1 line at the lower end of the measured range and fell below it at higher measured yields. This pattern suggests a mild regression-to-the-mean effect, with limited overestimation at the lower end and a more evident

underestimation of high seed yields. A few larger residuals, including some low-yield fertilized rainfed plots, appeared to represent isolated plot-level deviations rather than a coherent treatment-specific error pattern. Overall, these findings show that MI-based feature selection, combined with LR, provides an effective and interpretable approach for estimating safflower seed yield.

4.4. Comparison of Feature Selection and Regression Methods in Predicting Safflower Oil Yields

This section presents the results of a comparative analysis of feature selection methods and machine learning regression algorithms for estimating oil yield from VI in safflower plants. Table 8 presents the experimental results for safflower oil yield prediction, including the *R*, *MAE*, and *RMSE* values obtained across the five outer CV layers. The feature selection methods and regressors listed were selected from the model components that achieved the highest performance in each outer layer's inner CV process.

Table 8. Best outer CV results for oil yield estimation. The bold row indicates the fold that achieved the highest predictive performance.

Fold	Feature Selection	Regressor	<i>R</i>	<i>MAE</i> (t ha ^{−1})	<i>RMSE</i> (t ha ^{−1})	Selected VI
1	MI	LR	0.8607	0.0975	0.1194	11
2	RFE	LR	0.8059	0.0947	0.1221	6
3	MI	LR	0.8923	0.0784	0.1104	8
4	MI	LR	0.8805	0.0796	0.1038	8
5	MI	LR	0.8788	0.0901	0.1089	15

Figure 10 shows the feature indices determined by the MI-based feature selection method, sorted by importance, for the 3rd fold, where the highest overall performance was achieved in the Outer CV process.

The results presented in Table 8 indicate that the model achieved consistently high performance in estimating safflower oil yield across all Outer CV folds. The *R* coefficient, ranging from 0.8059 to 0.8923, indicates strong agreement between the observed and predicted oil yields. Furthermore, the MI+LR combination demonstrated consistently high performance across 4 of the 5 folds, suggesting that it offers a more reliable and effective framework than the alternative methods evaluated in the study. Although the lowest *RMSE* was obtained in the fourth fold, the third fold was considered the best overall, as it performed better on two of the three evaluation metrics, namely *R* and *MAE*. Accordingly, a Single CV was performed using the MI+LR combination identified in the third fold, with the eight selected VIs. The Single CV results for safflower oil yield estimation are presented in Table 9.

Table 9. Single CV results for oil yield estimation. The bold row indicates the mean values of the metrics across all folds.

Fold	<i>R</i>	<i>MAE</i> (t ha ^{−1})	<i>RMSE</i> (t ha ^{−1})
1	0.8610	0.0976	0.1198
2	0.8313	0.0934	0.1112
3	0.8923	0.0784	0.1104
4	0.8887	0.0785	0.1016
5	0.8815	0.0901	0.1077
Mean	0.8710	0.0876	0.1101

The results presented in Table 9 demonstrate that the MI+LR model, applied with the eight selected VI, achieved stable and high predictive performance for safflower oil yield estimation across all Single CV folds. The *R* values ranged from 0.8313 to 0.8923, indicating

a strong relationship between the observed and predicted values. In addition, the *MAE* and *RMSE* values remained within a relatively narrow range, suggesting that the model produced consistent prediction errors across the folds. Based on the mean performance values, the model yielded an overall *R* of 0.8710, an *MAE* of 0.0876, and an *RMSE* of 0.1101, confirming the reliability of the proposed approach for oil yield estimation. This agreement between measured and predicted oil yield values is further illustrated in Figure 8.

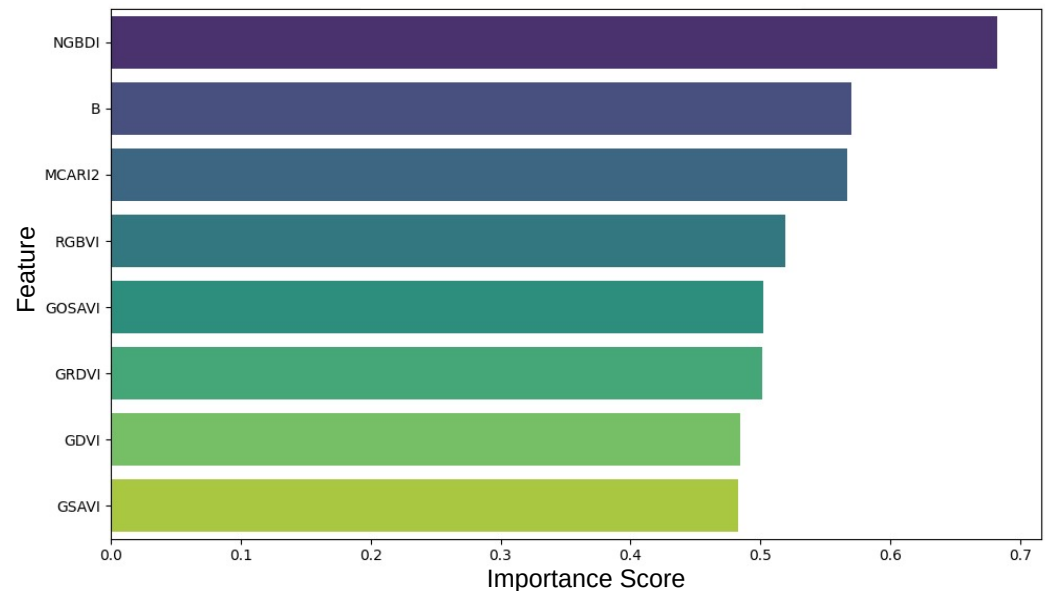


Figure 10. Feature importance ranking of the 8 selected VI for oil yield prediction, determined by MI feature selection in the best outer CV fold (Fold 3).

Figure 8 illustrates the predictive performance of the MI+LR model for safflower oil yield estimation under the 5-fold Single CV scheme. The scatter plot shows that most predicted values align with the measured oil yield pattern, indicating that the model captured the major changes in oil yield among the experimental plots. The fitted regression line exhibited a clear positive trend and remained close to the ideal prediction trend across the evaluated range. This visual result is consistent with the model's quantitative performance, which yielded mean values of $R = 0.8710$, $MAE = 0.0876$, and $RMSE = 0.1101$. The relatively narrow confidence interval suggests that the overall regression trend was stable, whereas the wider prediction interval reflects the natural uncertainty of individual plot-level estimations. The treatment-based visual assessment showed a pattern similar to that observed for seed yield. Rainfed plots were concentrated mainly in the lower oil-yield range, whereas plots under 65% and 100% irrigation occurred more frequently at higher measured values, although overlap remained among the irrigation levels. Fertilized and non-fertilized plots, together with the different rhizobacterial treatments, were distributed on both sides of the 1:1 line. Thus, no visually apparent treatment-specific systematic bias was associated with fertilization status or rhizobacterial treatment. The fitted line remained above the 1:1 line at lower measured oil yields and below it at higher values, indicating a mild tendency to overestimate low oil yields and underestimate high oil yields. The largest residuals were confined to a small number of individual plots and did not form a consistent treatment-specific cluster. Considering that oil yield is influenced by both seed yield and oil content, these results indicate that the selected UAV-derived VI provided sufficient spectral information to estimate this yield-related quality trait.

5. Discussion

The potential of UAV-based multispectral imagery and machine learning algorithms for estimating seed and oil yields in safflower was evaluated under varying irrigation, fertilization, and PGPR treatment conditions. The experimental design created a wide range of agronomic conditions, including different water regimes, bacterial applications, and both fertilized and unfertilized plots. Consequently, VI derived from multispectral images reflected yield-related spectral variation resulting from diverse field management practices. In this context, 62 VI were assessed using three feature selection methods, namely MI, RFE, and LASSO, in combination with six regression algorithms: LR, Decision Tree, Random Forest, SVR, KNN, and XGBoost.

The results indicate that the MI+LR combination provided the most consistent and effective performance for both seed and oil yield estimation. For seed yield prediction, the best-performing model combinations selected in the outer CV folds achieved R values between 0.7971 and 0.9082. MI+LR was selected as the best combination in four out of five outer CV folds, demonstrating that Mutual Information-based feature selection effectively identified informative VI for seed yield estimation. In the subsequent 5-fold Single CV evaluation, the MI+LR model achieved an average R of 0.8600, MAE of 0.2870, and $RMSE$ of 0.3765. These results indicate that selected VI derived from UAV-based multispectral imagery can successfully explain yield-related spectral variation in safflower.

A similar result was observed for oil yield prediction. In the outer CV results, R values ranged from 0.8059 to 0.8923, and MI+LR was again selected as the best-performing model in four out of five folds. The highest overall outer CV performance was achieved using eight VI selected by MI, resulting in $R = 0.8923$, $MAE = 0.0784$, and $RMSE = 0.1104$. In the Single CV evaluation, the MI+LR model achieved an average R of 0.8710, MAE of 0.0876, and $RMSE$ of 0.1101. Given that oil yield is derived from seed yield and oil content, these results indicate that multispectral VI can represent not only canopy vigor and biomass-related characteristics but also spectral patterns associated with final oil yield.

Irrigation and fertilization are among the most influential agronomic practices affecting crop productivity because they regulate plant water status, nutrient availability, photosynthetic activity, biomass accumulation, and ultimately seed and oil yield. Previous studies have shown that changes in irrigation and nitrogen management significantly influence canopy development and crop performance, which are reflected in vegetation indices derived from UAV multispectral imagery. For example, Peng et al. [41] demonstrated that nitrogen fertilization altered leaf area index (LAI) and yield in oilseed rape, while UAV-derived red-edge vegetation indices accurately captured these changes and enabled reliable yield estimation. Similarly, Jiang et al. [42] reported that reduced nitrogen application affected vegetation indices such as NDVI, RVI, and PSRI, reflecting differences in plant growth and grain yield. Moreover, Zhu et al. [43] showed that crop water status under different irrigation regimes could be effectively characterized using UAV multispectral imagery. Although these studies were conducted on different crops, they consistently demonstrate that water and nutrient management can induce physiological and structural changes that are reflected in canopy spectral characteristics.

In the present study, fertilization did not produce a consistent increase in final seed and oil yields across all treatment combinations. However, descriptive differences were observed among the plots in several growth-related characteristics, including dry biomass, branch number, capitulum number, capitulum weight, and stem diameter, as well as in seed oil content, which was evaluated as a quality parameter. These observations suggest that fertilization may have influenced plant development, canopy architecture, and seed quality, even when its effect on final yield was not consistent. These observations

suggest that fertilization may have influenced plant development and canopy architecture even when its effect on final yield was not consistent. Such physiological and structural variation may be reflected in canopy spectral characteristics and, consequently, in the vegetation indices extracted from UAV-based multispectral imagery. Therefore, although fertilization was not directly included as an input variable in the machine learning models, some of its effects on crop condition may have been indirectly represented through the spectral predictors. Because the imagery was acquired at harvest maturity, it represented the integrated canopy condition resulting from the applied irrigation, fertilization, and rhizobacterial treatments up to that stage. However, the single-date imaging design did not allow direct evaluation of treatment-specific growth dynamics throughout the growing season. A detailed temporal and treatment-specific analysis of these dynamics is therefore beyond the scope of the present prediction-focused study and should be addressed in future research. Consistent with the cited literature, the present results suggest that spectral variability associated with irrigation, fertilization, and rhizobacterial treatments was captured by UAV-derived vegetation indices, enabling the machine learning models to estimate safflower seed and oil yields without explicitly including these agronomic factors as input variables.

A notable finding of this study is that LR, when combined with MI-based feature selection, demonstrated superior and more consistent performance compared to more complex algorithms. Although Random Forest, Decision Tree, SVR, KNN, and XGBoost were also evaluated, their performance did not exceed that of the MI+LR approach. This indicates that the relevance of the selected input variables was more decisive than model complexity. In datasets, many indices are derived from similar spectral bands and may overlap. Therefore, the use of feature selection is important for reducing redundancy, improving interpretability, and increasing model stability.

The effectiveness of feature selection observed in this study is consistent with previous research. Li et al. [20] demonstrated that feature selection methods such as RFE, Boruta, and Pearson correlation-based selection enhance yield prediction using UAV-based hyperspectral imagery. Similarly, Zhu et al. [21] used feature selection methods to identify effective spectral variables for predicting oilseed rape yield. Li et al. [44] also emphasized that selecting informative features can improve model performance compared with using all available variables. Consistent with these findings, the present results indicate that MI-based feature selection can identify informative subsets of VI for safflower yield estimation.

The successful prediction of oil yield is another important result of this study. While most UAV-based yield estimation studies focus mainly on grain or seed yield, oil yield is a key economic trait for oilseed crops. Dilmurat et al. [45] demonstrated that remote sensing and machine learning can estimate seed protein and oil-related quality traits. The findings of the present study support this by showing that multispectral VI can also estimate safflower oil yield. This is especially important because safflower is a major oilseed crop in arid and semi-arid regions, yet it has not been widely studied in UAV-based yield-prediction research.

Building upon these findings, another important aspect of the present study is the inclusion of various combinations of irrigation, fertilization, and PGPR treatments within the same experimental framework. These treatments influence plant vigor, pigment status, canopy development, photosynthetic activity, and ultimately seed and oil yield. Previous research has demonstrated that the physiological changes induced by agronomic management practices and plant stress responses can be effectively monitored through canopy spectral reflectance. For instance, de Souza et al. [9] reported that alterations in plant pigments and spectral responses associated with PGPR can be quantified using remote sensing, while Kthiri et al. [46] showed that the effects of biological agents can be mon-

itored via plant spectral characteristics. Therefore, the present study extends previous research by demonstrating that UAV-based multispectral imaging, combined with machine learning, can successfully estimate safflower seed and oil yields under integrated irrigation, fertilization, and PGPR management conditions.

Overall, the results show that UAV-derived VI, when integrated with suitable feature selection and regression modeling, offers a practical, interpretable framework for estimating safflower seed and oil yields. The consistent performance of MI+LR across both target variables supports the use of the selected multispectral VI as reliable predictors in precision agriculture.

5.1. Comparison with the Current Literature

In recent years, remote sensing and machine learning have been widely used for crop yield estimation. Bian et al. [14] used UAV-based multispectral data and machine learning models to estimate wheat yield at the field scale. Kumar et al. [22] demonstrated the effectiveness of high-resolution UAV-based multispectral imagery for predicting maize yield. Similarly, Lukas et al. [23] reported that UAV-derived VI are valuable for estimating oilseed rape yield. Jiang et al. [24] indicated that integrating UAV multispectral imagery with additional data sources enhances cotton yield prediction. These studies indicate that UAV-based spectral data can effectively represent crop growth and yield-related characteristics.

The significance of the feature selection observed in this study is consistent with that observed in previous studies. Li et al. [44] demonstrated that feature selection enhances prediction performance by reducing redundant spectral information and isolating the most informative variables. Similarly, Zhu et al. [21] found that selecting effective wavelengths or spectral features improves yield prediction performance in oilseed crops. In the present study, MI-based feature selection enhanced both predictive performance and interpretability by identifying compact subsets of VI for seed and oil yield estimation.

This study contributes to the limited literature on safflower yield prediction. Most UAV-based yield estimation studies have focused on major field crops such as wheat, maize, cotton, soybean, and oilseed rape. In comparison, safflower has received less attention, especially regarding seed and oil yield prediction using UAV-based multispectral imagery and machine learning. Since both seed and oil yields are economically significant traits in safflower production, the ability to estimate these traits from spectral data represents a valuable advancement in precision agriculture research for oilseed crops.

A notable aspect of this study is the inclusion of various combinations of irrigation, fertilization, and PGPR treatment. Previous studies have shown that biological applications, nutrient management, and water availability significantly affect plant physiological status and spectral reflectance. By evaluating safflower under integrated agronomic conditions, this study establishes a comprehensive field-based framework for yield prediction. Therefore, the proposed approach may facilitate more accurate monitoring of crop performance under diverse management conditions.

5.2. The Limitations and Advantages of Work

The most important advantage of this study is that it presents an integrated approach that combines multispectral imagery, VI, feature selection methods, and machine learning algorithms. In particular, the systematic evaluation of high-dimensional vegetation index datasets and the comparison of different feature selection methods provide a significant methodological contribution to the modeling process. This approach offers a valuable contribution to the literature by developing data-representation-based modeling strategies.

Another key advantage is the high interpretability of the model's performance. The fact that simpler model structures can yield effective results when combined with appropriate

feature selection reduces the computational cost of the model and makes the results easier to interpret. This is a significant advantage, particularly for field applications and decision support systems.

One of the study's strengths is the joint assessment of biotic and abiotic factors. The integrated consideration of irrigation, fertilization, and PGPR applications contributes to a more realistic modeling of plant growth and yield formation. Such integrated approaches enable more accurate, practical results in precision agriculture applications.

Several limitations are present in this study. The MI+LR model demonstrated high predictive performance within the 2024 experimental dataset, and the cross-validation results showed that the model maintained stable accuracy across folds. Although the study was conducted during a single growing season and at one experimental site, the experimental design included different irrigation levels, fertilization conditions, and PGPR treatments, thereby generating a wide range of agronomic variability among the plots. This diversity supports the evaluation of the model under contrasting field management conditions within the scope of the present experiment. Nevertheless, environmental factors such as seasonal climatic variation, soil properties, local growing conditions, cultivar responses, and crop management practices may influence the relationship between UAV-derived VI and safflower seed and oil yields. Therefore, future validation using multi-year and multi-location datasets would further strengthen the assessment of the generalizability and practical robustness of the proposed MI+LR framework across broader production environments.

The UAV survey was conducted at the harvest maturity stage on 24 August 2024. This stage was selected because spectral responses at harvest maturity are expected to reflect the cumulative effects of irrigation, fertilization, and rhizobacterial treatments on canopy condition, biomass accumulation, senescence, and final yield formation. However, using a single survey date limits the ability to evaluate how prediction performance may vary across phenological stages. The choice of survey date can influence yield estimation because VI represent different physiological and structural crop characteristics at different growth stages. Therefore, future studies using multi-temporal UAV imagery across key phenological stages would be useful for determining the most informative imaging period and improving the robustness of yield prediction models.

Finally, the fact that model performance is highly dependent on data representation suggests it may vary across datasets and under different environmental conditions. Therefore, it is important to test the model's generalizability in future studies using larger datasets, different locations, and time-series data.

6. Conclusions

This study evaluated the prediction of safflower seed and oil yields using UAV-based multispectral imagery, VI, feature selection methods, and machine learning algorithms under different irrigation (0%, 35%, 65%, and 100%), fertilization, and PGPR treatment conditions. Therefore, the dataset encompassed a wide range of water, nutrient, and biological management conditions that affect safflower yield formation.

The results indicated that UAV-derived VI effectively captured yield-related spectral variation resulting from the treatment combinations. These findings suggest that the effects of varying irrigation levels, fertilizer applications, and PGPR treatments on safflower seed and oil yields were reflected in the multispectral data and could be accurately modeled using machine learning. Therefore, this study demonstrates that multispectral imaging is a viable tool for monitoring safflower yield responses to integrated water, fertilizer, and bacterial management strategies.

Among the evaluated methods, the LR model with MI-based feature selection demonstrated the most consistent performance for both target variables. The MI+LR model

achieved an average R of 0.8600, MAE of 0.2870, and $RMSE$ of 0.3765 for seed yield estimation, as well as an average R of 0.8710, MAE of 0.0876, and $RMSE$ of 0.1101 for oil yield estimation. The results suggest that VI offers reliable predictions of safflower seed and oil yield across diverse agricultural management conditions.

Another important finding is that reliable prediction performance was achieved using single-date multispectral imagery acquired at harvest maturity. This result indicates that UAV-based multispectral imaging can provide a rapid, non-destructive, and cost-effective tool for evaluating safflower productivity and estimating seed and oil yields at the plot scale.

In conclusion, this study demonstrates that UAV-derived VI and machine learning can be used to estimate safflower seed and oil yields under varying irrigation levels, fertilization practices, and PGPR applications. Future research will involve multi-season, multi-location, and cultivar-specific validation of the proposed framework. Additional treatment-specific statistical analyses will be performed to determine the irrigation, fertilizer, and PGPR combinations that yield the most favorable seed and oil production outcomes.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/agriculture16141566/s1>, Table S1. Complete list of the 62 vegetation indices, including their abbreviations, mathematical formulas, spectral bands, and references (PDF).

Author Contributions: Conceptualization, İ.B., F.A., G.A., A.Ş., A.N.T. and A.Ü.; methodology, İ.B., F.A., G.A., A.Ş., A.N.T. and A.Ü.; software, F.A. and G.A.; validation, İ.B., F.A., G.A., A.Ş., A.N.T. and A.Ü.; formal analysis, İ.B.; investigation, İ.B., F.A. and G.A.; data curation, İ.B., F.A., G.A., A.N.T. and A.Ü.; writing—original draft preparation, İ.B., F.A. and G.A.; writing—review and editing, A.Ş., A.N.T. and A.Ü.; visualization, F.A. and G.A.; supervision, A.Ş., A.N.T. and A.Ü. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) 2211-A National PhD Scholarship Program, and Erciyes University Scientific Research Projects Coordination Office under Project Nos. FDK-2023-13105 and FDK-2025-14764.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request. The source code for data processing and model implementation is publicly available at https://github.com/azizoglu/safflower_yield_prediction (accessed on 19 July 2026).

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT 5.5 by OpenAI to assist in generating safflower-related icon elements used in the general block diagrams presented in Figures 1 and 3. The authors reviewed and edited the output and take full responsibility for the content of this publication. Grammarly was used only for English language editing and readability improvement.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

PGPR	Plant Growth-Promoting Rhizobacteria
UAV	Unmanned Aerial Vehicle
VI	Vegetation Index/Indices
ERUTAM	Erciyes University Agricultural Research and Application Center
RFE	Recursive Feature Elimination

MI	Mutual Information
CV	Cross-Validation
R	Pearson Correlation Coefficient
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
NIR	Near-infrared

References

1. Tardieu, F.; Simonneau, T.; Muller, B. The physiological basis of drought tolerance in crop plants: A scenario-dependent probabilistic approach. *Annu. Rev. Plant Biol.* **2018**, *69*, 733–759. [\[CrossRef\]](#)
2. Azizoğlu, F.; Azizoğlu, G.; Toprak, A.; Sağlam, C. Predicting phenological stages with deep neural network models to increase efficiency in tomato production. In Proceedings of the 2024 32nd Signal Processing and Communications Applications Conference (SIU), Antalya, Türkiye, 15–18 May 2024; pp. 1–4. [\[CrossRef\]](#)
3. Wheeler, T.; von Braun, J. Climate change impacts on global food security. *Science* **2013**, *341*, 508–513. [\[CrossRef\]](#)
4. Zhou, S.; Williams, A.P.; Lintner, B.R.; Berg, A.M.; Zhang, Y.; Keenan, T.F.; Cook, B.I.; Hagemann, S.; Seneviratne, S.I.; Gentile, P. Soil moisture-atmosphere feedbacks mitigate declining water availability in drylands. *Nat. Clim. Change* **2021**, *11*, 38–44. [\[CrossRef\]](#)
5. Du, T.; Kang, S.; Zhang, J.; Davies, W.J. Deficit irrigation and sustainable water-resource strategies in agriculture for China's food security. *J. Exp. Bot.* **2015**, *66*, 2253–2269. [\[CrossRef\]](#)
6. Roberts, D.P.; Mattoo, A.K. Sustainable agriculture-enhancing environmental benefits, food nutritional quality and building crop resilience to abiotic and biotic stresses. *Agriculture* **2018**, *8*, 8. [\[CrossRef\]](#)
7. Orhan, E.; Esitken, A.; Ercisli, S.; Turan, M.; Sahin, F. Effects of plant growth promoting rhizobacteria (PGPR) on yield, growth and nutrient contents in organically growing raspberry. *Sci. Hortic.* **2006**, *111*, 38–43. [\[CrossRef\]](#)
8. Er, H.; Kuşlu, Y. Effect of integrated use of recycled wastewater and plant growth-promoting rhizobacteria (PGPR) on the quality characteristics of safflower (*Carthamus tinctorius* L.) under deficit irrigation in semi-arid conditions. *J. Agric. Sci.* **2025**, *163*, 625–636. [\[CrossRef\]](#)
9. de Souza, A.E.S.; Barbosa Júnior, M.R.; de Almeida Moreira, B.R.; da Silva, R.P.; Lemos, L.B. UAV multispectral data: A reliable approach for managing phosphate-solubilizing bacteria in common bean. *Agronomy* **2022**, *12*, 2284. [\[CrossRef\]](#)
10. Singh, S.; Boote, K.J.; Angadi, S.V.; Grover, K.K. Estimating water balance, evapotranspiration and water use efficiency of spring safflower using the CROPGRO model. *Agric. Water Manag.* **2017**, *185*, 137–144. [\[CrossRef\]](#)
11. Amini, H.; Arzani, A.; Bahrami, F. Seed yield and some physiological traits of safflower as affected by water deficit stress. *Int. J. Plant Prod.* **2013**, *7*, 597–614. [\[CrossRef\]](#)
12. Khairunniza-Bejo, S.; Mustaffa, S.; Wan Ismail, W.I. Application of artificial neural network in predicting crop yield: A review. *J. Food Sci. Eng.* **2014**, *4*, 1–9.
13. Islam, M.R.; Abdullah, H.M.; Rahman, M.F.; Islam, M.; Tuhin, A.K.; Ashiquzzaman, M.; Islam, K.S.; Geisseler, D. Mitigation of water-deficit stress in soybean by seaweed extract: The integrated approaches of UAV-based remote sensing and a field trial. *Drones* **2025**, *9*, 487. [\[CrossRef\]](#)
14. Bian, C.; Shi, H.; Wu, S.; Zhang, K.; Wei, M.; Zhao, Y.; Sun, Y.; Zhuang, H.; Zhang, X.; Chen, S. Prediction of field-scale wheat yield using machine learning method and multi-spectral UAV data. *Remote Sens.* **2022**, *14*, 1474. [\[CrossRef\]](#)
15. Li, F.; Guo, Y.; Ma, Y.; Lv, N.; Gao, Z.; Wang, G.; Zhang, Z.; Shi, L.; Zhao, C. Accurate estimation of spring maize aboveground biomass in arid regions based on integrated UAV remote sensing feature selection. *Agronomy* **2026**, *16*, 219. [\[CrossRef\]](#)
16. Barzin, R.; Pathak, R.; Lotfi, H.; Varco, J.; Bora, G.C. Use of UAS multispectral imagery at different physiological stages for yield prediction and input resource optimization in corn. *Remote Sens.* **2020**, *12*, 2392. [\[CrossRef\]](#)
17. Azizoglu, F.; Toprak, A.; Sağlam, C.; Yetişir, H.; Ünlükara, A.; Sekerci, A.; Hancı, F.; Çetin, N. Hail damage assessment in tomato crops using UAV-based vegetation indices and machine learning-based regression. *Comput. Electron. Agric.* **2026**, *250*, 111883. [\[CrossRef\]](#)
18. Kešelj, K.; Stamenković, Z.; Kostić, M.; Aćin, V.; Ivezić, A.; Ivanišević, M.; Magazin, N. Temporal dynamics of UAV multispectral vegetation indices for accurate machine learning-based wheat yield prediction. *AgriEngineering* **2026**, *8*, 71. [\[CrossRef\]](#)
19. Shafiee, S.; Lied, L.M.; Burud, I.; Dieseth, J.A.; Alsheikh, M.; Lillemo, M. Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. *Comput. Electron. Agric.* **2021**, *183*, 106036. [\[CrossRef\]](#)

20. Li, Z.; Chen, Z.; Cheng, Q.; Duan, F.; Sui, R.; Huang, X.; Xu, H. UAV-based hyperspectral and ensemble machine learning for predicting yield in winter wheat. *Agronomy* **2022**, *12*, 202. [\[CrossRef\]](#)
21. Zhu, H.; Lin, C.; Dong, Z.; Xu, J.L.; He, Y. Early yield prediction of oilseed rape using UAV-based hyperspectral imaging combined with machine learning algorithms. *Agriculture* **2025**, *15*, 1100. [\[CrossRef\]](#)
22. Kumar, C.; Mubvumba, P.; Huang, Y.; Dhillon, J.; Reddy, K. Multi-stage corn yield prediction using high-resolution UAV multispectral data and machine learning models. *Agronomy* **2023**, *13*, 1277. [\[CrossRef\]](#)
23. Lukas, V.; Huňady, I.; Kintl, A.; Mezera, J.; Hammerschmiedt, T.; Sobotková, J.; Brtnický, M.; Elbl, J. Using UAV to identify the optimal vegetation index for yield prediction of oil seed rape (*Brassica napus* L.) at the flowering stage. *Remote Sens.* **2022**, *14*, 4953. [\[CrossRef\]](#)
24. Jiang, C.; Guo, X.; Li, Y.; Lai, N.; Peng, L.; Geng, Q. Multimodal deep learning models in precision agriculture: Cotton yield prediction based on unmanned aerial vehicle imagery and meteorological data. *Agronomy* **2025**, *15*, 1217. [\[CrossRef\]](#)
25. Allen, R.G.; Pereira, L.S.; Raes, D.; Smith, M. *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*; FAO Irrigation and Drainage Paper 56; FAO: Rome, Italy, 1998.
26. Ünlü, E.; Yilmaz, S.; Yetişir, H.; Karim, A.A.; Gün, B.; Idris, A.B. Characterization of multi-trait plant growth-promoting rhizobacteria isolated from alfalfa rhizosphere and evaluation of their efficacy on tomato and watermelon growth. *Discov. Agric.* **2024**, *2*, 117. [\[CrossRef\]](#)
27. Edwards, U.; Rogall, T.; Blöcker, H.; Emde, M.; Böttger, E.C. Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res.* **1989**, *17*, 7843–7853. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Marin, D.; Silva Ferraz, G.; Santana, L.; Barbosa, B.; Barata, R.; Osco, L.; Ramos, A.; Guimarães, P. Detecting coffee leaf rust with UAV-based vegetation indices and decision tree machine learning models. *Comput. Electron. Agric.* **2021**, *190*, 106476. [\[CrossRef\]](#)
29. Liu, Y.; Feng, H.; Yue, J.; Fan, Y.; Bian, M.; Ma, Y.; Jin, X.; Song, X.; Yang, G. Estimating potato above-ground biomass by using integrated unmanned aerial system-based optical, structural, and textural canopy measurements. *Comput. Electron. Agric.* **2023**, *213*, 108229. [\[CrossRef\]](#)
30. Liu, Y.; Feng, H.; Yue, J.; Li, Z.; Yang, G.; Song, X.; Yang, X.; Zhao, Y. Remote-sensing estimation of potato above-ground biomass based on spectral and spatial features extracted from high-definition digital camera images. *Comput. Electron. Agric.* **2022**, *198*, 107089. [\[CrossRef\]](#)
31. Danaci, C.; Avci, D.; Tuncer, S. Diagnosis of pes planus from X-ray images: Enhanced feature selection with deep learning and machine learning techniques. *Biomed. Signal Process. Control* **2025**, *106*, 107769. [\[CrossRef\]](#)
32. Cover, T.M.; Thomas, J.A. Entropy, relative entropy and mutual information. In *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 1991.
33. Muthukrishnan, R.; Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. In Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24–26 October 2016; pp. 18–20. [\[CrossRef\]](#)
34. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [\[CrossRef\]](#)
35. Su, X.; Yan, X.; Tsai, C.-L. Linear regression. *WIREs Comput. Stat.* **2012**, *4*, 275–294. [\[CrossRef\]](#)
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
37. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
38. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [\[CrossRef\]](#)
39. Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1984. [\[CrossRef\]](#)
40. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
41. Peng, Y.; Zhu, T.; Li, Y.; Dai, C.; Fang, S.; Gong, Y.; Wu, X.; Zhu, R.; Liu, K. Remote prediction of yield based on LAI estimation in oilseed rape under different planting methods and nitrogen fertilizer applications. *Agric. For. Meteorol.* **2019**, *271*, 116–125. [\[CrossRef\]](#)
42. Jiang, Y.; Wei, H.; Hou, S.; Yin, X.; Wei, S.; Jiang, D. Estimation of Maize Yield and Protein Content under Different Density and N Rate Conditions Based on UAV Multi-Spectral Images. *Agronomy* **2023**, *13*, 421. [\[CrossRef\]](#)
43. Zhu, J.; Li, S.; Wu, W.; Zhao, P.; Ao, X.; Chen, H. Methodological study on maize water stress diagnosis based on UAV multispectral data and multi-model comparison. *Agronomy* **2025**, *15*, 2318. [\[CrossRef\]](#)
44. Li, Z.; Zhou, X.; Cheng, Q.; Zhai, W.; Mao, B.; Li, Y.; Chen, Z. An integrated feature selection approach to high water stress yield prediction. *Front. Plant Sci.* **2023**, *14*, 1289692. [\[CrossRef\]](#) [\[PubMed\]](#)

45. Dilmurat, K.; Sagan, V.; Maimaitijiang, M.; Moose, S.; Fritschi, F.B. Estimating crop seed composition using machine learning from multisensory UAV data. *Remote Sens.* **2022**, *14*, 4786. [[CrossRef](#)]
46. Kthiri, Z.; Jabeur, M.B.; Harbaoui, K.; Karim, A.A.; Chamekh, Z.; Chairi, F.; Serret, M.D.; Araus, J.L.; Hamada, W. Comparative performances of beneficial microorganisms on the induction of durum wheat tolerance to Fusarium head blight. *Microorganisms* **2021**, *9*, 2410. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.