



Machine learning approach for satellite-based subfield canola yield prediction using floral phenology metrics and soil parameters

Hansanee Fernando¹ · Thuan Ha¹ · Kwabena Abrefa Nketia¹ · Anjika Attanayake¹ · Steven Shirliff¹

Accepted: 11 January 2024
© The Author(s) 2024

Abstract

Early monitoring of within-field yield variability and forecasting yield potential is critical for farmers and other key stakeholders such as policymakers. Remote sensing techniques are progressively being used in yield prediction studies due to easy access and affordability. Despite the increasing use of remote sensing techniques for yield prediction in agriculture, there is still a need for medium-resolution satellite imagery when predicting canola yield using a combination of crop and soil information. In this study, we investigated the utility of remotely sensed flowering information from PlanetScope (at 4 m) satellite imagery combined with derived soil and topography parameters to predict canola yield. Our yield prediction model was trained and validated using data from 21 fields managed under variable rate seed and fertilizer application, including cleaned harvester yield maps, soil, and topography maps. To quantify the flowering intensity of canola, 9 vegetation indices (VIs) were calculated using spectral bands from PlanetScope imagery acquired for the reproductive stages of canola. We created five random forest regression models using different subsets of covariates, including VIs, soil, and topography features, to predict canola yield within the season. Using a random forest regression algorithm, we recorded accuracies ranging from poor to best performing using coefficient of determination and root mean squared error (R^2 : 0.47 to 0.66, RMSE: 325 to 399 kg ha⁻¹). The optimal subset of covariates identified electrical conductivity (EC), Normalized Difference Yellowness Index, and Canola Index as the key variables explaining within-spatial variability in canola yield. Our final model exhibited a validation R^2 of 0.46 (RMSE = 730 kg ha⁻¹), demonstrating the potential of medium-resolution satellite imagery during the flowering stage to detect and quantify sub-field spatial and temporal floral phenology changes when predicting canola yield.

Keywords Random forest · Yield prediction · Canola · Flowering intensity · Soil and topography · PlanetScope

✉ Steven Shirliff
steve.shirliff@usask.ca

¹ Department of Plant Sciences, College of Agriculture and Bioresources, University of Saskatchewan, Saskatoon S7N 5A8, Canada

Introduction

Canola is the third-largest oil crop in the world and widely used as edible oil, animal feed, and biofuel (Canola Council of Canada, 2021; Zhang et al., 2020). Canada, being the leading global producer of canola, achieved a peak revenue of 19.4 million metric tons in 2020 (Statistics Canada, 2021), with Saskatchewan contributing approximately 55% of Canada's total production. Thus, canola yield forecasts are imperative to making more informed decisions on a farmer, regional, or national level.

Remote sensing technologies have revolutionized yield estimation approaches by providing fast, accurate, and cost-effective datasets, surpassing traditional methods like surveys or crop cuttings (Lobell et al., 2019). Satellite imagery enables the estimation of field parameters such as soil properties, topographic indicators, and crop responses, facilitating the simulation of within-field spatial variability for optimized farm management practices (Liaghat & Balasundram, 2010). However, using remotely sensed imagery (i.e. multiple spectral bands) for within-field spatial variability assessments of crop yields require improved calibrations of spatial, radiometric, and temporal resolutions (Enclona et al., 2004).

Vegetation indices, which are derived from multiple spectral bands, are widely utilized in remote sensing studies to establish functional relationships with biophysical variables using empirical, mechanistic, or combined approaches (Chlingaryan et al., 2018; Weiss et al., 2020). Vegetative indices such as Normalized Difference Vegetation Index (NDVI) cannot be used to map canola flowers due to their distinctive spectral properties (Sulik & Long, 2015). For instance, a reflectance curve of a flower pixel—within a yellow canopy—is distinctively different from a similar pixel within a green canopy (Fernando et al., 2022). To address this limitation, several studies have developed new vegetation indices to map yellow flowers, such as Normalized Difference Yellowness Index (NDYI), Blue-NDVI (BNDVI), High-resolution Flowering Index (HrFI) and Canola Index (CI) (Ashourloo et al., 2019; Chen et al., 2019; Fernando et al., 2022; Sulik & Long, 2016; Zhang et al., 2021). For instance, Ashourloo et al (2019) employed Sentinel-2 time series curves to automatically map and detect flowering dates. A common feature among these studies is the utilization of the unique spectral properties of yellow flowers to capture crucial phenological information required for predicting yield potentials before harvest (Fernando et al., 2021; Sulik & Long, 2016; Zhang et al., 2021). The peak flowering period of canola has also been demonstrated to have a strong relationship with its seed yield (d'Andrimont et al., 2020; Fernando et al., 2021; Sulik & Long, 2016). Additionally, cumulative flowering intensity has been identified as a strong indicator of yield potential compared to a single-date regression approach (Zhang et al., 2021).

Spatial variability within large agricultural areas, influenced by climate, topography, soil, and biotic factors, (Jiang & Thelen, 2004), significantly impacts average seed yield (Sakamoto, 2020; Sulik & Long, 2016). In light of the growing adoption of precision agriculture, identifying within-field spatio-temporal variability to manage variable rate applications to facilitate optimum timely resource allocations are important. Accurate estimation of within-field variability requires multiple layers of precise geo-spatial data. Data-driven processes utilizing machine learning algorithms (MLAs) can leverage large volumes of data to identify patterns and relationships that may not be easily discernible through human analysis alone (Géron, 2019). Continuously refining models based on new data not only improves accuracy over time but also enable better decision-making and automation of data processing and analytic pipelines, which reduces the risk of human error while

facilitating faster and more efficient decision-making processes. MLAs have proven to be highly valuable in predicting crop yields by analyzing the complex relationships between various environmental factors that affect crop growth (Chlingaryan et al., 2018). Random forest regression (RFR) is a MLA widely used in the domain of agricultural remote sensing studies, as it simultaneously analyzes a large number of input variables compared to simple empirical approaches such as multiple regression. RFR is also computationally less expensive than complex empirical models such as neural networks (Marques Ramos et al., 2020).

To the best of our knowledge, within-season yield prediction of canola seed yield has not been attempted on a sub-field scale using medium-resolution satellite imagery, soil, and ancillary data. We hypothesize that (1) the temporal and spatial heterogeneities of canola flowering intensity within a field can be indicative of seed yield and (2) indices derived from medium-resolution satellite imagery is a reasonable predictor of the sub-field grain yield. Thus the objective of this study was to; (1) develop spectral index-based time series curves to analyze the variability of canola flowering intensity within and between fields, (2) create a random forest model that uses quantified floral intensity, soil, and topography data to predict canola seed yield, and (3) assess the significance of input variables in predicting seed yield through three RFR models and variable importance plots.

Materials and methods

Study area

The study area is located within Rose Valley, Saskatchewan, Canada ($52^{\circ} 28' N$, $103^{\circ} 80' W$), which included 21 canola fields covering approximately 20 km^2 (Fig. 1). Yield maps for 2019 for these fields were acquired from a precision agriculture company, with a spatial resolution of 2 m. Additional information about canola, including variety, planting, and harvesting dates, is provided in Table 1. The fields were managed with variable rate fertilizer applications (Fig. 2).

Data

Spatial data maps

Spatial data layers, including canola seed yield, electrical conductivity (EC) (surface soil and deep soil), elevation, and topography index, were obtained from CropPro (a precision agriculture company) in Canada. Topography index indicates the curvature of the topology, and is developed through the use of proprietary methods of ADMS 64 software version 9.9 (GK Technology Inc., 2017). Harvest data, collected using a combined harvester with a GPS-tagged yield monitor, was pre-processed to generate an artifact-free yield map. Elevation data was acquired using a mobile vehicle equipped with real-time kinematic positioning (RTK) technology, which logged elevation measurements. Additionally, an EM38-MK2 scanner was employed to capture data for the development of EC maps. (Geonics Limited, 2013). The vehicle surveyed each field with a 24.4 m swaths taking measurement every 10 s with a maximum speed of 45 km hr^{-1} . These spatial data layers were pre-processed and resampled to a resolution of 2 m using ADMS 64 software version 9.9 (GK Technology Inc., 2017).

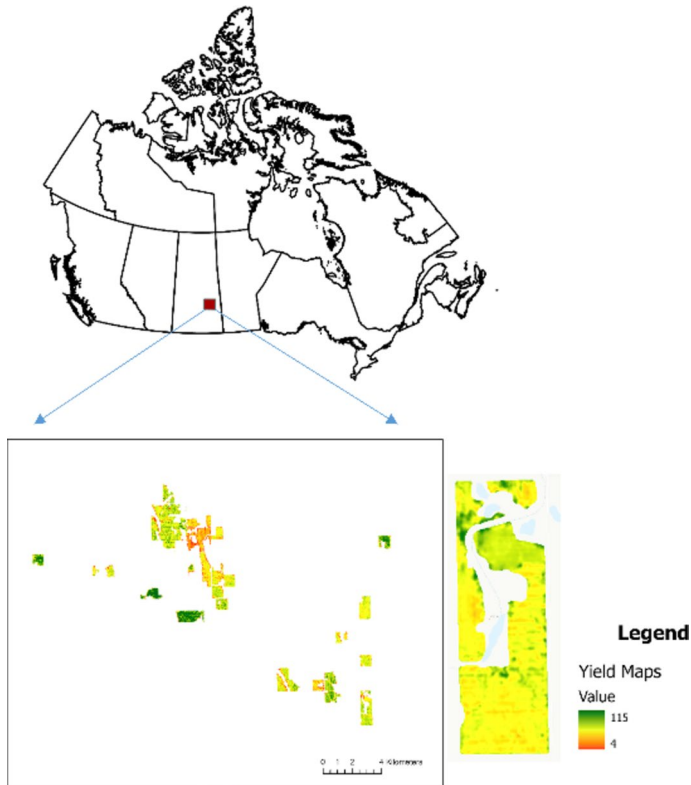


Fig. 1 Layout map indicating the study area with the spatial variability of canola seed yield of one field. The colour scale indicates the canola seed yield in Bu/Ac

Satellite datasets

Image acquisition In this study, we used 3B analytic 4-band (Blue 455–515 nm; Green 500–590 nm; Red 590–670 nm; NIR 780–860 nm) surface reflectance (SR) products from the 1st generation PlanteScope constellation (PS2: Dove classic) (Frazier & Hemingway, 2021). The SR imagery were acquired as 16-bit GeoTIFF scenes, with the values scaled by 10 000 and a ground sampling distance of 3.7 m. These multispectral images were corrected for geometric, radiometric, and atmospheric distortions indicated by the image provider (Planet Labs Inc, 2021). A total of 434 scenes were initially collected for the canola growing period, i.e. from May 12 to September 06, 2019. Ortho scenes with > 80% cloud cover were not acquired for this study. The PS2 imagery were then processed using ArcGIS Pro 10.6 to derive specific vegetation indices.

Image preprocessing and cloud masking Pre-processed orthorectified individual ortho scenes were mosaicked to produce a single image for the study area. Cloud Index (CdI) was calculated for every image date to threshold cloud and shadow affected pixels within individual images (Zhai et al., 2018). CdI-based thresholds were used to develop a binary image, which was then used to filter clouds from the corresponding raster.

Table 1 Information on 21 canola fields for year 2019 used in the study

Field ID	Crop	Area (ha)	Date seeded	Date harvested
01	Canola-Liberty (L233P)	146	14-May	Harvested Oct 9
02	Canola-Liberty (L233P)	69	13-May	Harvested Oct 9
03	Canola-RR (DKTF 92 SC)	168	09-May	Harvested Oct 18
04	Canola-RR (DKTF 92 SC)	170	10-May	Harvested Oct 20
05	Canola-RR (DKTF 92 SC)	47	09-May	Harvested Oct 18
06	Canola-RR (DKTF 92 SC)	67	10-May	Harvested Oct 9
08	Canola-Liberty (L233P)	65	15-May	Harvested Oct 18
09	Canola-Liberty (L233P)	486	15-May	Harvested Oct 20
10	Canola-RR (DKTF 92 SC)	24	09-May	Harvested Oct 9
11	Canola-RR (DKTF 92 SC)	61	09-May	Harvested Oct 9
12	Canola-RR (DKTF 92 SC)	73	09-May	Harvested Oct 10
13	Canola-Liberty (L233P)	138	09-May	Harvested Oct 9
18	Canola-Liberty (L233P)	148	10-May	Harvested Oct 17
19	Canola-Liberty (L233P)	57	12-May	Harvested Oct 10
20	Canola-Liberty (L233P)	154	12-May	Harvested Oct 17
23	Canola-Liberty (L233P)	156	11-May	Harvested Oct 17
24	Canola-Liberty (L233P)	63	11-May	Harvested Oct 11
43	Canola-Liberty (L233P)	42	10-May	Harvested Oct 20
50	Canola-Liberty (L233P)	119	11-May	Harvested Oct 11
55	Canola-RR (DKTF 92 SC)	63	09-May	Harvested Oct 10
10	Canola-Liberty (L233P)	55	10-May	Harvested Oct 6

Calculation of spectral indices Nine vegetation indices (VIs; Table 2) were utilized to quantify canopy reflectance and establish an empirical relationship with canola flowering intensity. These indices were categorized as yellowness indices (YI) including NDYI, CI, BI, RBNI, MYI, and HrFI and Greenness Indices (GIs) including NDVI, BNDVI, and VARI. Yellowness indices allow us to assess floral dynamics while GIs were used to evaluate vegetation dynamics. The VIs were calculated based on cloud-masked mosaicked images, resulting in a data stack of nine spectral index layers for a single imaging date. Time series VI curves were first generated for the entire growing season, and the reproductive stage identified using NDYI. Floral phenology was analyzed by the temporal patterns of VIs to measure the strength of yellow flowers. To determine the overall intensity of flowering, we computed the cumulative intensity by calculating the integrated area under the temporal curve (AUC). This was achieved by using the ‘trapezoid’ function (MLmetrics) in RStudio (Makowski et al., 2019). The highest value of the index during the reproductive season was used to identify the peak flowering intensity.

Data extraction

To extract predictor variables, a vector polygon of each field was created using the yield map as a reference layer. The mosaicked yield map, consisting of 21 fields, was subdivided into four classes based on quartile values: Class 1 (201–2690 kg/ha), Class 2 (2690–3564

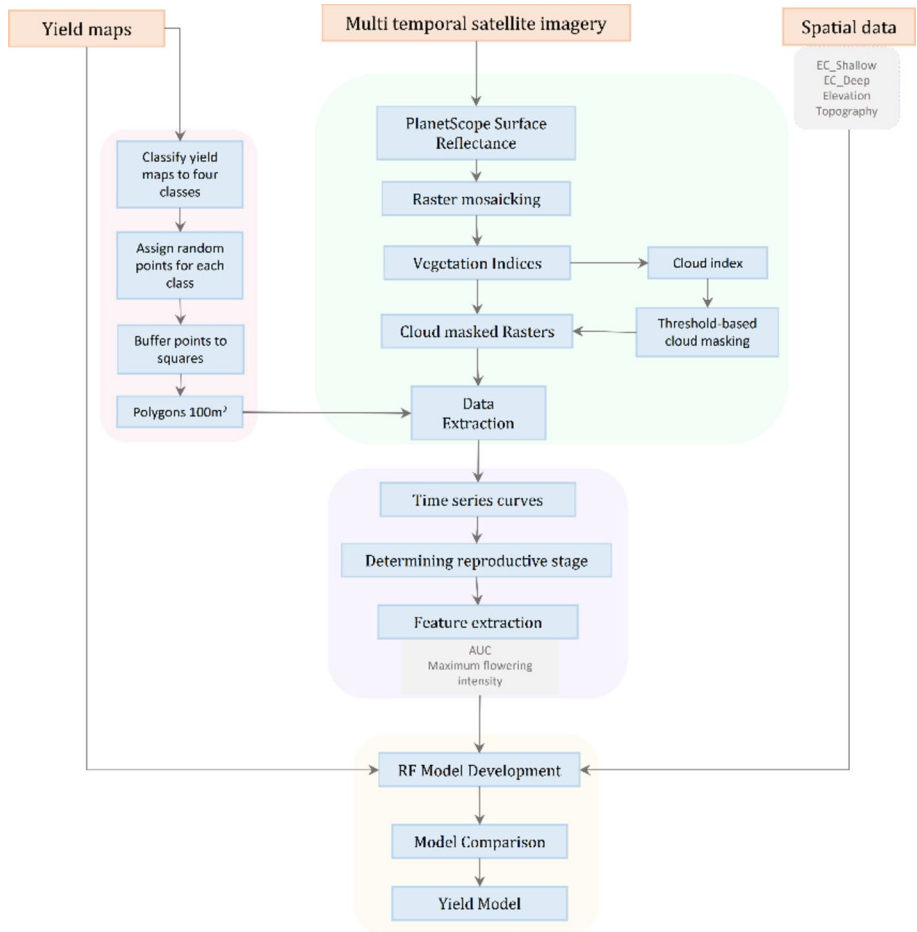


Fig. 2 The workflow of image processing, data analysis, and yield modeling procedure used for developing a random forest regression model for yield prediction

kg/ha), Class 3 (3564–4640 kg/ha), and Class 4 (4640–7733 kg/ha). 2500 points were assigned to each class using stratified random sampling. A 10 m buffer was applied to each point, resulting in square-shaped polygons with an area of 10 m². A vector layer containing 10 000 square polygons was then used to extract zonal statistics from the input data layers.

Yield model

The dataset was divided into a training subset (80%, $n \approx 8000$) and an external validation subset (20%, $n \approx 2000$) for the prediction model (Gholamy et al., 2018). A ten-fold cross-validation random forest regression model was utilized. Two hyperparameters, namely the number of trees (ntree) and the number of features considered at each split (mtry), were optimized based on the lowest root mean squared error (RMSE). The covariates were categorized into three groups: YI, GI, and Soil. The BI, despite being a yellowness index, was excluded

Table 2 Equations used for the estimation of various vegetation indices. NDYI, CI, BI, RBNI, MYI, HrFI are grouped as yellowness indices (YI). BNDVI, NDVI, VARI are grouped as vegetation indices (VI)

Spectral Index	Equation	References
Normalized Difference Yellowness Index (NDYI)	$\frac{(Green-Blue)}{(Green+Blue)}$	(1) Sulik and Long (2016)
Canola Index (CI)	$NIR(Red + Green)$	(2) Ashourloo et al. (2019)
Blooming Index (BI)	$\frac{(Red+Green)/Blue}{\frac{Green}{Blue} \times (Red-Blue+10^4)}$	(3) Chen et al. (2019)
Red Blue Normalizing Index (RBNI)	$\frac{(Red-Blue)}{(Red+Blue)}$	(4) Fernando et al. (2022)
Modified Yellowness Index (MYI)	$\frac{(Red \times Green)}{Blue}$	(5) Fernando et al. (2022)
High-resolution Flowering Index (HrFI)	$(Red - Blue) \times (Green - Blue)$	(6) Fernando et al. (2022)
Blue–NDVI (BNDVI)	$\frac{(NIR-Blue)}{(NIR+Blue)}$	(7) Sulik and Long (2016)
Normalized Difference Vegetation Index (NDVI)	$\frac{(NIR-Red)}{(NIR+Red)}$	(8) Tucker (1979)
Visible Atmospherically Resistant Index (VARI)	$\frac{(Green-Red)}{(Green+Red-Blue)}$	(9) Gitelson et al. (2002)
Cloud Index (CId)	$\frac{(3 \times NIR)}{(Green+Blue+Red)}$	(10) Zhai et al. (2018)

from the model development as it exhibited an inverse relationship with flowering intensity. To evaluate the contribution of each data group, five RFR models were constructed (Table 3) and compared using coefficient of determination (R^2). All the models were trained and validated via the ‘caret’ R package (Kuhn, 2008) within RStudio Version 3.6.1 (RStudio, 2018).

In addition to comparing data types (VIs, and soil variables) through different RFR models, the importance of each variable in the finalized model was assessed using the “VarImp” function from ‘caret’ R package (Kuhn, 2008). This function calculates the contribution of each predictor by recording the mean squared error of out-of-bag data for each tree, after permuting each predictor variable. The differences are then averaged and normalized by the standard error. To facilitate comparison between input features, the variable importance plots were scaled as relative importance from 0 to 100, where a score of 100 represents the most important variable, and a score of 0 represents the least important variable. Detailed descriptions of the input variables can be found in the supplementary materials Table S1. Furthermore, the importance and stability of information obtained from VIs and soil data were additionally analyzed using the “holding-out” method. This involved iterating model3 (Soil Model), model4 (Flower and Soil Model), and model5 (Flower Model), a number of times (in this case 21 times), with each iteration holding out a different field for validation and utilizing the remaining fields for training. The resulting distribution of correlation (R) values was examined to assess the stability of these models. After selecting the optimal model, a feature reduction step was performed by eliminating variables with correlations exceeding 0.9. For further exploration, the relationship of individual variables to canola seed yield was explored using simple linear regression for each of the 21 fields.

Table 3 Variables used in developing the five random forest regression models

	RFR1: Flower-Green-Soil Model	RFR2: Flower-Green Model	RFR3: Soil-Only Model	RFR4: Flower-Soil Model	RFR5: Flower-Only Model
Model explanation	Model built using all VI-based input parameters (quantified canola intensity), soil and topography parameters	Model built with only VI-based input parameters	Model built with only soil and topography parameters	Model built with selected yellowness indices and soil parameters	Model built with only selected yellowness indices
Indices	NDYI CI HrFI MYI RBNI BI NDVI VARI BNDVI	NDYI CI HrFI MYI RBNI BI NDVI VARI BNDVI		NDYI CI HrFI MYI RBNI	NDYI CI HrFI MYI RBNI
Soil parameters	EC Topography Elevation		EC Topography Elevation	EC Topography Elevation	

Results

Yield model

All RFR trained models, except for Soil-Only model (RFR3), were able to explain approximately 65% of the variability in yield (R^2) with similar RMSE, averaging at 350 kg ha^{-1} (Table 4). The Soil-Only model, explained 47% of the yield variability with RMSE of 918 kg ha^{-1} . During model validation, Flower-Soil model (RFR4) exhibited the highest correlation (R^2 0.46, RMSE 730 kg ha^{-1}), closely followed by the Flower-Green-Soil model (RFR1; R^2 0.45, RMSE 760 kg ha^{-1}). Despite the similar predictive power (R^2) between these two models, the number of inputs differed significantly. The Flower-Soil model, with 14 predictors, demonstrated a slightly better performance to the Flower-Green-Soil model, which utilized 23 predictors.

Despite the lower performance of the Soil-Only model, incorporating soil-related parameters alongside spectral predictors improved the model's predictive power. This is evident from the higher validation metrics observed in the Flower-Soil model ($R^2=0.46$) and the Flower-Green-Soil model ($R^2=0.45$) compared to the Soil-Only model ($R^2=0.18$). The Flower-Green model (RFR2) and the Flower-Only model (RFR5), which solely utilized spectral information as predictor variables, achieved a validation R^2 of 0.27 and 0.29, respectively, between the actual and predicted seed yields.

Cross-validation results reveal that small changes in the training data can significantly influence the model power (Fig. 3). The Soil-Only model shows a wide range of R values (-0.3 to 0.70), while the Flower-Soil model, exhibits a narrower range of R values (0.3 to 0.8).

In Flower-Soil model, the number of predictors were reduced from 14 to 9 by eliminating highly correlated variables ($R > 0.9$) 'Max.CI', 'EC_Deep', 'CI_AUC', 'MYI_AUC', and 'Max.MYI'. The correlation matrix of the variables is presented in Supplementary Materials Fig. S1. This variable reduction reduced the R of the validation plot by 19% from 0.46 (14 predictors) to 0.37 (9 predictors) and increased the RMSE from 730 to 824 kg ha^{-1} . Thus, in the final yield model, all 14 variables were used. The validation plot of the RFR4 model is presented in Fig. 4a.

Table 4 Comparison of performance metrics of the random forest regression models

Models	Number of predictors	mtry/ntree	Training R^2	Training RMSE (kg ha^{-1})	Validation R^2	Validation RMSE (kg ha^{-1})
RFR1: Flower-Green-Soil	23	12/500	0.66	318	0.45	761
RFR2: Flower-Green	19	10/500	0.65	343	0.27	872
RFR3: Soil-Only	4	2/500	0.47	399	0.18	918
RFR4: Flower-Soil	14	8/500	0.66	325	0.46	730
RFR5: Flower-Only	10	2/500	0.64	358	0.29	865

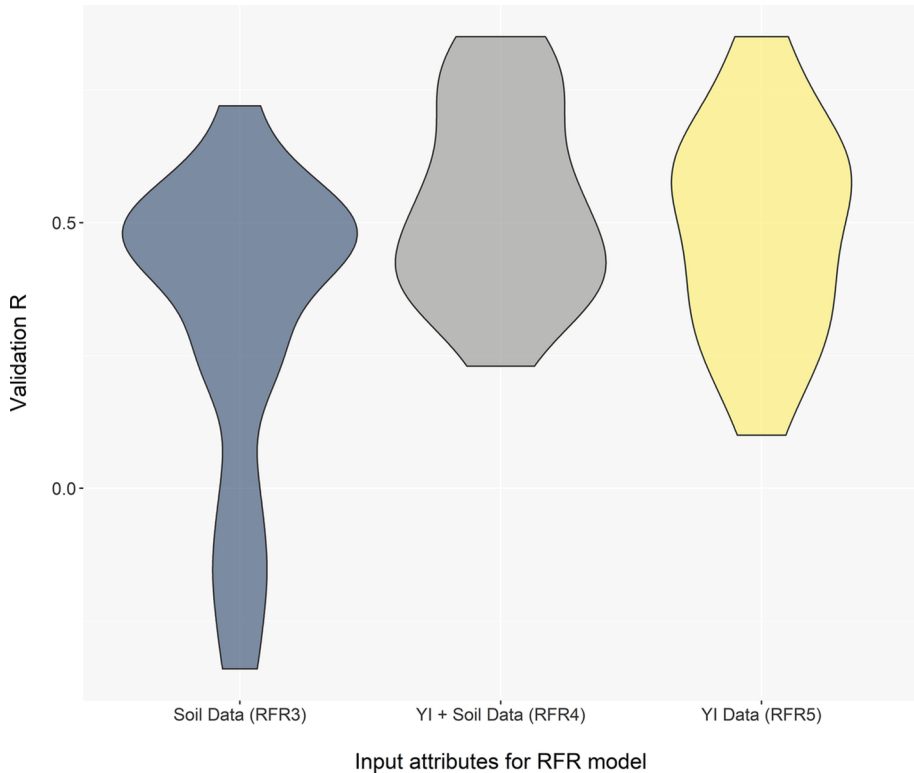


Fig. 3 Violin plots illustrating the distribution of validation R obtained from the leave-one-field out cross validation (Pearson's correlation coefficient) for RFR3, RFR4, and RFR5 models. The models were iterated 21 times, each time holding out an individual field out for validation. Wider vertical distribution in R value implies that the exclusion of a single field during model iteration can result in substantial changes in model performance

Variable importance

The three most important variables identified were Shallow EC (topsoil electrical conductivity), Max.CI (peak flowering intensity captured by the Canola Index), and NDYI_AUC (cumulative flowering intensity measured through NDYI) (Fig. 4b). Significance of shallow EC as a predictor is further supported by variable importance plots in the Flower-Green-Soil, Soil-Only, and Flower-Soil models, where it consistently ranks as the most important variable (Figures not included). In models incorporating spectral indices, Max.CI and NDYI_AUC also ranked among the top three variables, albeit with slight variations in ranking across different models. The importance of using different data types (Flower, Leaf and Soil) is further highlighted by the variable importance plots, where Shallow EC ranked 1st followed by either peak flowering intensity or cumulative intensity. Notably, the EC of the subsoil was found to be less important (Fig. 4b). The relationship between EC_Shallow and canola seed yield across different fields is presented in Supplementary Materials Fig. S2. The figures show moderate correlations ranging from 0.5 to 0.7. Similarly, the relationship between Max.CI and canola seed yield across different fields is presented in Supplementary Materials Fig. S3. Max.CI, representing peak flowering intensity, exhibits

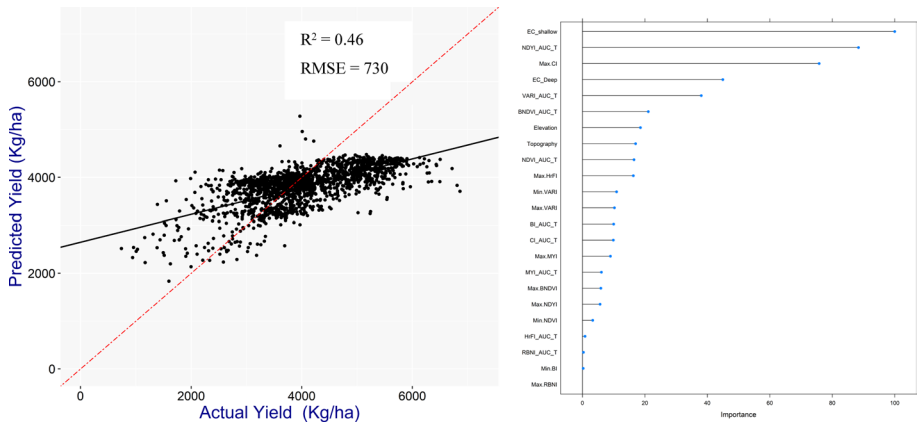


Fig. 4 **a** Validation plot of the theoretical performance of the random forest regression model built using yellowness indices, soil, and topography variables (RFR4) for the canola seed yield with selected features. The validation resulted in an R^2 of 0.46 and an RMSE of 730 kg ha^{-1} , **b** Normalized variable importance RFR4. The values are scaled from 0 to 100 (most important). EC_Shallow: Electrical Conductivity of the topsoil, Max.CI: Maximum flowering intensity quantified through CI, NDYI_AUC: Cumulative flowering intensity quantified through NDYI, EC_Deep: Electrical Conductivity of the subsoil, CI_AUC: Cumulative flowering intensity quantified through CI, Max.HrFI: Maximum flowering intensity quantified through HrFI, MYI_AUC: Cumulative flowering intensity quantified through MYI, Max.NDYI: Maximum flowering intensity quantified through NDYI, Elevation: Elevation measured in feet, Max.MYI: Maximum flowering intensity quantified through MYI, Topography: Topography index, RBNI_AUC: Cumulative flowering intensity quantified through RBNI, Max.RBNI: Maximum flowering intensity quantified through RBNI, HrFI_AUC: Cumulative flowering intensity quantified through HrFI

a positive correlation with seed yield, ranging from 0.16 to 0.86 with most fields having a R of above 0.45, indicating the importance of the peak flowering intensity variable in yield prediction. AUC_NDYI, representing cumulative flowering intensity, shows a similar relationship to Max.CI, with most R values above 0.45.

The correlation coefficients between the input variables and canola seed yields across all 21 fields demonstrated a consistent trend in decreasing R values as the variable importance decreases (Fig. 5). This trend holds true until the 11th ranked variable, which corresponded to topography. Surprisingly, despite topography exhibiting the lesser overall correlation with seed yield, it is assigned a greater variable importance compared to variables like 'AUC_RBNI', 'Max.RBNI', and 'AUC_HrFI', which showed stronger correlations with seed yield.

Spatial and temporal variability of spectral information

The temporal dynamics of nine VIs, as depicted in Fig. 6, coincide with the changes in flowering intensity over time. At the onset of flowering, indices such as BI, NDVI, and VARI exhibited a decline in values as flowering intensity increases, reaching a depression at the peak flowering period (Julian Date 200). Subsequently, these indices gradually increased as the plants progressed into the pod stage. Notably, among the yellowness indices, only BI demonstrated a negative correlation with flowering. The depressions observed in BI and VARI are more pronounced compared to NDVI across all fields. Conversely, the greenness index BNDVI and other yellowness indices (CI,

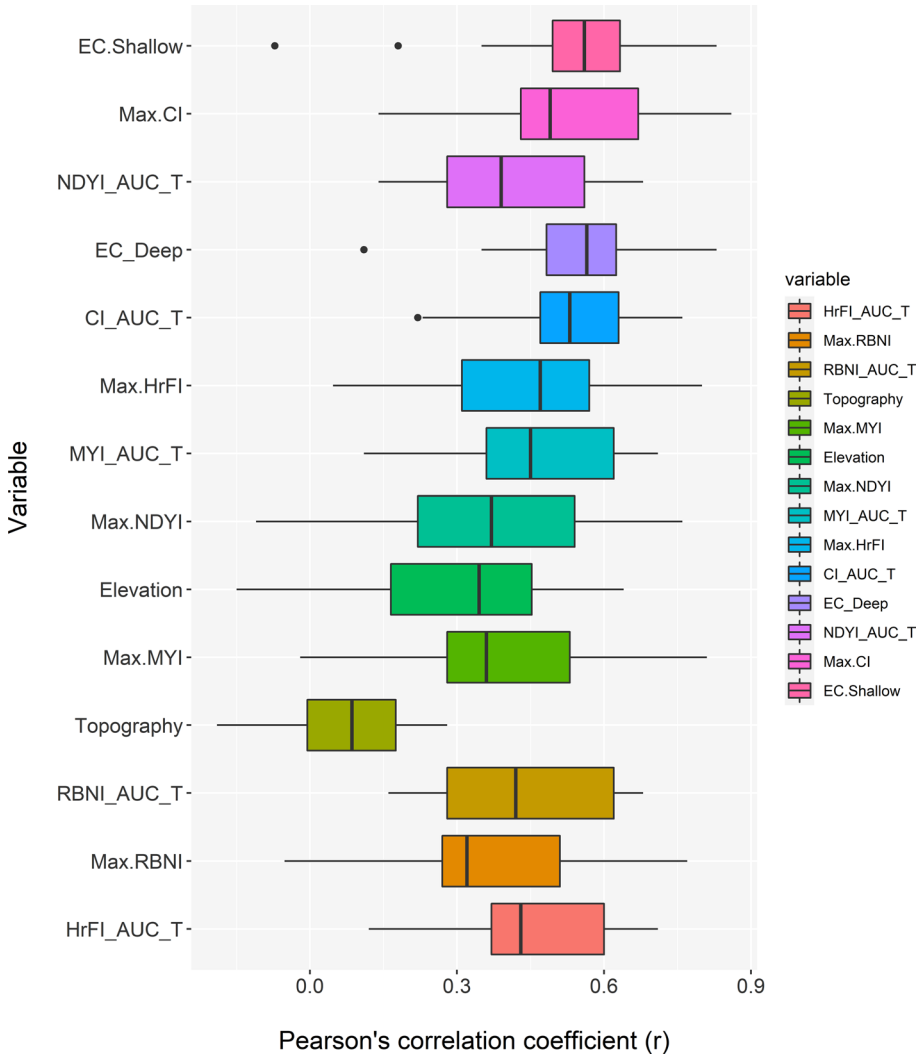


Fig. 5 Variability in Pearson's correlation coefficient between the input feature and the canola seed yield for 21 canola fields. EC_Shallow: Electrical Conductivity of the topsoil, Max.CI: Maximum flowering intensity quantified through CI, NDYI_AUC: Cumulative flowering intensity quantified through NDYI, EC_Deep: Electrical Conductivity of the subsoil, CI_AUC: Cumulative flowering intensity quantified through CI, Max.HrFI: Maximum flowering intensity quantified through HrFI, MYI_AUC: Cumulative flowering intensity quantified through MYI, Max.NDYI: Maximum flowering intensity quantified through NDYI, Elevation: Elevation measured in feet, Max.MYI: Maximum flowering intensity quantified through MYI, Topography: Topography index, RBNI_AUC: Cumulative flowering intensity quantified through RBNI, Max.RBNI: Maximum flowering intensity quantified through RBNI, HrFI_AUC: Cumulative flowering intensity quantified through HrFI

HrFI, MYI, NDYI, RFI) exhibited a positive relationship with flowering, wherein higher flowering intensity corresponded to elevated index values, peaking around Julian Date 200. Furthermore, the differences in index values between the four yield classes were

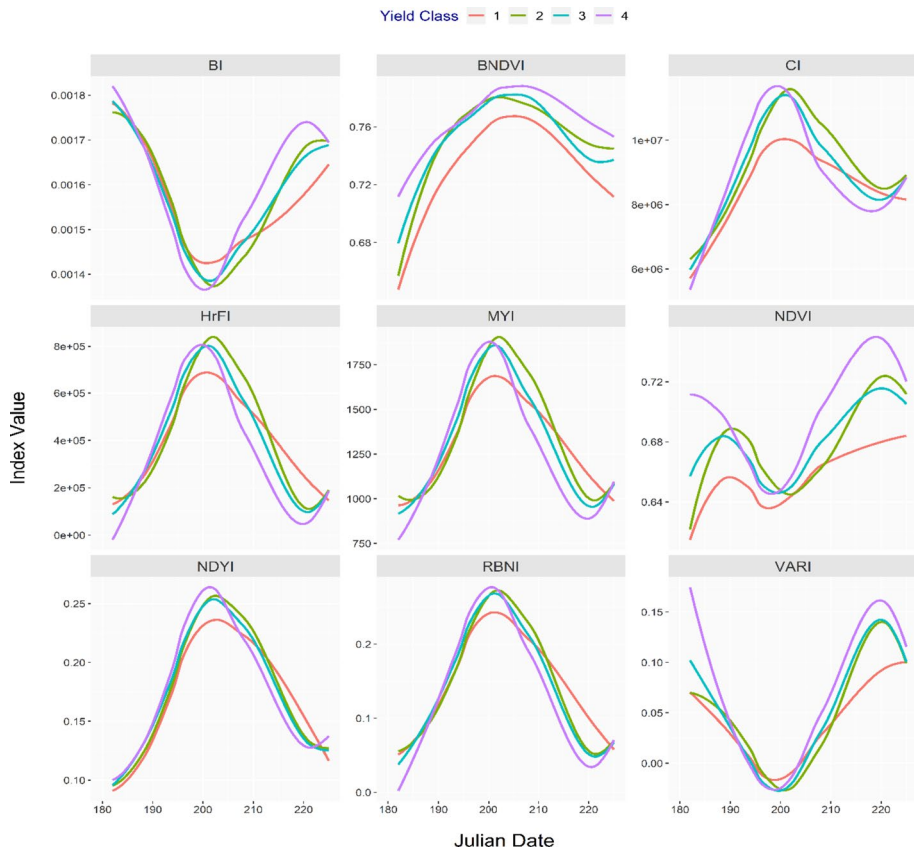


Fig. 6 Example time series curves of VIs depicting the index value change with flowering phenology for Field ID 19. The lines are smoothed using the “Loess” function from Library “ggplot2”. Four yield classes: Class 1 (201–2690), Class 2 (2690–3564), Class 3 (3564–4640), and Class 4 (4640–7733)

most distinct during the peak flowering period, particularly in class 1, and this pattern was consistent across most yellowness indices.

Discussion

The combination of soil and spectral information in predicting canola yield demonstrate superior predictive power compared to their individual usage (Table 4). The limited vertical distribution observed in the Flower-Soil model (Fig. 3) further confirms the stability of the model when both soil and yellowness spectral information are utilized, in contrast to using either Flower-Only or Soil-Only models. The performance of the Flower-Only model was comparable or slightly better when combined with the Greenness model. The underperformance of using only greenness indices during flowering can be attributed to the weak relationship between vegetation indices and canopy reflectance during peak flowering (Sulik & Long, 2016).

The validation plot of the Flower-Soil model (Fig. 4a) exhibited a trend of underestimating canola seed yield values exceeding 4500 kg/ha. This can be attributed to the saturation effect observed in vegetation indices at higher biomass levels, where the increase in flowering intensity does not correspond proportionally to an increase in the yellowness index value. Furthermore, most of the training data being concentrated around or below the average yield of 4500 kg ha⁻¹ contributes to the model's limited ability to extrapolate to yield values outside the range of the training data. This is a common drawback of machine learning algorithms, where their predictive performance may be constrained by the training data distribution. To improve the model's performance in estimating high-yield scenarios, it is essential to consider the representativeness of training data and explore approaches that address the saturation effect and enable better extrapolation capabilities.

Spectral information demonstrated stronger predictive capabilities for canola yield compared to the soil information used in this study. This observation can be attributed to the distinct nature of these data types. Soil information represents the pre-growing season condition of the soil, providing insight into its suitability for plant growth. In contrast, spectral information captures the dynamic response of plants throughout the growing season, encompassing the influence of soil, genetics, and environmental factors. Consequently, spectral information indirectly reflects the soil conditions and the canola's response to its growing environment. Thus, it is reasonable to conclude that spectral covariates enhance predictive power more effectively than soil variables.

The canola fields used in this study were subject to variable rate applications, but the specific input rates were not utilized as variables in the model development. It is important to note that management decisions regarding crop inputs can influence crop yield, and thus, they may have impacted the performance of the model. In a comprehensive 40-year study by Wang (2021), a knowledge-driven model called GEM (Genetic, Environment, and Management) was developed to emphasize the significance of incorporating management information for accurate yield predictions. However, even in the absence of site-specific rate information in the present study, the developed model achieved a R^2 value of 0.46 in predicting yield. In using MLAs in forecasting yield, there is a primary focus on improving interpretability and explainability of the algorithmic processes (Gilpin et al., 2018). RFR models offer the advantage of ranking variables based on their importance, which contributes to a better understanding of the data. For instance, the variable importance plot revealed the significance of shallow EC and peak and cumulative flowering intensities in predicting canola yield.

Shallow EC, CI.Max (Peak flowering intensity captured by the CI), and AUC_NDYI (Cumulative flowering intensity captured by the NDYI) were identified as the most important features consistently across different model iterations, highlighting their significance for predicting canola seed yield. EC shallow reflects the suitability of soil for optimum plant growth as saline soils hinders water movement into plant roots (Machado & Serralheiro, 2017). CI.Max and AUC_NDYI represents reproductive potential of the crop, where each flower produces one silique that produce multiple canola seed, hence a direct indicator of yield. Similar results have been reported in the literature, where it was demonstrated that canola seed yield prediction is improved when spectral information is used with precipitation and soil data (Wen et al., 2021).

Feature selection in MLAs reduces the dimensions and enhances computational efficiency (Goodfellow et al., 2016). Removing highly correlated variables ($R > 0.9$) reduced the model power by 19%, suggesting correlated predictors have some impact on random forest models as the data is randomly selected when developing decision trees. Highly correlated variables can affect the ranking of feature importance, wherein despite high

correlation of AUC_HrFI with the yield (Fig. 5), it is ranked as the least important variable. The time series curve of HrFI is very similar to other yellowness indices, especially with NDYI, which ranked as the 3rd most variable (Fig. 6).

The time series curves of the vegetation indices also exhibited the expected patterns, demonstrating slight differences among the four yield classes. Time series curves of average normalized difference yellowness index (NDYI) for four yield classes are presented in Supplementary Materials Fig. S4. It is worth noting that the impact of radiometric calibrations may be insignificant since the canola fields are in the same area. However, when using PlanetScope imagery across a larger area, the lack of agreement between individual satellites could affect data quality. Nonetheless, several studies have shown that despite quality differences between scenes, PlanetScope imagery can produce similar or even improved results compared to Landsat and Sentinel-2 imagery (Mansaray et al., 2021; Moon et al., 2021; Shimizu et al., 2020).

The use of PlanetScope satellite imagery in this study enabled the identification of temporal and spatial variability in flowering within the canola fields, despite the variation in radiometric and geometric quality compared to Landsat and Sentinel imagery (Frazier & Hemingway, 2021). The developed yellowness indices showed a moderate correlation with canola yield at the individual field level (Figs. S2 and S3), also showing slight differences between the four yield classes. When using PlanetScope imagery across a larger area, the lack of agreement between individual satellites could affect data quality. Nonetheless, several studies have shown that despite quality differences between scenes, PlanetScope imagery can produce similar or even improved results compared to Landsat and Sentinel-2 imagery (Mansaray et al., 2021; Moon et al., 2021; Shimizu et al., 2020).

Canola is harvested about 50 days after flower initiation (Canola Council of Canada, 2011), and this allows farmers with a window for decision-making regarding harvesting or swathing and pest management, optimizing yield potential. Obtaining the input data layers (elevation, EC, topography) used in this study would no longer be difficult since more farmers are moving toward the practice of precision agriculture. Additionally, using vegetation indices from satellite data to determine reproductive potential is far less labour-intensive and time-consuming than manually counting canola flowers. While this study only used the information from the flowering stage, it is highly possible that the information from both the vegetative and pod maturity stages could have improved the predictive power of the model. Nevertheless, it is critical to map and quantify canola reproductive stage as an indicator of yield potential, as the unique spectral reflectance of canola flowers provides an invaluable opportunity to distinguish between the reproductive stage and its vegetative stage readily.

The accuracy of yield estimation was found to be higher when using integrated flower accumulation in UAV imagery compared to a single image at a specific time-step (Zhang et al., 2021). This is similar in satellite imagery as well, since regressing spectral indices to yield on individual dates had lower R^2 (<0.3) (Fig. not included) than the developed random forest model.

A One potential avenue for improvement in future studies is the utilization of freely available satellite imagery sources such as Landsat and Sentinel-2 data, which can be directly accessed through cloud-based platforms like Google Earth Engine (GEE). This would alleviate the current limitation of restricted access to PlanetScope imagery and reduce computational time, making it more convenient for researchers. Additionally, previous research has demonstrated successful identification of peak flowering dates using Sentinel-1 and Sentinel-2 imagery with an accuracy of 1–4 days, providing further possibilities for enhancing the methodology (d'Andrimont et al., 2020). Furthermore, expanding

the current process to include a larger dataset across multiple provinces would enable the incorporation of variations in weather and phenology, leading to more robust and generalizable results.

Supplementary materials

The following are available online. Table S1 Input variables used in building the random forest regression. Figure S1 Correlation matrix of the input variables used in developing the random forest regression. Figure S2 Scatter plots and linear regression line between the canola seed yield and electrical conductivity of the topsoil. Figure S3 Scatter plots and linear regression line between the canola seed yield and the peak flowering intensity quantified through Canola Index (CI). Figure S4 Time series curves of average normalized difference yellowness index (NDYI) for four yield classes.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11119-024-10116-1>.

Acknowledgements The authors would like to acknowledge the contributions from CropPro Consulting.

Author contributions Conceptualization, HF, TH and SJS; methodology, HF, TH, AA, and KAN; software, HF; formal analysis, HF; data curation, HF; writing—original draft preparation, HF; writing—review, and editing, HF, TH, AA and SJS; supervision, SJS; project administration, SJS; funding acquisition, SJS. All authors have read and agreed to the published version of the manuscript.

Funding The authors would like to acknowledge the P2IRC funding 422082, Canada First Research Excellent Fund (CFREF), Canola Council, and Natural Sciences and Engineering Research Council (NSERC). H.F is funded through Western Grains Research Foundation and Mitacs Accelerate program.

Data availability Some data layers are private and are not available to public.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ashourloo, D., Shahrabi, H. S., Azadbakht, M., Aghighi, H., Nematollahi, H., Alimohammadi, A., & Matkan, A. A. (2019). Automatic canola mapping using time series of sentinel 2 images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 156, 63–76.
- Canola Council of Canada. (2011). Canola needs a month after flowering to mature. <https://www.canolacouncil.org/canola-watch/2011/08/04/canola-needs-a-month-after-flowering-to-mature/#:~:text=Growing%20conditions%20play%20a%20major,are%20coming%20out%20of%20flower.>
- Canola Council of Canada. (2021). About Canola- Industry Overview. <https://www.canolacouncil.org/about-canola/industry/>

- Chen, B., Jin, Y., & Brown, P. (2019). An enhanced bloom index for quantifying floral phenology using multi-scale remote sensing observations. *ISPRS Journal of Photogrammetry and Remote Sensing*, *156*, 108–120. <https://doi.org/10.1016/j.isprsjprs.2019.08.006>
- Chlingaryan, A., Sukkariéh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69. <https://doi.org/10.1016/j.compag.2018.05.012>
- d'Andrimont, R., Taymans, M., Lemoine, G., Ceglar, A., Yordanov, M., & van der Velde, M. (2020). Detecting flowering phenology in oil seed rape parcels with Sentinel-1 and -2 time series. *Remote Sensing of Environment*, *239*, 111660. <https://doi.org/10.1016/j.rse.2020.111660>
- Enclona, E. A., Thenkabail, P. S., Celis, D., & Diekmann, J. (2004). Within-field wheat yield prediction from IKONOS data: A new matrix approach. *International Journal of Remote Sensing*, *25*, 377–388. <https://doi.org/10.1080/0143116031000102485>
- Fernando, H., Ha, T., Attanayake, A., Benaragama, D., Nketia, K. A., Kanmi-Obembe, O., & Shirtliffe, S. J. (2022). High-resolution flowering index for canola yield modelling. *Remote Sensing*, *14*(18), 4464. <https://doi.org/10.3390/rs14184464>
- Fernando, H., Ha, T., Duddu, H., Benaragama, D., Attanayake, A., Olakorede, K.-O., et al. (2021). Canola yield simulation through digitalized flower number using high-resolution UAV-RGB imagery. *Earth and Space Science Open Archive*. <https://doi.org/10.1002/essoar.10508314.1>
- Frazier, A. E., & Hemingway, B. L. (2021). A technical review of planet smallsat data: Practical considerations for processing and using planetscope imagery. *Remote Sensing*, *13*, 3930. <https://doi.org/10.3390/rs13193930>
- Geonics Limited. (2013). EM38-MK2. <http://www.geonics.com/html/em38.html>
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media Inc.
- Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. https://scholarworks.utep.edu/cs_techrep/1209/#:~:text=Empirical%20studies%20show%20that%20the,of%20the%20data%20for%20training
- Gitelson, A. A., Kaufman, Y. J., Stark, R., & Rundquist, D. (2002). Novel algorithms for remote estimation of vegetation fraction. *Remote Sensing of Environment*, *80*, 76–87. <https://doi.org/10.1078/0176-1617-01176>
- GK Technolgy Inc. (2017). The ultimate Ag data mapping solution broken into task-oriented modules. <https://gktechinc.com/adms-product/>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* <http://www.deeplearningbook.org>
- Jiang, P., & Thelen, K. (2004). Effect of soil and topographic properties on crop yield in a north-central corn-soybean cropping system. *Agronomy Journal*. <https://doi.org/10.2134/agronj2004.0252>
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*, 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Liaghat, S., & Balasundram, S. (2010). A review: The role of remote sensing in precision agriculture. *American Journal of Agricultural and Biological Science*. <https://doi.org/10.3844/ajabssp.2010.50.55>
- Lobell, D. B., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T., & Murray, S. (2019). Eyes in the sky, boots on the ground: Assessing satellite- and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economics*, *102*, 202–219. <https://doi.org/10.1093/ajae/aaz051>
- Machado, R. M. A., & Serralheiro, R. P. (2017). Soil salinity: Effect on vegetable crop Growth. Management practices to prevent and mitigate soil salinization. *Horticulturae*, *3*, 30. <https://doi.org/10.3390/horticulturae3020030>
- Makowski, D., Ben-Shachar, M., & Lüdtke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*, 1541. <https://doi.org/10.21105/joss.01541>
- Mansaray, A. S., Dzialowski, A. R., Martin, M. E., Wagner, K. L., Gholizadeh, H., & Stoodley, S. H. (2021). Comparing PlanetScope to Landsat-8 and Sentinel-2 for sensing water quality in agricultural watersheds. *Remote Sensing*, *13*, 1847. <https://doi.org/10.3390/rs13091847>
- Marques Ramos, A. P., Prado Osco, L., Elis Garcia Furuya, D., Nunes Gonçalves, W., Cordeiro Santana, D., Pereira Ribeiro Teodoro, L., et al. (2020). A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices. *Computers and Electronics in Agriculture*, *178*, 105791. <https://doi.org/10.1016/j.compag.2020.105791>
- Moon, M., Richardson, A. D., & Friedl, M. A. (2021). Multiscale assessment of land surface phenology from harmonized Landsat 8 and Sentinel-2, PlanetScope, and PhenoCam imagery. *Remote Sensing of Environment*, *266*, 112716. <https://doi.org/10.1016/j.rse.2021.112716>

- Planet Labs Inc. (2021, June 12). Planet imagery product specifications. https://assets.planet.com/docs/Planet_Combined_Imagery_Product_Specs_letter_screen.pdf
- RStudio, T. (2018). RStudio: Integrated development environment for R. <http://www.rstudio.com/>
- Sakamoto, T. (2020). Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing*, *160*, 208–228. <https://doi.org/10.1016/j.isprsjprs.2019.12.012>
- Shimizu, K., Ota, T., Mizoue, N., & Saito, H. (2020). Comparison of multi-temporal planetscope data with Landsat 8 and Sentinel-2 data for estimating airborne LiDAR derived canopy height in temperate forests. *Remote Sensing*, *12*, 1876. <https://doi.org/10.3390/rs12111876>
- Statistics Canada. (2021). Table: 32-10-0359-01 Estimated areas, yield, production, average farm price and total farm value of principal field crops, in metric and imperial units.
- Sulik, J. J., & Long, D. S. (2015). Spectral indices for yellow canola flowers. *International Journal of Remote Sensing*, *36*, 2751–2765. <https://doi.org/10.1080/01431161.2015.1047994>
- Sulik, J. J., & Long, D. S. (2016). Spectral considerations for modeling yield of canola. *Remote Sensing of Environment*, *184*, 161–174. <https://doi.org/10.1016/j.rse.2016.06.016>
- Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150. [https://doi.org/10.1016/0034-4257\(79\)90013-0](https://doi.org/10.1016/0034-4257(79)90013-0)
- Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, *236*, 111402. <https://doi.org/10.1016/j.rse.2019.111402>
- Wen, G., Ma, B.-L., Vanasse, A., Caldwell, C. D., Earl, H. J., & Smith, D. L. (2021). Machine learning-based canola yield prediction for site-specific nitrogen recommendations. *Nutrient Cycling in Agroecosystems*, *121*, 241–256. <https://doi.org/10.1007/s10705-021-10170-5>
- Zhai, H., Zhang, H., Zhang, L., & Li, P. (2018). Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *144*, 235–253. <https://doi.org/10.1016/j.isprsjprs.2018.07.006>
- Zhang, J., Zhao, B., Yang, C., Shi, Y., Liao, Q., Zhou, G., et al. (2020). Rapeseed stand count estimation at leaf development stages with UAV imagery and convolutional neural networks. *Frontiers in Plant Science*, *11*, 617. <https://doi.org/10.3389/fpls.2020.00617>
- Zhang, T., Vail, S., Duddu, H., Parkin, I. A. P., Guo, X., Johnson, E. N., & Shirtliffe, S. J. (2021). Phenotyping flowering in canola (*Brassica napus* L.) and estimating seed yield using an unmanned aerial vehicle-based imagery. *Frontiers in Plant Science*, *12*, 686332. <https://doi.org/10.3389/fpls.2021.686332>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.