

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

# Remote Sensing Applications: Society and Environment

journal homepage: [www.elsevier.com/locate/rsase](http://www.elsevier.com/locate/rsase)

## Estimation of soybean grain yield from multispectral high-resolution UAV data with machine learning models in West Africa

Tunrayo R. Alabi<sup>a,\*</sup>, Abush T. Abebe<sup>a</sup>, Godfree Chigeza<sup>b</sup>, Kayode R. Fowobaje<sup>a</sup><sup>a</sup> International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria<sup>b</sup> International Institute of Tropical Agriculture (IITA), Lusaka, Zambia

## ARTICLE INFO

## Keywords:

Soybean  
 Gray level co-occurrence matrix (GLCM)  
 Yield prediction  
 Machine learning models  
 Vegetation indices

## ABSTRACT

Soybean (*Glycine max* (L.) Merr.) is a leguminous and oil crop with rapidly growing importance and demand in Africa following the increasing demand for oil and livestock and poultry feed in sub-Saharan Africa. However, soybean productivity is low in most countries of sub-Saharan Africa, especially in West Africa, where productivity is below one ton per ha. Hence, concerted soybean varietal development and testing efforts have been underway by the International Institute of Tropical Agriculture (IITA), collaborating with the various African and US-based soybean breeding programs. Integrating new varietal evaluation approaches based on advanced phenotyping techniques into IITA's soybean breeding program is crucial for designing efficient crop genetic improvement techniques. Hence, this work aims to investigate machine learning (ML) models and Unmanned Aerial vehicles (UAVs) to aid rapid high throughput phenotypic workflow for soybean yield estimation. We acquired multispectral images through a Sequoia® camera aboard a senseFly eBee X UAV from five variety trials during the 2020 growing season in Nigeria. UAV-based spectral bands, canopy height, vegetation indices (VI), and texture features were generated by gray level co-occurrence matrix (GLCM) and integrated to predict crop grain yield using five machine learning (ML) regression models, including Cubist, Extreme Gradient Boosting (XGBoost), Stochastic Gradient Boosting (GBM), Support vector machine (SVM), and Random Forest (RF). The main findings are the textural information generated using gray level co-occurrence matrix (GLCM) slightly outperformed predictors based mainly on vegetation indices (VI) and provided a promising alternative to the conventional use of VI in crop yield estimation. All the five ML models performed moderately well in predicting grain yield for all the soybean trials investigated, though the Cubist and RF model stood out, with  $R^2$  reaching 0.89. The study provides a framework to perform crop breeding trial assessments more effectively and consistently at high spatial scales that African crop breeding programs did not commonly apply. The workflow can also be successfully modified and applied for high throughput phenotyping of breeding platforms in other crops.

### 1. Introduction

Soybean (*Glycine max* (L.) Merr.) has enormous economic benefits in African smallholder farming systems because of its high

\* Corresponding author.

E-mail address: [t.alabi@cgiar.org](mailto:t.alabi@cgiar.org) (T.R. Alabi).

<https://doi.org/10.1016/j.rsase.2022.100782>

Received 12 October 2021; Received in revised form 12 May 2022; Accepted 20 May 2022

Available online 25 May 2022

2352-9385/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

nutritional value, raw material for edible oil processing factories, and soil-enriching potential because of its symbiotic N<sub>2</sub> fixation (Sinclair et al., 2014). Soybean seeds are high in protein content and therefore have excellent nutritional values for the low-income population of the developing world who can not afford animal source proteins. Soybean was a crop recently introduced into the African farming system with minimal options of high-yielding varieties, especially for smallholder farmers. Smallholder farmers are commonly experiencing lower yields than the global average (FAO, 2021; Khojely et al., 2018) and other soybean-producing Western countries (Diers and Scaboo, 2019; Santos, 2019). Therefore, the role of high-yielding soybean varieties in ensuring nutritional security and diversifying their on-farm income sources is of paramount importance. Thus, the International Institute of Tropical Agriculture soybean breeding program (IITASBP) has been working to develop new, improved soybean varieties for sub-Saharan Africa (SSA). IITASBP employs hybridization and development of breeding lines and the introduction of exotic germplasm from various countries to evaluate and identify high-yielding and well-adapted soybean varieties. The program has played a substantial role in providing advanced soybean breeding lines for the national soybean breeding programs in SSA countries, including Nigeria, which led to the release of several IITA varieties, most of which are still under production by farmers (Chigeza et al., 2019). As part of this effort, the IITASBP implemented two advanced yield trials viz., Advanced Variety Trial Early Set (AVTES), and Medium Set (AVTMS), and one Pan-African Variety Trial (PAVT), and two preliminary variety trials (PVT) i.e., set 01 (PVT01) and Set 02 (PVT02) in the 2020 cropping season at Ibadan, Nigeria.

In crop improvement programs, breeders evaluate breeding lines for high yield, resistance to diseases, and abiotic stresses. Plant phenotyping involves evaluating various plant attributes, such as its biophysical properties, leaf arrangement and biochemical traits, to identify critical determinants of yield and growth parameters (Johansen et al., 2019; Yang et al., 2017). Establishing good quality phenotypic data from field trials is still a constraint to enhancing the efficiency of breeding programs (Singh et al., 2016). Several recent authors have reported the shortcomings of traditional techniques of obtaining crop traits, such as yield, leaf color, aboveground biomass, and chlorophyll content (Fukano et al., 2021; Yang et al., 2017). Some drawbacks of conventional phenotyping based on manual sampling techniques are that they are not cost-effective, labor-intensive, and destructive. Moreover, the standard procedure of manual phenotyping often involves visual assessment, which can introduce subjectivity into data collection, thereby limiting accuracy and capacity (Sankaran et al., 2018). In contrast, digital phenotyping methods can consistently and rapidly acquire extensive data that are impracticable with manual measurements.

Remote sensing technologies have been employed to gather consistent non-destructive agricultural data in a near real-time manner (Chang et al., 2021; Johansen et al., 2019) in many applications. Unfortunately, satellite-based imagery products have some inherent disadvantages, such as low spatial resolution, atmospheric cloud conditions, and data acquisition frequency, for application in detailed plant or plot-level assessment of breeding trials (Chang et al., 2021). In the recent literature, Unmanned Aerial Vehicles (UAV) technologies have been proven to possess great potential to address the limitations of traditional field assessment or conventional satellite remote sensing approaches in measuring crop traits in breeding programs (Chang et al., 2017). The UAV approach offers high spectral resolution images over time and space, containing detailed canopy and other phenological features than the conventional satellite products (Sagan et al., 2019; Sidike et al., 2018). It also provides new opportunities to acquire vast, consistent, higher spatio-temporal resolution phenotyping data. UAV systems and machine learning-based high throughput phenotyping of plants are paramount in developing high precision and low-cost crop genetic programs. Recent progress in multispectral imaging and drone technology presents a cheaper platform for obtaining high-precision phenotyping data (Araus et al., 2018; Maimaitijiang et al., 2020)

Several studies reported that UAV acquired datasets such as spectral, textural, and structural features have successfully predicted various plant traits, such as grain yield, biomass, plant density, emergence, and senescence (Hassan et al., 2019; Malambo et al., 2018; Radelović et al., 2020; Roth and Streit, 2018). Recently, Chang et al. (2021) employed UAV-based high-throughput phenotyping methods to accurately estimate tomato yield, while Johansen et al. (2019) used the same approach to evaluate salinity stress tolerance in wild tomato cultivars. Many other crops such as maize, wheat, sorghum, and dry bean have been assessed successfully using drone-based high throughput phenotyping approaches for some of their phenotypic parameters. Makanza et al. (2018) utilized Red-green-blue (RGB) imagery based on UAV to appraise crop biomass and crop senescence in a maize trial and found a moderately high performance in predicting both traits. Additional examples of the applications of UAV based high throughput phenotyping include evaluating dry bean responses to drought stress and nitrogen deficit (Sankaran et al., 2018); genomic prediction modeling of sorghum plant height (Watanabe et al., 2017); and discriminating vigor of different barley genotypes (Di Gennaro et al., 2017).

Furthermore, several studies that used UAV data for soybean traits prediction and modeling are now available. Toda et al. (2021) reported proper determination of genetic variation of soybean for growth parameters using UAV-generated data for genomic prediction of soybean biomass. Fukano et al. (2021) also reported the significant beneficial effect of soybean cultivar traits on wheat yield, examining the benefits of soybean-wheat rotation based on UAV-derived vegetative indices. Maimaitijiang et al. (2020) reported a successful soybean yield estimation based on canopy spectral information obtained from UAV images from an experimental site in Columbia, Missouri, USA. Additionally, Herrero-Huerta et al. (2020) employed an array of vegetation indices (VI) and structural features to model soybean productivity. Furthermore, including GLCM features can boost prediction or classification accuracy (Iqbal et al., 2021; Kwak and Park, 2019; Räsänen and Virtanen, 2019).

Machine learning (ML) algorithms combined with UAV data have shown outstanding achievement in estimating and modeling crop traits, such as yield, biomass, and height (Herrero-Huerta et al., 2020). These methods employ advanced statistical techniques to model complex non-linear functions between spectral information and biophysical features. For example, Maimaitijiang et al. (2020) effectively used an array of regression ML models to estimate soybean grain yield in the humid climates of the USA. Equally, Eugenio et al. (2020) utilized the Multi-Layer Perceptron algorithm to predict soybean yield in the soybean growing region of Brazil and obtained satisfactory results. Moreover, Herrero-Huerta et al. (2020) accurately estimated the grain yield of the soybean trial from a site in Indiana, USA, using eXtreme Gradient Boosting (XGBoost) and Random forest regression models. However, few studies have

used ML models, UAV-derived vegetation indices, and textural features to estimate soybean yield for rapid phenotypic pipelines within African farming systems and soybean breeding programs. Thus, there is a critical and urgent need to adopt spectral vegetative indices and texture derived from UAV sensors in predicting crop phenotypes to modernize the soybean breeding program of IITA with implications to enhance the efficiency in the breeding programs of other crops and countries.

Therefore, this study aims to assess the use of UAV-derived vegetation indices (VI) and texture information from GLCM in combination with structural height to predict soybean yield as an aid to rapid high throughput phenotypic workflow.

## 2. Materials and methods

### 2.1. Description of the study area

The soybean yield trials were carried out at the International Institute of Tropical Agriculture (IITA) experimental research station plot in Ibadan (Latitude 07.5°N, Longitude 003.9°E), Oyo State, Nigeria (Fig. 1). IITA is a not-for-profit organization committed to agricultural research for food security in Africa for more than 50 years. About half of the 1000-ha research station is primarily forest, while about a third of the station comprises agricultural experiment fields. Researchers grow cassava, maize, cowpea, banana, and soybean crops for breeding or agronomic trials. The mean annual rainfall of the experimental station is 1370 mm, with the minimum and maximum temperatures of 22.1 and 31.5 °C, respectively. The soil type of the breeding trial site is mainly Ferric Luvisols, with a soil pH of between 6 and 6.5 (Oladoye, 2015). Soil texture is sandy loam with relatively low water holding capacity.

### 2.2. Experimental setup

Five soybean yield trials were conducted at the experimental site used for this study (Fig. 1), comprising two advanced and two preliminary trials during the rainy season of 2020. The advanced trials included two sets: i.e., Advanced Variety Trial-Early Set (AVTES) and Advanced Variety Trial-Medium Set (AVTMS). The preliminary tests included: Preliminary Variety Trial Set-01 (PVT01) and Preliminary Variety Trial Set-02 (PVT02). An additional variety trial called Pan African Variety Trials (PAVT) was included in the study. The AVTES trial consisted of 16 genotypes, including the standard check varieties. The genotypes in this trial were specially selected for early maturity, with days to maturity ranging between 90 and 100 days. The AVTMS consisted of 45 soybean genotypes mainly selected for medium maturity, with maturity days varying between 100 and 120 days. The PVT01 and PVT02 consisted of 50 and 60 entries combining early and medium maturing genotypes with the standard checks. The genotypes that best performed at this stage were promoted to advanced yield trials in 2021. The PAVT conducted in Nigeria in the cropping season of 2020 was composed of 45 soybean varieties that are best performing and registered in different African countries, including the standard checks recently registered in the country. The PAVT is a variety testing network of 59 public- and private sector partners from 24 countries across 113 locations, jointly coordinated by Feed the Future Soybean Innovation Lab (SIL) and the IITA soybean breeding program. Each trial used in this study was laid out in an alpha lattice design with three replications and planted with a standard soybean plot size of four rows of 4 m in length, each with inter and intra-row spacings of 50 cm and 5 cm, respectively (Fig. 2e and f). The inorganic fertilizers N, P, K, and S were applied at 20, 12, 100, and 15 kg ha<sup>-1</sup>. All the genotypes evaluated in the five trials were treated with a commercial Rhizobia inoculum called Nodumax (a product of IITA Business Incubation Platform IITA-BIP) at the recommended 100 g of inoculum

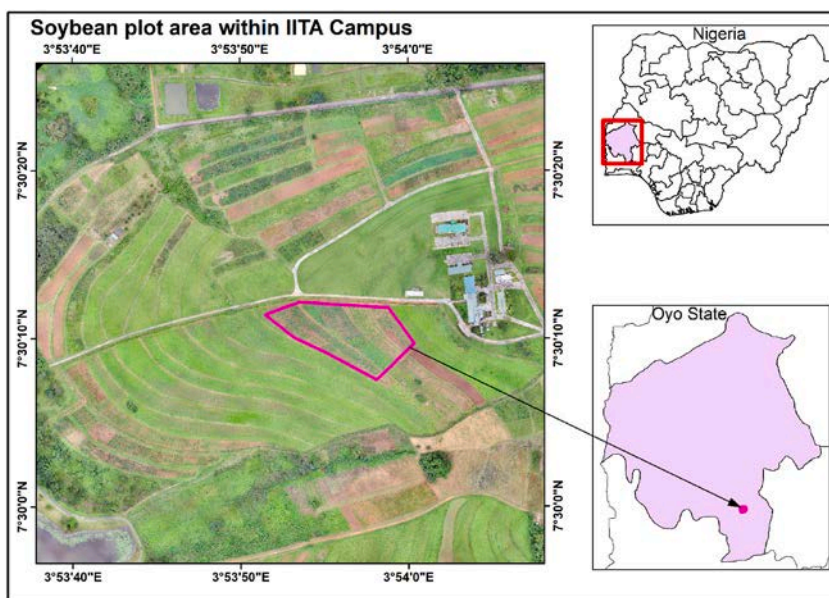


Fig. 1. UAV image showing the Soybean plot area (pink border), IITA Ibadan, Oyo State, Nigeria. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

for 10 kg of seed.

### 2.3. UAV image collection

This study used a senseFly eBee X fixed-wing drone with a 116 cm wingspan weighing 1.1 to 1.4 Kg (Fig. 2d). The drone flight missions had aboard a Parrot Sequoia multispectral camera integrated with an RGB camera (Fig. 2c). The Sequoia sensor is a self-calibrating radiometric system with four bands (near-infrared (NIR): 770–810 nm; red: 640–680 nm; green: 530–570 nm; red edge: 730–740 nm). Furthermore, it uses a sunshine sensor that synchronizes brightness values with the Inertial Measurement Unit (IMU) and onboard GPS. In addition, we employed a senseFly Geobase station (Fig. 2b) that enables high-precision positioning systems during every flight. The GeoBase system helped achieve an accuracy of about 3 cm without Ground Control Points (GCPs).

Flight planning was designed with the Emotion Software. The flight missions were performed twice during the growing stages of soybean on 21 September (about 60 days after planting (DAP) at the start of flowering (R1 stage) and 27 October 2020 (95 DAP, at the commencement of pod setting (R6 stage)). The drone flight took place between 11:00 a.m. and 2:00 p.m. local time in clear weather conditions. The flight parameters were set to lateral overlap at 60% during the flight operations, while longitudinal overlap was at 80%, as recommended for optimal UAV image overlap (Böhler et al., 2018).

### 2.4. UAV image processing

After each flight, Emotion Software was used to carry out post-flight processing. Pix4D mapper version 4.6.4 was used to generate orthomosaic images. Capturing accurate geo-referenced information of multispectral images is a significant advancement incorporated into the Parrot Sequoia camera. The Pix4D processed the orthomosaic images through geo-referencing, camera alignment, dense point cloud development, digital surface model (DSM), and digital terrain model (DTM) generation (Maimaitijiang et al., 2020). Moreover, to produce accurate orthomosaic images, the Pix4D mapper automatically computed GCPs by matching tie points positions of images. Furthermore, it performed radiometric calibration using vignetting correction, lens distortion information, and the irradiance values obtained by the sunshine sensor during the flight. All multispectral reflectance bands were generated at a spatial resolution of 12 cm and RGB mosaic images at 2.6 cm.

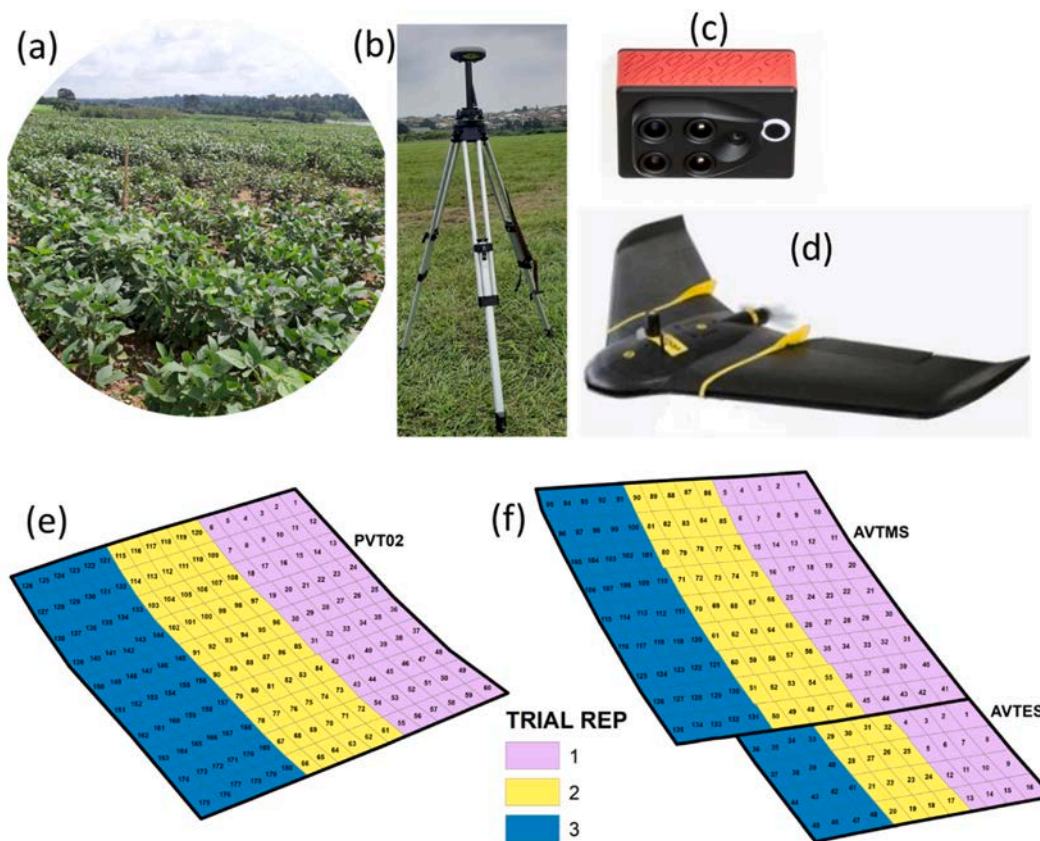


Fig. 2. (a) Soybean plots, (b) Sensefly Geobase, (c) Parrot Sequoia camera, (d) Sensefly eBee X fixed-wing drone, (e & f) PVT02, AVTMS, and AVTES field layout showing plot no. and replication.



## 2.5. Texture and vegetation indices extraction

### 2.5.1. Canopy texture information

Many crop classification and trait predictions utilize textural features as input layers using machine learning models. These features help reduce the pixel noise effect and give valuable information about a spatial object's architectural arrangement, color, and intensities (Iqbal et al., 2021). Gray level co-occurrence matrix (GLCM), proposed by Haralick et al. (1973), is a well-known statistical technique for handling remotely sensed data to classify landcover and model vegetation structure (Kwak and Park, 2019). Haralick et al. (1973) identified fourteen texture information from an image object, of which some correlate, indicating redundant information. Hence in this study, seven commonly used GLCM features were considered: (1) mean, (2) standard deviation, (3) homogeneity, (4) dissimilarity, (5) entropy, (6) angular second moment, and (7) variance. The seven texture parameters used in the present work are well described in Haralick et al. (1973) and Kwak and Park (2019). As implemented in R Software, we used the "glcm" package to calculate the GLCM metrics (Zvoleff, 2020). Texture parameters obtained from NIR, Rededge, Red, and Green reflectance bands were the input variables for soybean yield prediction.

### 2.5.2. Spectral vegetative indices

In recent literature, vegetative indices (VIs) have been used in UAV-based crop phenotyping workflow, apart from canopy texture information (Randelović et al., 2020). These VIs are indicators for the rate of photosynthesis, level of chlorophyll, leaf area index (LAI), and green biomass. They are also commonly used to assess growth parameters and crop yield (Hassan et al., 2019). Twenty VIs were used in this study and are described briefly in Table 1. For the computation of VIs, we used the function "spectralIndices" from the RStoolbox package within R software (Leutner et al., 2019; Suab and Avtar, 2020).

### 2.5.3. Plot polygon generation and spectral feature extraction

Plot boundary polygons were generated using on-the screen digitizing from the UAV-RGB (Fig. 3). Pixel values of spectral bands,

**Table 1**  
Description of vegetation indices.

Index	Description	Formula	Reference
CTVI	Corrected Transformed Vegetation Index	$(NDVI + 0.5) / \sqrt{ABS(NDVI + 0.5)}$	Perry & Lautenschlager (1984)
DVI	Difference Vegetation Index	$SL * NR - RD$	Richardson & Weigand (1977)
GEMI	Global Environmental Monitoring Index	$((NR^2 - RD^2) * 2 + (NR * 1.5) + (RD * 0.5)) / (NR + RD + 0.5) * (1 - (((NR^2 - RD^2) * 2 + (NR * 1.5) + (RD * 0.5)) / (NR + RD + 0.5) * 0.25)) - ((RD - 0.125) / (1 - red))$	Pinty & Verstraete (1992)
GCVI	Green Chlorophyll Vegetation Index	$(NR/GR)-1$	Gitelson et al. (2005)
GNDVI	Green Normalized Difference Vegetation Index	$(NR - GR)/(NR + GR)$	Gitelson & Merzlyak (1998)
MCARI	Modified Chlorophyll Absorption Ratio Index	$((RE - RD) - (RE - GR)) * (RE/RD)$	Daughtry et al. (2000)
MSAVI	Modified Soil Adjusted Vegetation Index	$NR + 0.5 - (0.5 * \sqrt{((2 * NR + 1)^2 - 8 * (NR - (2 * RD)))})$	Qi et al. (1994)
MSAVI2	Modified Soil Adjusted Vegetation Index 2	$(2 * (NR + 1) * \sqrt{((2 * NR + 1)^2 - 8 * (NR - (2 * RD)))}) / 2$	Qi et al. (1994)
NDRE	Normalized Difference Red Edge Index	$(NR - RE)/(NR + RE)$	Barnes et al. (2000)
NDVI	Normalized Difference Vegetation Index	$(NR - RD)/(NR + RD)$	Rouse et al. (1974)
NDWI	Normalized Difference Water Index	$(GR - NR)/(GR + NR)$	McFEETERS (2007)
NRVI	Normalized Ratio Vegetation Index	$(RD/NR - 1)/(RD/NR + 1)$	Baret & Guyot (1991)
RECI	Red edge chlorophyll index	$(NR/RE)-1$	Gitelson et al. (2003)
RVI	Ratio Vegetation Index	$RD/NR$	Jordan (1969)
SAVI	Soil Adjusted Vegetation Index	$(NR - RD) * (1 + SB)/(NR + RD + SB)$	Huete (1988)
SR	Simple Ratio Vegetation Index	$NR/RD$	Birth & McVey (1968)
TTVI	Thiam's Transformed Vegetation Index	$\sqrt{ABS((NR - RD)/(NR + RD) + 0.5)}$	Thiam (1997)
TVI	Transformed Vegetation Index	$\sqrt{((NR - RD)/(NR + RD) + 0.5)}$	Deering et al. (1975)
WDVI	Weighted Difference Vegetation Index	$NR - SL * RD$	Richardson & Weigand (1977)
WDRVI	Wide Dynamic Range Vegetation Index,	$(SL * NR - RD)/(SL * NR + RD)$	Gitelson (2004)

NR = Near Infrared, GR = Green, RD = Red, RE = Rededge, SL = slope of the soil line, SB =Soil brightness factor, ABS = absolute.

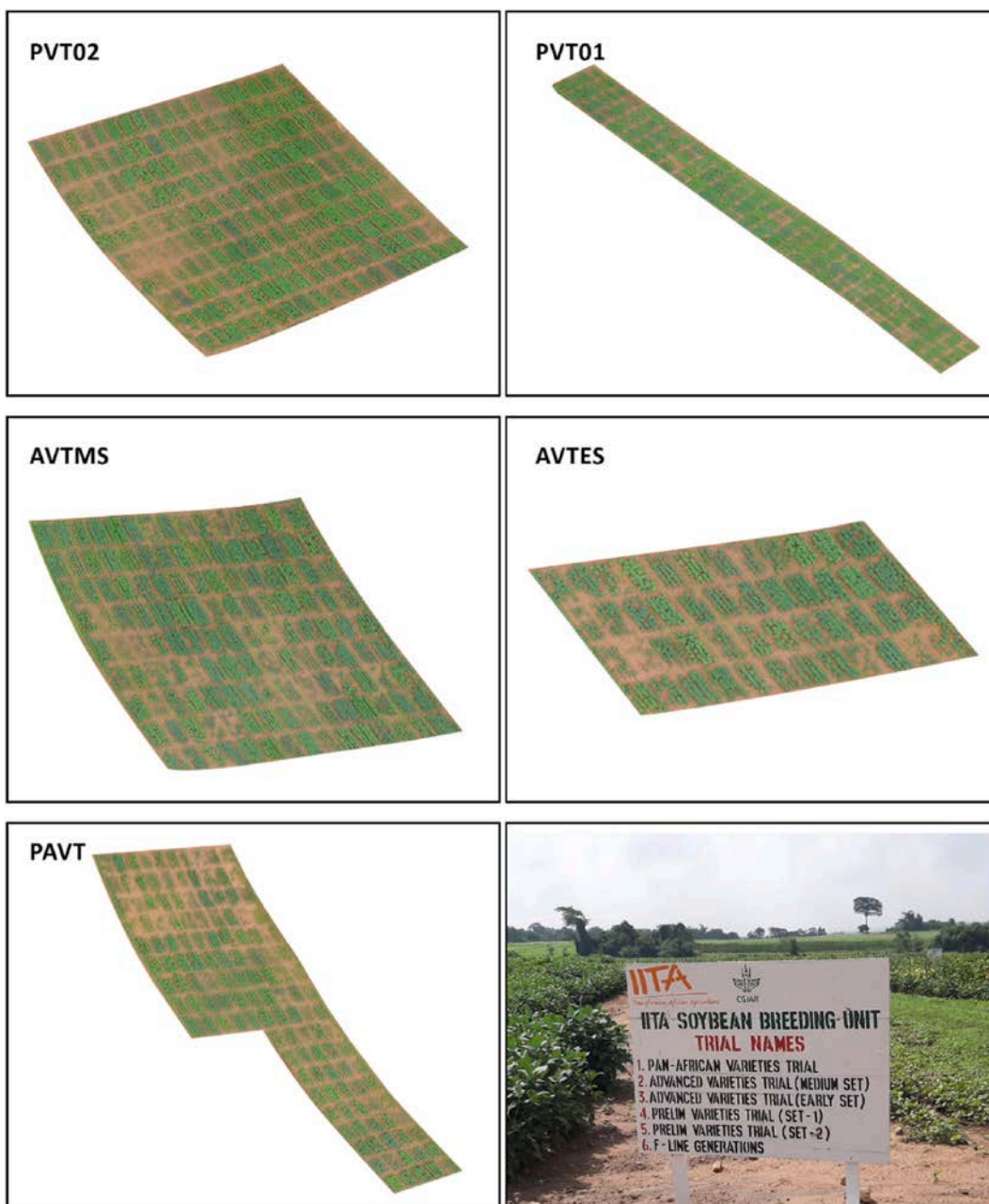


Fig. 3. Plot images of the five soybean trials acquired by the UAV on 21 September 2020.

vegetative indices, canopy height, and textural features were extracted from the plot polygons. To avoid the border effects caused by weeds or soil around each plot, we created a negative buffer of 0.7m away from the plot boundary using the ArcGIS 10.7 buffer tool (Fig. 4). Pixel values for testing and training were then obtained within the polygon buffer.

### 2.6. Predictive model description

Five predictive machine learning models were used to estimate soybean yield from five advanced breeding trials. A brief description of these machine learning models is presented in the following sections.

#### 2.6.1. Cubist

Cubist is a regression procedure commonly used for prediction (Quinlan and Quinlan, 1992). Primarily it functions by forming a tree structure and converts each path of the tree into a rule. Each rule is transformed into a regression model using the data subset specified by the rules. The different rules are trimmed or possibly merged, and the promising variables for the linear regression

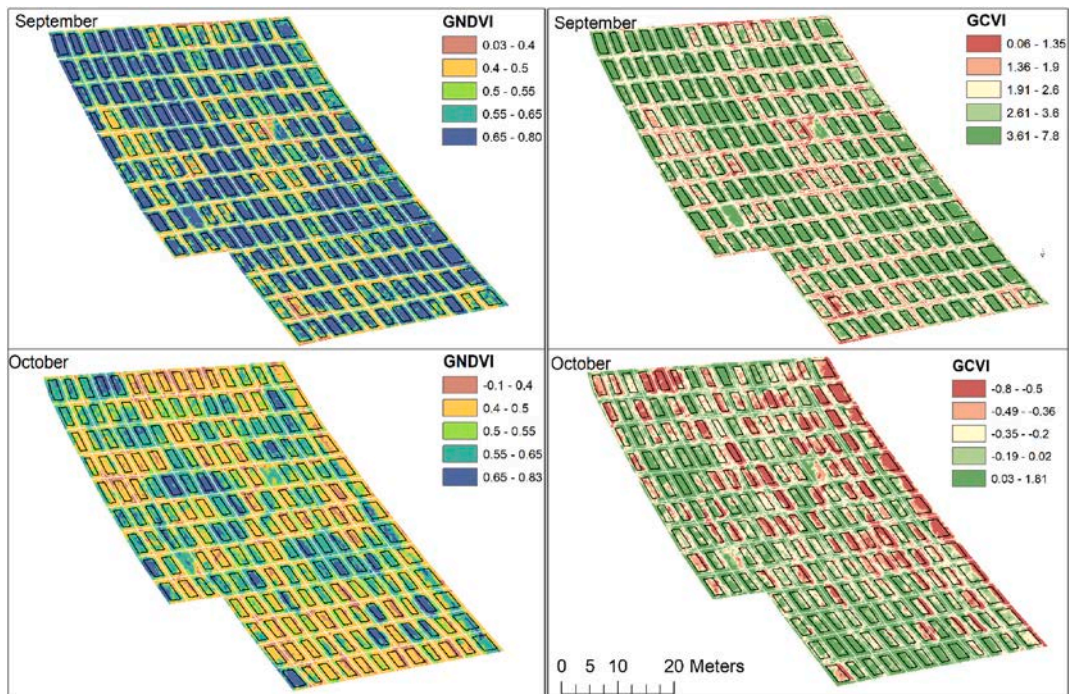


Fig. 4. Plot layout showing GNDVI and GCVI for September and October. White polygons show individual plot boundaries, while black polygons are 70 cm buffers within each plot to exclude borders in feature extraction.

procedure are the dependent variables used in the parts of the rule that were trimmed out. Tuning this model requires two parameters, namely committees and neighbors. The R package *caret* (*caret::train()* function) has bindings to find the optimal values of the parameters.

### 2.6.2. Random forest (RF)

RF is a commonly employed ML model for classification and regression initially introduced by Breiman (2001). As an ensemble ML algorithm, it constructs several independent decision trees for model fitting. The tree or model prediction with the maximum votes becomes the outcome for a particular class or value. A certain number of input variables are selected randomly from each node, and the subsets are used to calculate the best model output (Herrero-Huerta et al., 2020).

### 2.6.3. Stochastic gradient boosting (GBM)

GBM is a widely used ML technique successfully implemented across many areas of modeling studies and involves resampling observations and features in each round (Freeman et al., 2015). RF constructs an ensemble of deep independent trees, while GBM creates an ensemble of shallow sequential trees with each tree using the previous model estimations for improvement. The combination of these numerous weak consecutive predictions often produces a robust model. The main principle of stochastic gradient boosting is the sequential addition of new models to the ensemble. A different weak, base-learner model is trained to minimize the entire ensemble's error at each iteration. A simple GBM model contains boosting and tree-specific hyperparameters. There are two boosting hyperparameters, i.e., *learning rate* and the *number of trees*. On the other hand, tree-specific hyperparameters include the *minimum number of observations in terminal nodes* and *tree depth*.

### 2.6.4. Extreme Gradient Boosting (XGBoost)

XGBoost is an open-source ML model that efficiently implements gradient boosting decision trees (Chen and Guestrin, 2016). XGBoost accelerates the boosted tree construction using parallelization and utilizes a more efficient tree searching procedure. The weight of each input to the trained model is considered when boosted trees are built to obtain accurate feature scores. Some hyperparameters have to be tuned to develop a robust XGBoost model. They include *learning rate*, which helps shrink the weight at each iteration and *regularisation alpha*, a parameter to prevent overfitting. Other essential parameters include the *maximum tree depth*, *subsample ratio* that indicates the proportion of observations used for each iteration to the total training samples, and *colsample\_bytree* denoting the percentage of features sampled randomly at each step.

### 2.6.5. Support vector machines (SVM)

SVM are supervised learning techniques based on vectors separated by hyperplanes or decision boundaries. SVMs are founded on statistical learning theory and are very powerful for complex classification or regression problems (Cortes and Vapnik, 1995). SVMs are highly flexible, using various kernel functions and can estimate complex non-linear decision boundaries. The present work uses the

highly flexible radial basis kernel to implement SVM. Two optimization parameters, namely: Sigma ( $\sigma$ ) and cost (C), were tuned using the *train* function of the caret package for prediction in this study.

### 2.7. Feature integration for yield prediction

Diverse feature integration techniques have been explored to explain the influence of various types of data fusion on the performance of ML prediction models on crop traits estimation (Maimaitijiang et al., 2020). Additionally, several authors have substantiated the capability of combined spectral and structure data in crop yield prediction (Stanton et al., 2017; Bendig et al., 2015). Hence, this study explored the prediction model accuracies on the three variants of datasets where spectral bands (GREEN, RED, NIR, and REDEDGE) and canopy height were common to the different versions of datasets. The first consisted of GLCM features coupled with spectral bands (SP) and canopy height (CH), hereafter called the GLCM model. The second was based on integrating vegetation indices (VI), spectral bands (GREEN, RED, NIR, and REDEDGE), and canopy height, subsequently called the VI model. The third was the integration of GLCM and VI datasets resulting in the GLCM + VI model.

The two UAV flights on 21 September and 27 October 2020, corresponding to R1 and R6 of the soybean growth stages, respectively, generated 106 predictors. Eight of these parameters were from the multispectral bands of the two dates (GREEN, NIR, RED, REDEDGE), while 40 vegetation indices were obtained from the reflectance bands for the two dates (Table 1). The GLCM-derived variables formed 56 input layers to the models. The last two parameters came from the canopy height derived from the point cloud analysis during the photogrammetric workflow in Pix4D software.

### 2.8. Development of predictive models

All the regression ML models employed in the study have different hyperparameters for optimal performance. The k-fold cross-validation technique is an excellent approach for hyperparameter tuning, proven for its superiority in estimating prediction error (Eugenio et al., 2020; Fushiki, 2009). This approach also prevents the overfitting of predictive models. Hence, the study's overall predictive model development process involved 10-fold cross-validation repeated three times and implemented using the *caret::train* function. Modeling data is often divided into training and testing subsets of the entire data to avoid overfitting problems. Recent studies in ML predictive modeling used various ratios such as 80/20, 60/40, 70/30, and 50/50 to subset field data into train and test datasets. Nguyen et al. (2021) evaluated different ML techniques with several ratios and reported that 70/30 gave the best result. Thus, following the recommended ratio of data splitting, the field data in this study was divided into a 70/30 ratio for training and testing samples, respectively.

### 2.9. Predictive model performance evaluation

Four commonly used statistical indicators of accuracy, i.e., the coefficient of determination ( $R^2$ ), root mean square error (RMSE), mean absolute error (MAE), and normalized RMSE (NRMSE%) were computed to assess quantitatively the performance of all the yield prediction models. NRMSE% was calculated by dividing RMSE by the measured soybean grain yield range. The computation of these metrics is expressed in the following equations, where  $E_i$  and  $G_i$  are the predicted and observed grain yield, respectively.

$$R^2 = \left( \frac{\sum_{i=1}^n (G_i - \bar{G})(E_i - \bar{E})}{\sqrt{\sum_{i=1}^n (G_i - \bar{G})^2} \sqrt{\sum_{i=1}^n (E_i - \bar{E})^2}} \right)^2 \quad (1)$$

$$MAE = \frac{\sum_{i=1}^n |G_i - E_i|}{n} \quad (2)$$

$$RMSE = \left[ \frac{\sum_{i=1}^n (G_i - E_i)^2}{n} \right]^{1/2} \quad (3)$$

$$NRMSE\% = \frac{RMSE}{y_{max} - y_{min}} * 100 \quad (4)$$

### 2.10. Statistical analysis

A descriptive summary for each soybean trial was analyzed using *tidyverse* and *moments* libraries of R-software. The statistical distribution of soybean yield in the variety trials was visualized using a boxplot, and *multcompBoxplot* and the *Tukey HSD* function were used to estimate statistical significance using R statistical software version 4.1.

### 2.11. Feature importance computation

We performed a feature importance analysis that identifies significant input variables contributing to soybean grain yield prediction using the Boruta package in R (Kursa et al., 2021). The Boruta package, an advanced variable selection technique for finding all relevant features, is one of the most accurate and robust feature selection methods (Degenhardt et al., 2019; Sanchez-Pinto et al., 2018; Speiser et al., 2019). The technique is built around the random forest regression model. It removes the statistically proven irrelevant features (Kursa and Rudnicki, 2010).



### 3. Results

This study employed five ML models, i.e., RF, Cubist, SVM, GBM, and XGBoost, for soybean yield prediction using extracted canopy structural features (height), vegetation indices, and textural elements derived from multispectral UAV imagery.

#### 3.1. Performance among regression models for soybean grain yield prediction

Mean quantitative measures of model performance in soybean grain yield prediction are shown in Fig. 5. Considering all the performance metrics ( $R^2$ , RMSE, NRMSE, and MAE), Cubist exhibited the best prediction accuracy and was followed closely by the Random Forest (RF) across the three data sets viz., GLCM, VI, and GLCM + VI. The Extreme Gradient Boosting (XGBoost), across all datasets, was the third-best performing model in soybean grain yield predictive ability. Stochastic Gradient Boosting (GBM) was the least accurate estimate of soybean grain yield among the five models.

The GLCM dataset exhibited the highest  $R^2$  value along with the RF and Cubist models in the soybean yield prediction (Fig. 5). In contrast, no noticeable differences were found in RMSE values using GLCM and GLCM + VI datasets for the other three models (GBM, SVM, and XGBoost).

Furthermore, the Cubist model's superior performance is presented in Table 2, where its  $R^2$  varied from 0.73 to 0.89 for the GLCM dataset across the five trials. The RF algorithm closely followed this, with  $R^2$  ranging from 0.62 to 0.89. Moreover, these two models' superior performance trends are displayed in Table 2; NRMSE varied from 6.2 to 10.7% for the Cubist and 7.1–13% for the RF model.

#### 3.2. Influence of feature integration on soybean yield prediction

To understand the influence of different configurations of feature types on the prediction power of the ML models, we trained and tested the prediction algorithms with three subsets of datasets, namely, texture information (GLCM), vegetation indices (VI), and integration of both datasets (GLCM + VI). As depicted in Fig. 5 and Table 2, the model based on the texture information (GLCM) offered the most accurate soybean yield prediction, slightly outperforming the model based on the combined datasets (GLCM + VI). Using all performance metrics ( $R^2$ , RMSE, MAE, and NRMSE), the VI-based model displayed the lowest accuracy in estimating yield. The mean  $R^2$  obtained using the VI-based model ranged from 0.44 to 0.66, while those observed using the GLCM model ranged from 0.55 to 0.73. Equally, the percentage error of grain yield estimation (NRMSE%) from the VI-based model (8.5–18.7%) was consistently higher than those achieved with the GLCM model (7.8–14.5%) (Table 2). Integrating both VI and GLCM data did not significantly increase the performance metrics, probably due to the critical input variables such as canopy height and spectral bands common to both models.

#### 3.3. Yield prediction for the different preliminary and advanced soybean variety trials

The prediction accuracy varied across trials by the different yield estimation ML models. There exist minor differences between the two best performing dataset configurations, as can be seen from the mean  $R^2$  and NRMSE (Table 2). Hence, to elucidate the prediction patterns across the different trials, we concentrate on the GLCM + VI model. The models predicted that soybean grain yield in the PVT01 trial achieved the lowest mean  $R^2$  of 0.50. However, the most accurately predicted trial was PVT02, judged by the NRMSE values of 8% compared to 15% obtained for the AVTES. Although a slightly lower mean  $R^2$  (0.7) was achieved under PVT02 relative to 0.72 for the AVTES, the NRMSE of 8.0% for PVT02 was much better than the 15% obtained for AVTES. Overall, the ML models' performance for soybean yield estimation for the studied five trials compares consistently well and, in most cases, is not significantly

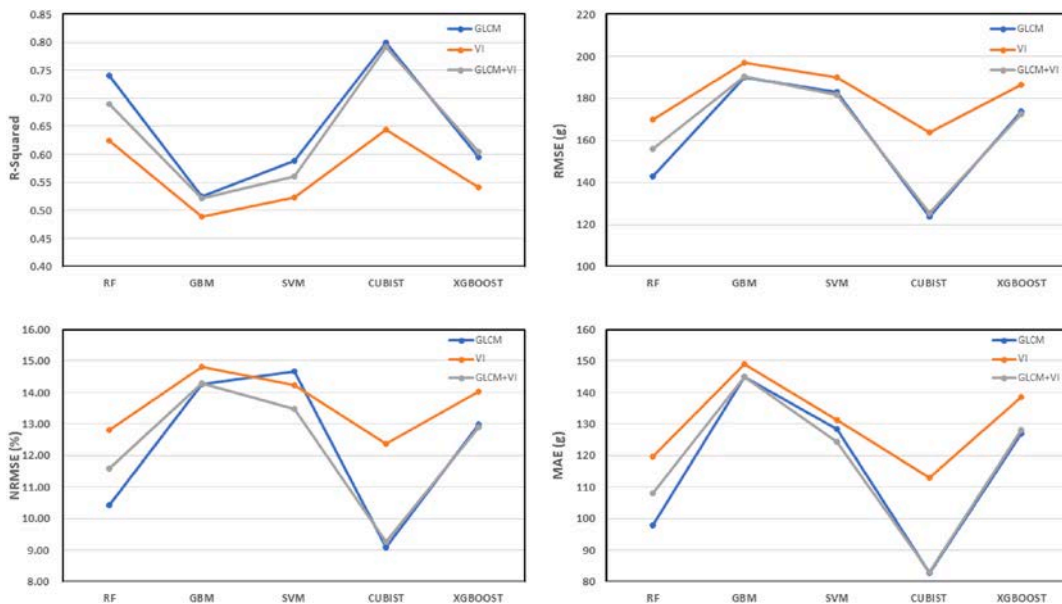


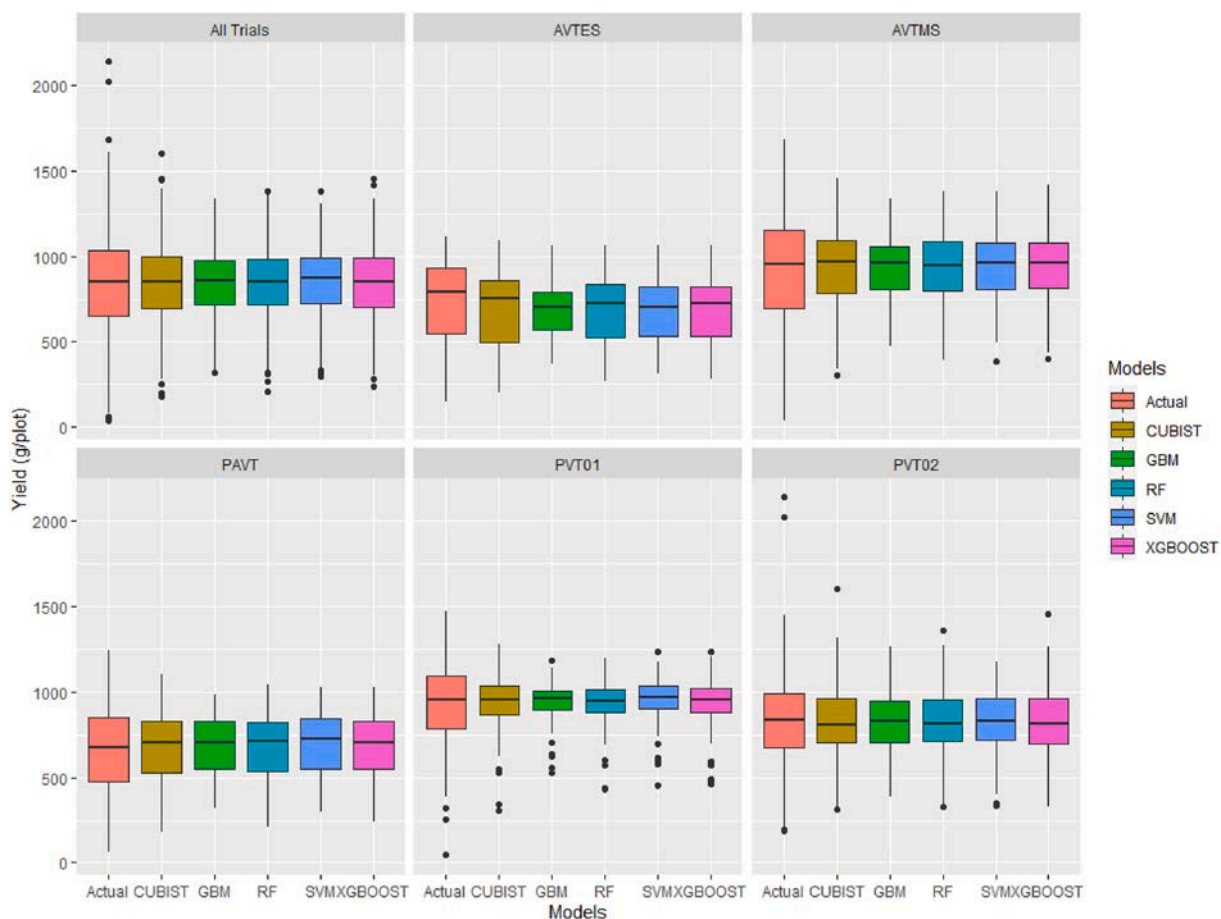
Fig. 5. Mean of model performance metrics ( $R^2$ , RMSE (g/plot), NRMSE (%), and MAE (g/plot) across the five experiments.

**Table 2**  
Performance metrics of the five ML models using GLCM, VI, and GLCM + VI dataset for five soybean trials.

	R <sup>2</sup>					NRMSE (%)					
	MODEL	AVTES	AVTMS	PAVT	PVT01	PVT02	AVTES	AVTMS	PAVT	PVT01	PVT02
GLCM	RF	0.89	0.62	0.77	0.67	0.77	10.2	13.1	11.1	10.5	7.1
	GBM	0.58	0.44	0.61	0.38	0.62	18.7	15.5	14.4	13.8	8.9
	SVM	0.64	0.44	0.61	0.64	0.62	17.4	15.5	14.2	17.4	9.0
	Cubist	0.89	0.74	0.83	0.73	0.81	9.9	10.7	9.5	9.1	6.3
	XGBOOST	0.68	0.50	0.69	0.39	0.71	16.2	14.6	12.6	13.7	7.9
	<b>Mean</b>	<b>0.73</b>	<b>0.55</b>	<b>0.70</b>	<b>0.56</b>	<b>0.71</b>	<b>14.5</b>	<b>13.9</b>	<b>12.3</b>	<b>12.9</b>	<b>7.8</b>
VI	RF	0.62	0.56	0.71	0.53	0.71	17.9	13.9	12.3	12.2	7.9
	GBM	0.52	0.41	0.58	0.35	0.59	20.0	15.9	14.8	14.1	9.3
	SVM	0.57	0.44	0.62	0.37	0.61	18.9	15.4	13.9	13.9	9.1
	Cubist	0.63	0.58	0.74	0.54	0.73	17.6	13.3	11.7	11.8	7.5
	XGBOOST	0.56	0.46	0.63	0.42	0.64	19.2	15.2	13.7	13.3	8.8
	<b>Mean</b>	<b>0.58</b>	<b>0.49</b>	<b>0.66</b>	<b>0.44</b>	<b>0.66</b>	<b>18.7</b>	<b>14.7</b>	<b>13.3</b>	<b>13.1</b>	<b>8.5</b>
GLCM + VI	RF	0.77	0.60	0.75	0.59	0.74	14.1	13.3	11.4	11.5	7.56
	GBM	0.59	0.43	0.59	0.38	0.62	18.4	15.6	14.6	13.8	9.03
	SVM	0.70	0.46	0.62	0.40	0.63	15.7	15.2	14.0	13.6	8.92
	Cubist	0.87	0.74	0.83	0.70	0.82	10.3	10.7	9.5	9.6	6.22
	XGBOOST	0.68	0.52	0.69	0.46	0.68	16.3	14.3	12.7	12.9	8.22
	<b>Mean</b>	<b>0.72</b>	<b>0.55</b>	<b>0.70</b>	<b>0.50</b>	<b>0.70</b>	<b>15.0</b>	<b>13.8</b>	<b>12.4</b>	<b>12.3</b>	<b>8.0</b>

different (Table 2).

ANOVA was performed for the different soybean trials. The means were separated with *HSDTukey* tests (at *P* < 0.05 level of significance) to evaluate further the predicted soybean grain yield difference produced with the combined datasets (GLCM + VI). The



**Fig. 6.** Boxplot showing the distribution of measured soybean grain yield and ML models predicted yields for the five trials.

boxplot (Fig. 6) of the variation among the estimated soybean grain yield by the five ML algorithms revealed that the output of each model was not statistically different. The pattern of the predicted soybean yields was also following the distribution of the observed grain yield (Fig. 6). These findings suggest that all ML models effectively detected grain yield differences among the five varietal evaluations. The boxplot further revealed slight differences in grain yield distribution and patterns measured in the five experiments'. There were no outliers within the distribution of the predicted and actual yields of the PAVT and AVTES and a few outliers for the AVTMS (Fig. 6). As displayed in the box plots, the performance of genotypes for grain yield was higher in AVTMS, PVT01, and PVT02s than in AVTES and PAVT.

The scatterplots of yield estimates developed from the different regression models were compared with the corresponding actual values measured in the variety trials (Fig. 7). The distribution of the scatterplots for the various experiments was similar. The coefficient of determination ( $R^2$ ) obtained from Cubist, the best performing model, is shown on the graph, ranging from 0.77 to 0.84, suggesting good predicting power by the models. As can be noticed from the scatterplots, most models overestimated the actual grain yield at low measured values; while underestimating them at high actual yields. The extent of this discrepancy differed with the different experiments. For instance, in AVTES, the extremely low yields were more accurately predicted by Cubist, RF, and XGBoost. A similar trend of performance can be seen in PAVT and PVT01.

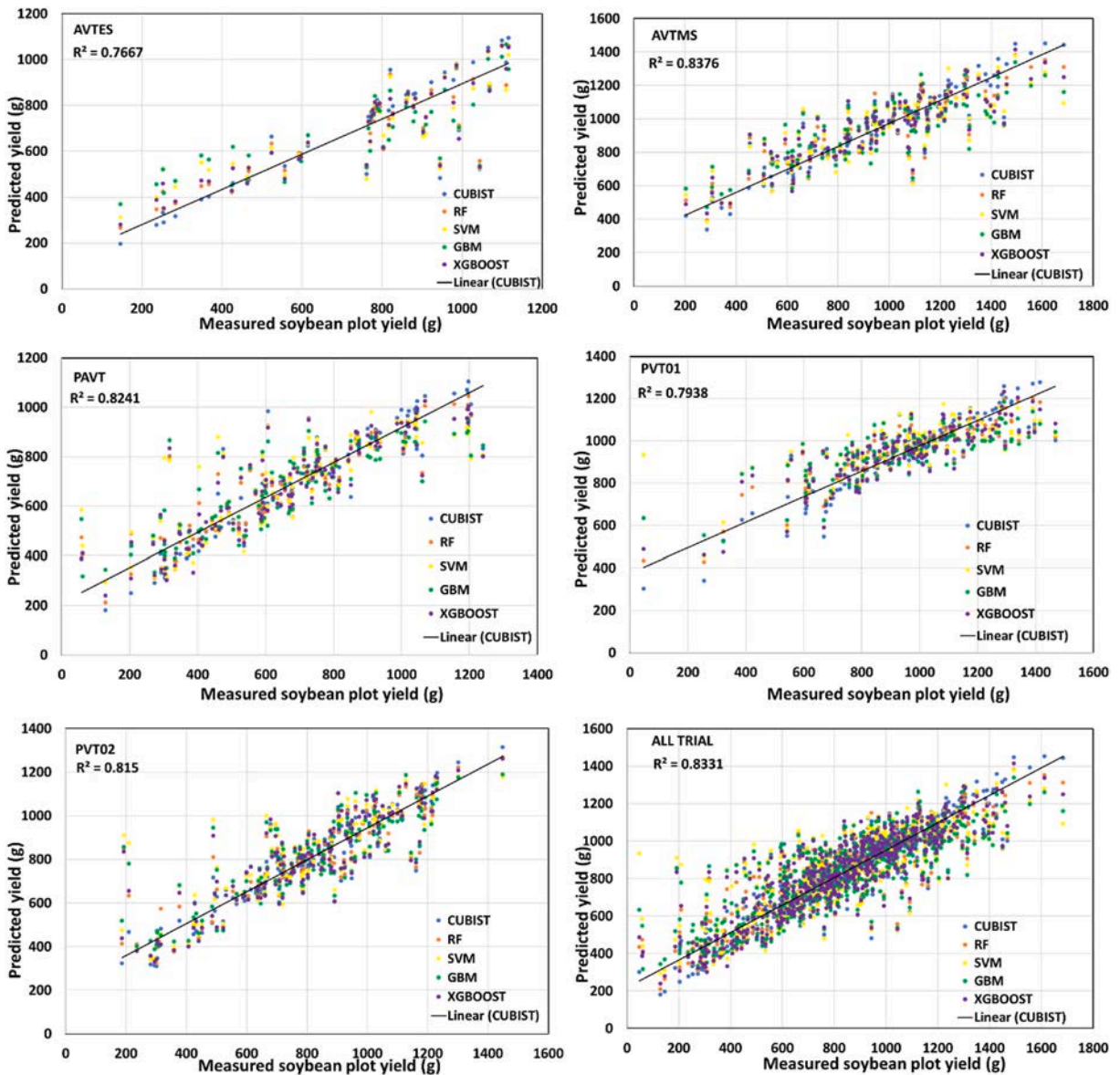


Fig. 7. Comparison of the five models measured against predicted soybean plot yield for the breeding experiments. The best fit line was obtained from the Cubist model.



3.4. Visual assessment of the spatial variation of soybean yield prediction map

Effective grain yield prediction models must capture spatial variabilities caused by changes in terrain, irrigation, soil fertility, irrigation, and other environmental factors influencing crop development (Rischbeck et al., 2016). Hence, predicted yield maps were compared with the measured grain yield map for the two trials, as shown in Fig. 8. At the left uppermost corner, the measured plot yield map showed spatial variability of soybean grain yield for the two advanced variety trials (AVTMS and AVTES). The AVTES trial consisted of 16 genotypes, including standard check varieties, while the AVTMS consisted of 45 soybean genotypes. As seen in Fig. 8, the five models captured the spatial variation of grain yield shown for the measured plots. The models successfully delineated fields with high, medium, and low grain yields for AVTMS and AVTES. This result indicates that the five regression models were adaptable over space and accurately detected the within-field heterogeneity.

3.5. Impact of input features on soybean yield prediction

The impact of different predictor variables on yield prediction by the Random Forest algorithm is presented in Fig. 9. Variable importance obtained from the models that used vegetation indices coupled with canopy height and spectral bands (VI model) is presented in Fig. 9a. Fig. 9b shows the feature importance obtained from the model that utilized textural data combined with canopy height and spectral bands (GLCM model). Fig. 9a and b showed that the canopy height of the two soybean phenological stages

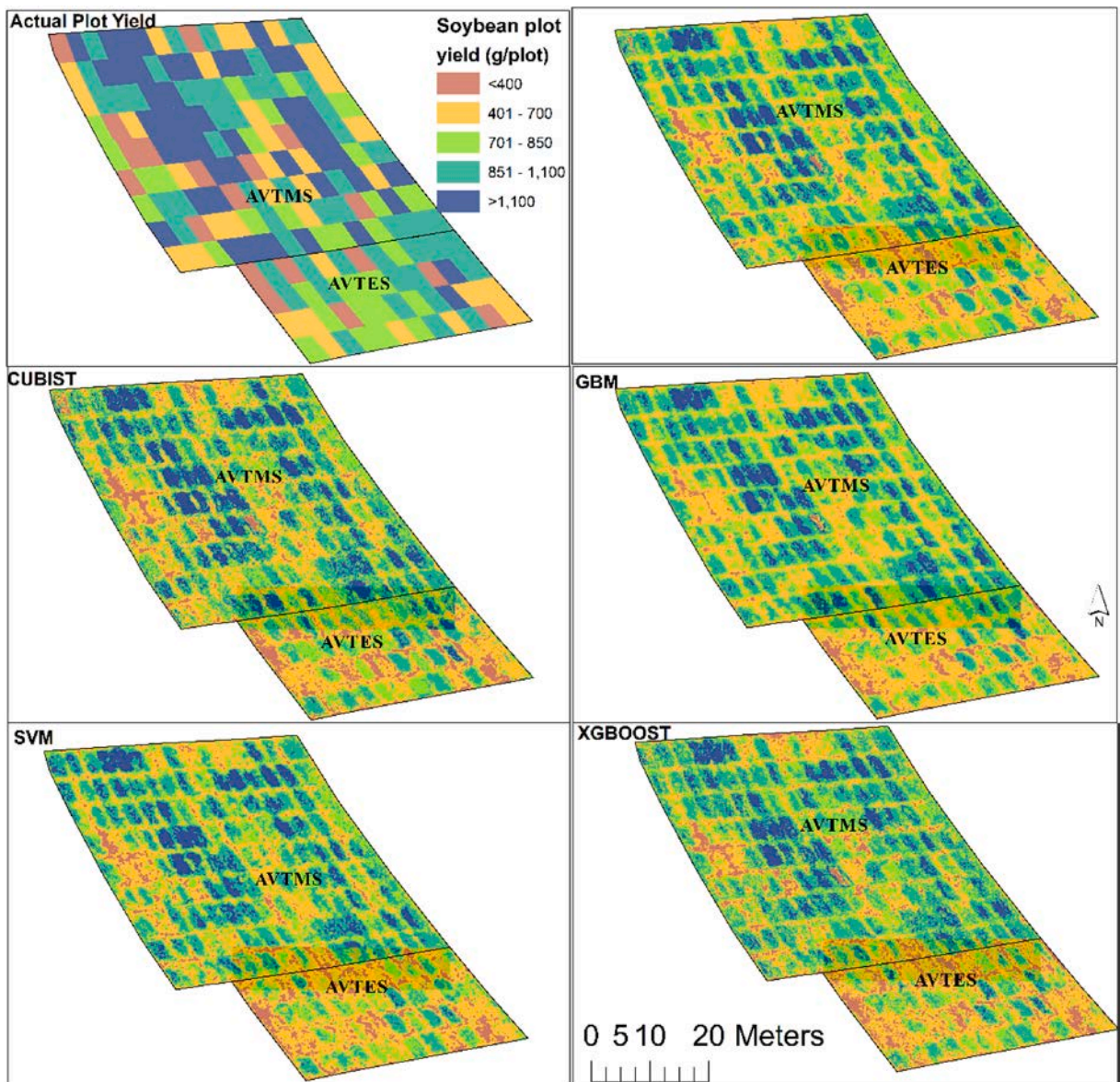


Fig. 8. Comparison of actual soybean plot yield against prediction maps by five ML models (Cubist, RF, SVM and XGBOOST) for two Trials (AVTES and AVTMS).



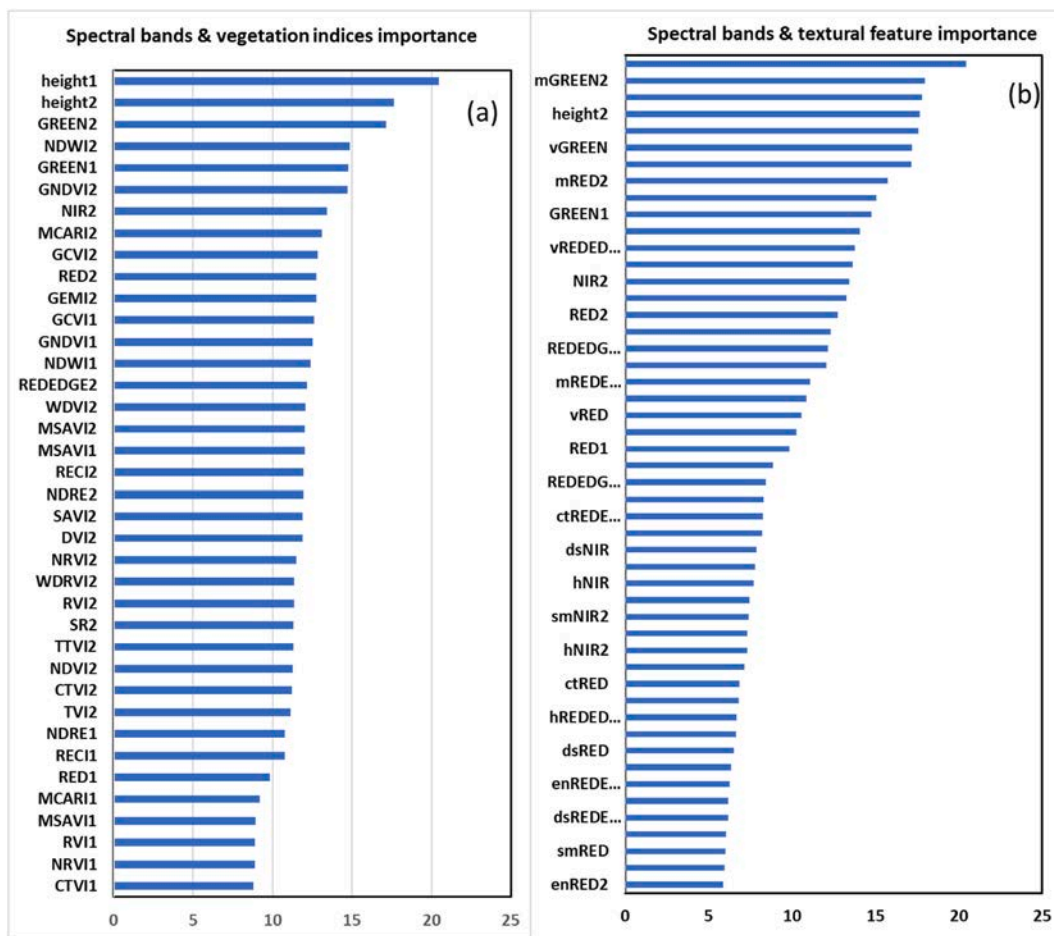


Fig. 9. (a) The importance of spectral bands, canopy height and vegetation indices; (b) spectral bands, canopy height, and texture attributes based on the GLCM on soybean yield prediction. Features obtained in September (soybean R1 stage) are distinguished with “1” added to their names, while “2” was added to October features (Soybean R6 stage). Only about 75% of the predictors are shown in both cases. The GLCM features (Fig. 9b) are shortened (e.g. mGreen = mean of Green, vGreen = variance of Green, enGreen = entropy of Green, dsGreen = dissimilarity of Green, ctGreen = contrast of Green, hGreen = homogeneity of Green, smGreen = second moment of Green). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

appeared highly influential in soybean yield prediction. They ranked 1st and 2nd among the predictors of the VI model. They ranked high 1st and 4th among the predictor variables’ importance obtained with the GLCM model. Spectral bands (GREEN2, GREEN1, NIR2) were among the most critical predictors of soybean grain yield as they occurred 3rd, 5th and 7th in order of importance (Fig. 9a).

Consequently, the VIs based on GREEN and NIR bands (NDWI, GNDVI, GCVI) were prominently influential among input variables to the VI model. Surprisingly, NDVI, the most widely applied index in vegetation and crop prediction studies, did not rank among the best 20 estimators of soybean grain yield.

The variable importance plots (Fig. 9a and b) also reveal the special significance of estimators collected during the later growth stage of soybean (R6). For example, the four most influential VIs (NDWI2, GNDVI2, MCARI2, and GCVI2) displayed in Fig. 9a were based on the UAV data collected in October (R6 growth stage), signifying that the later soybean reproductive phase is crucial for yield modeling. Furthermore, among the ten most significant textural features (mGreen2, vGreen2, Green2, vRed2, mRed2, mNIR2, vRededge2) influencing grain yield detection (Fig. 9b), seven of them were from the data collected during the soybean R6 stage in October, further confirming the great utility of variables collected at that soybean reproductive phase.

#### 4. Discussion

Soybean grain yield prediction is of paramount importance as it can inform policy on agriculture, climate change adaptation, and understanding of crop phenotyping. Several researchers have demonstrated the utility of ML models combined with UAV-based imagery to estimate crop grain yields. This study used five ML models for quantitative soybean grain yield prediction. Generally, our results show that the five ML regressions models (RF, Cubist, XGBoost, GBM, and SVM) can estimate soybean yields accurately. However, among the five ML models, the performance of Cubist was the best, closely followed by the RF models. Some regression studies show that the Cubist model outperformed RF. For instance, Zhou et al. (2019) observed a superior performance of the Cubist

models over the RF, while [Noi et al. \(2017\)](#) recommended both the Cubist and RF models for predicting air temperature over other regression methods. [Wang et al. \(2016\)](#) similarly reported the complementary performance of both Cubist and RF algorithms. Overall, the  $R^2$  obtained for soybean grain yield prediction in this study was similar to recent studies ([Eugenio et al., 2020](#); [Herrero-Huerta et al., 2020](#); [Randelović et al., 2020](#)). [Maimaitijiang et al. \(2020\)](#) employed some ML regression models in their study to predict soybean yield. They reported that the Deep Neural Network (DNN) and RF were outstanding in predicting soybean yield under irrigated and non-irrigated conditions. They reported  $R^2$  values of about 0.85 and 0.83 for irrigated and rainfed soybean, respectively, similar to the prediction by Cubist and RF models in the present work. Furthermore, they found lower values of  $R^2$  for Partial Least Squares Regression (PLSR) and SVM than what was obtained for RF and DNN. They concluded that although some models gave a better performance, prediction accuracy enhancement was insignificant.

The impact of the inclusion of diverse feature attributes in crop trait modeling has been explored in many studies ([da Silva et al., 2020](#); [Rischbeck et al., 2016](#); [Toda et al., 2021](#); [Zheng et al., 2019](#)). This study evaluated the power of UAV-derived crop features such as canopy height, spectral bands, vegetation indices, and GLCM in soybean yield prediction. The findings revealed that crop height data derived from the 3D point cloud model are valuable in soybean grain yield prediction. The extreme importance of UAV-derived canopy height for crop yield estimation or classification has been highlighted in recent literature ([Tao et al., 2020](#); [Yu et al., 2016](#)). [Sagan et al. \(2019\)](#) tested the influence of crop height on soybean yield prediction. They proposed that plant height information is a promising alternative to traditional vegetation indices in crop yield prediction. Furthermore, [Kedia et al. \(2021\)](#) reported that incorporating UAV-derived canopy height feature increased overall accuracy from 80 to 93%; while mapping invasive vegetation types in the arid regions of the USA.

This study further elucidates the potential of canopy textural attributes (GLCM) as essential input variables for soybean yield modeling. Our results suggest that GLCM variables are promising alternatives to the popular vegetation indices (VIs), as the GLCM-based model outperformed the VI model. Among the indicators often used in crop yield prediction and identification research, vegetation indices (VIs) are the most commonly used. Several studies have documented that VIs are notable indicators of crop maturity, stress, and many other attributes ([Ballester et al., 2017](#); [Sankaran et al., 2018](#); [Yeom et al., 2019](#); [X. Zhou et al., 2017](#)). However, some studies have demonstrated that NDVI, one of the most popular VI, has limitations as it gets saturated with dense vegetation cover ([Kedia et al., 2021](#); [Thenkabail et al., 2000](#); [Zheng et al., 2019](#)). Hence, other indicators with a more comprehensive range are proved efficient for applications in landcover discrimination and crop attributes modeling ([Garonna et al., 2009](#); [Yeom et al., 2019](#); [Zhang and Liu, 2014](#)). [Maimaitijiang et al. \(2020\)](#) reported that coupling texture information in the ML models significantly enhanced soybean yield estimation accuracy. The texture attributes enable a better description of crop spatial configurations, color, and intensities. Several researchers have established the potential of texture information in crop identification and trait estimation ([Böhler et al., 2018](#); [Iqbal et al., 2021](#); [Kwak and Park, 2019](#); [Zheng et al., 2019](#)). The studies by [Kwak and Park \(2019\)](#) revealed an improved accuracy of crop identification using texture information with ML models, while [Zheng et al. \(2019\)](#) reported significant performance of textural indices over vegetation indices in modeling the aboveground biomass in rice. This study revealed that applying texture information coupled with spectral and canopy height data proved superior to the VI-model for soybean yield prediction.

In addition, the use of UAV imagery for soybean grain yield prediction in this study also helped to compare the input variables at two phenological stages of the crop. Our findings suggest that phenological attributes obtained during the R6 stage significantly featured in yield prediction. The findings are consistent with other studies that reported the importance of phenological stages for successful soybean trait modeling ([Eugenio et al., 2020](#); [Herrero-Huerta et al., 2020](#); [Yoosefzadeh-Najafabadi et al., 2021](#)). Several studies have suggested the optimal time window for soybean yield prediction to be from flowering (R2) to commencement of podding (R5) ([Eugenio et al., 2020](#); [Gao et al., 2020](#); [WU et al., 2013](#)). [Ma et al. \(2001\)](#) observed that the best development phase for predicting soybean yield was the initial seed-filling stage (R5). We collected the UAV data at the beginning (R1) and the consummation of the reproductive phase (R6), which might be responsible for the observed excellent yield prediction. However, the results in this study slightly differed from that of [Eugenio et al. \(2020\)](#), who reported the best modeling stage at (V6), but in agreement with other researchers who reported better yield prediction from phenotypic traits measured at the later stages of soybean development. [Herrero-Huerta et al. \(2020\)](#) analyzed prediction errors in soybean grain yield changes based on UAV data collected at different soybean growth stages. They concluded that estimators collected at a later growth stage of the crop showed better performance in predicting grain yield.

Additionally, [Randelović et al. \(2020\)](#) compared soybean plant density prediction at two middle growth stages (V4 and R3) of soybean using a random forest model. They found better predictions from VIs collected from UAV data at the R3 stage. Authors still differ in the best stage of development crucial to soybean yield modeling attributing differences to location and cultivar characteristics. Additional research will be needed to ascertain the optimal phases and conditions for the best soybean trait prediction.

The box plots based on the ML algorithms revealed several outliers in the PVT01 & PVT02, indicating a significant level of mixtures. This might be because the materials in these trials consisted of soybean genotypes from different maturity (early and medium maturity) groups, which were evaluated separately at later advanced variety trial stages. Conversely, the absence of outliers in the PAVT and AVTES and few outliers in AVTMS might be because the trials were composed of varieties that have passed through several stages of selection and grouped based on maturity leading to the similarity in their performances; consequently, no or few outliers. The fact that materials with different genetic backgrounds and maturity groups performed differently indicate that genetic differences in the genotypes have implications on yield predictions in soybean or other crops.

In line with this, [Shook et al. \(2021\)](#) reported the insignificance of short-season variables for yield predictions of late-maturing genotypes. Ensemble-Stacking (ES) algorithm based on either the full or selected spectra reflectance can be effective in performing early selection (R5 stage) using yield prediction and identifying the best performing genotypes out of a large number of genotypes ([Yoosefzadeh-Najafabadi et al., 2021](#))

## 5. Conclusions

Our findings demonstrate the enormous potential of high-resolution drone multispectral images to predict soybean yield based on texture information, vegetation indices, and reflectance bands. The five ML regression models (Cubist, RF, SVM, GBM, and XGBoost) combined with UAV-derived multispectral reflectance data were employed to predict grain yield in soybean variety trials accurately. The main findings are that GLCM-based models slightly outperformed VI-based predictors and provided a promising alternative to the conventional use of VIs in crop yield estimation. All the five ML models performed moderately well in all the soybean variety trials investigated, though the Cubist and RF models stood out, with  $R^2$  reaching 0.89. The study provides an effective practical approach for crop variety evaluations that African-based crop breeding programs have not commonly used. With advances in UAV technology, more comprehensive studies involving the collection of additional traits such as plant height, biomass, and chlorophyll concentration can provide further insights into these modern methods. Finally, this modeling framework can be easily modified and implemented for many other crops to modernize the variety of testing techniques of the breeding programs.

## Author contributions

“Conceptualization, TA, and AA; methodology, TA; software, TA and FK; Experimental design, AA, and GC; formal analysis, TA; resources, A.A and GC; data curation, AA; writing—original draft preparation, TA; writing—review and editing, TA; A.A visualization, FK; supervision, GC; funding acquisition, AA and CG.

## Funding

The soybean trials used in this study were funded by USAID, while the Pan African Variety Trial was supported by Feed the Future Soybean Innovation Lab, University of Illinois; while the International Institute of Tropical Agricultural (IITA) provided logistics, research facilities, and administrative support for the implementation of the trials.

## Data availability statement

All the data used in this study are included in this published article and can be made available upon request.

## Ethical statement

The authors declare that all ethical practices have been followed in relation to the article’s development, writing, and publication.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We would like to acknowledge the Soybean Innovation Lab for facilitating and financially supporting the Pan African Variety Trial (PAVT) in Nigeria. We want to acknowledge the different National Soybean Breeding Programs that contributed to the soybean varieties used in the Pan-African Variety Trial.

## References

- Araus, J., Kefauver, S.C., M, Z.-A., Olsen, M.S., Cairns, J.E., 2018. Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci.* 23 (5), 451–466. <https://doi.org/10.1016/j.TPLANTS.2018.02.001>.
- Ballester, C., Hornbuckle, J., Brinkhoff, J., Smith, J., Quayle, W., 2017. Assessment of in-season cotton nitrogen status and lint yield prediction from unmanned aerial system imagery. *Rem. Sens.* 9 (11), 1149. <https://doi.org/10.3390/RS9111149>. 2017, Vol. 9, Page 1149.
- Baret, F., Guyot, G., 1991. Potentials and limits of vegetation indices for LAI and APAR assessment. *Remote Sensing of Environment* 35 (2–3), 161–173. [https://doi.org/10.1016/0034-4257\(91\)90009-U](https://doi.org/10.1016/0034-4257(91)90009-U).
- Barnes, E.M., Clarke, T.R., Richards, S.E., 2000. Coincident Detection of Crop Water Stress, Nitrogen Status and Canopy Density Using Ground Based Multispectral Data. Fifth International Conference on Precision Agriculture, Madison, WI. [https://www.researchgate.net/profile/Peter-Waller/publication/43256762\\_Coincident-detection-of-crop-water-stress-nitrogen-status-and-canopy-density-using-ground-based-multispectral-data/links/55ac358c08ae481aa7ff4da7/Coincident-detection-of-crop-water-stress-nitrogen-status-and-canopy-density-using-ground-based-multispectral-data.pdf](https://www.researchgate.net/profile/Peter-Waller/publication/43256762_Coincident-detection-of-crop-water-stress-nitrogen-status-and-canopy-density-using-ground-based-multispectral-data/links/55ac358c08ae481aa7ff4da7/Coincident-detection-of-crop-water-stress-nitrogen-status-and-canopy-density-using-ground-based-multispectral-data.pdf).
- Bendig, J., Yu, K., Aasen, H., Bolten, A., Bennertz, S., Broscheit, J., Gnyp, M.L., Bareth, G., 2015. Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *Int. J. Appl. Earth Obs. Geoinf.* 39, 79–87. <https://doi.org/10.1016/j.JAG.2015.02.012>.
- Birth, G.S., McVey, G.R., 1968. Measuring the color of growing turf with a reflectance Spectrophotometer1. *Agron. J.* 60 (6), 640–643. <https://doi.org/10.2134/AGRONJ1968.0002196200600060016X>.
- Böhler, J.E., Schaepman, M.E., Kneubühler, M., 2018. Crop classification in a heterogeneous arable landscape using uncalibrated UAV data. *Rem. Sens.* 10 (8) <https://doi.org/10.3390/rs10081282>.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45 (1), 5–32. <https://doi.org/10.1023/A:1010933404324>. 45(1), 2001.
- Chang, A., Jung, J., Maeda, M.M., Landivar, J., 2017. Crop height monitoring with digital imagery from Unmanned Aerial System (UAS). *Comput. Electron. Agric.* 141, 232–237. <https://doi.org/10.1016/J.COMPAG.2017.07.008>.
- Chang, A., Jung, J., Yeom, J., Maeda, M.M., Landivar, J.A., Enciso, J.M., Avila, C.A., Anciso, J.R., 2021. Unmanned aircraft system- (UAS-) based high-throughput phenotyping (HTP) for tomato yield estimation. *J. Sens.* <https://doi.org/10.1155/2021/8875606>, 2021.
- Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. *Proceed. ACM SIGKDD Int. Conf. Knowled. Discov. Data Min.* 785–794. <https://doi.org/10.1145/2939672.2939785>. 13-17-August-2016.

- Chigeza, G., Boahen, S., Gedil, M., Agoyi, E., Mushoriwa, H., Denwar, N., Gondwe, T., Tesfaye, A., Kamara, A., Alamu, O.E., Chikoye, D., 2019. Public sector soybean (Glycine max) breeding: advances in cultivar development in the African tropics. *Plant Breed.* 138 (4), 455–464. <https://doi.org/10.1111/pbr.12682>.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20 (3), 273–297. <https://doi.org/10.1007/BF00994018>, 1995 20:3.
- da Silva, E.E., Rojo Baio, F.H., Ribeiro Teodoro, L.P., da Silva Junior, C.A., Borges, R.S., Teodoro, P.E., 2020. UAV-multispectral and vegetation indices in soybean grain yield prediction based on in situ observation. *Remote Sens. Appl.: Soc. Environ.* 18, 100318 <https://doi.org/10.1016/J.RSASE.2020.100318>.
- Daughtry, C.S.T., Walthall, C.L., Kim, M.S., De Colstoun, E.B., McMurtrey, J.E., 2000. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Remote Sensing of Environment* 74 (2), 229–239. [https://doi.org/10.1016/S0034-4257\(00\)00113-9](https://doi.org/10.1016/S0034-4257(00)00113-9).
- Deering, D.W., Rouse, J.W., Haas, R.H., Schell, J.A., 1975. Measuring “forage production” of grazing units from landsat MSS data. In: *Proceedings of the 10th International Symposium on Remote Sensing of Environment*, pp. 1169–1178.
- Degenhardt, F., Seifert, S., Szymczak, S., 2019. Evaluation of variable selection methods for random forests and omics data sets. *Briefings Bioinf.* 20 (2), 492–503. <https://doi.org/10.1093/BIB/BBX124>.
- Di Gennaro, S.F., Rizza, F., Badeck, F.W., Berton, A., Delbono, S., Gioli, B., Toscano, P., Zaldei, A., Matese, A., 2017. UAV-based high-throughput phenotyping to discriminate barley vigour with visible and near-infrared vegetation indices. <https://doi.org/10.1080/01431161.2017.1395974>, 39,15–16,5330-5344. <https://doi.org/10.1080/01431161.2017.1395974>.
- Diers, B., Scaboo, A., 2019. Soybean breeding in africa. *Afr. J. Food Nutr. Sci.* 19 (5), 15121–15125. <https://doi.org/10.18697/ajfand.88.SILFarmDoc03>.
- Eugenio, F.C., Grohs, M., Venancio, L.P., Schuh, M., Bottega, E.L., Ruoso, R., Schons, C., Mallmann, C.L., Badin, T.L., Fernandes, P., 2020. Estimation of soybean yield from machine learning techniques and multispectral RPAS imagery. *Remote Sens. Appl.: Soc. Environ.* 20 (April) <https://doi.org/10.1016/j.rsase.2020.100397>.
- FAO, 2021. FAOstat. FAO, Rome, Italy. <https://www.fao.org/faostat/en/#data/QCL>. accessed 12/11/2021.
- Freeman, E.A., Moisen, G.G., Coulston, J.W., Wilson, B.T., 2015. Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. *Can. J. For. Res.* 46 (3), 323–339. <https://doi.org/10.1139/cjfr-2014-0562>.
- Fukano, Y., Guo, W., Aoki, N., Ootsuka, S., Noshita, K., Uchida, K., Kato, Y., Sasaki, K., Kamikawa, S., Kubota, H., 2021. GIS-based analysis for UAV-supported field experiments reveals soybean traits associated with rotational benefit. *Front. Plant Sci.* 12 (May), 1–11. <https://doi.org/10.3389/fpls.2021.637694>.
- Fushiki, T., 2009. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* 21 (2), 137–146. <https://doi.org/10.1007/S11222-009-9153-8>, 2009 21:2.
- Gao, F., Anderson, M., Daughtry, C., Karnieli, A., Hively, D., Kustas, W., 2020. A within-season approach for detecting early growth stages in corn and soybean using high temporal and spatial resolution imagery. *Remote Sensing of Environment* 242, 111752. <https://doi.org/10.1016/J.RSE.2020.111752>.
- Garonna, I., Fazey, I., Brown, M.E., Pettorelli, N., 2009. Rapid primary productivity changes in one of the last coastal rainforests: the case of Kahua, Solomon Islands. *Environ. Conserv.* 36 (3), 253–260. <https://doi.org/10.1017/S0376892909990208>.
- Gitelson, A.A., 2004. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. *J. Plant Physiol.* 161 (2), 165–173. <https://doi.org/10.1078/0176-1617-01176>.
- Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* 160 (3), 271–282. <https://doi.org/10.1078/0176-1617-00887>.
- Gitelson, A.A., Merzlyak, M.N., 1998. Remote sensing of chlorophyll concentration in higher plant leaves. *Adv. Space Res.* 22 (5), 689–692. [https://doi.org/10.1016/S0273-1177\(97\)01133-2](https://doi.org/10.1016/S0273-1177(97)01133-2).
- Gitelson, A.A., Viña, A., Ciganda, V., Rundquist, D.C., Arkebauer, T.J., 2005. Remote estimation of canopy chlorophyll content in crops. *Geophys. Res. Lett.* 32 (8), 1–4. <https://doi.org/10.1029/2005GL022688>.
- Haralick, R.M., Dinstein, I., Shanmugam, K., 1973. Textural features for image classification. *IEEE Transact. Sys. Man and Cybernet.* 3 (6), 610–621. <https://doi.org/10.1109/TSMC.1973.4309314>.
- Hassan, M.A., Yang, M., Fu, L., Rasheed, A., Zheng, B., Xia, X., Xiao, Y., He, Z., 2019. Accuracy assessment of plant height using an unmanned aerial vehicle for quantitative genomic analysis in bread wheat. *Plant Methods* 15 (1), 1–12. <https://doi.org/10.1186/s13007-019-0419-7>.
- Herrero-Huerta, M., Rodriguez-Gonzalez, P., Rainey, K.M., 2020. Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean. *Plant Methods* 16 (1), 1–16. <https://doi.org/10.1186/s13007-020-00620-6>.
- Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment* 25 (3), 295–309. [https://doi.org/10.1016/0034-4257\(88\)90106-X](https://doi.org/10.1016/0034-4257(88)90106-X).
- Iqbal, N., Mumtaz, R., Shafi, U., Zaidi, S.M.H., 2021. Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms. *PeerJ Comp. Sci.* 7, e536. <https://doi.org/10.7717/peerj-cs.536>.
- Johansen, K., Morton, M.J.L., Malbeteau, Y.M., Aragon, B., Al-Mashharawi, S.K., Ziliani, M.G., Angel, Y., Fiene, G.M., Negrão, S.S.C., Mousa, M.A.A., Tester, M.A., McCabe, M.F., 2019. Unmanned aerial vehicle-based phenotyping using morphometric and spectral analysis can quantify responses of wild tomato plants to salinity stress. *Front. Plant Sci.* 10, 370. <https://doi.org/10.3389/FPLS.2019.00370>.
- Jordan, C.F., 1969. Derivation of leaf-area index from quality of light on the forest floor. *Ecology* 50 (4), 663–666. <https://doi.org/10.2307/1936256>.
- Kedia, A.C., Kapos, B., Liao, S., Draper, J., Eddinger, J., Updike, C., Frazier, A.E., 2021. An integrated spectral–structural workflow for invasive vegetation mapping in an arid region using drones. *Drones* 5 (1), 19. <https://doi.org/10.3390/DRONES5010019>.
- Khojely, D.M., Ibrahim, S.E., Sapey, E., Han, T., 2018. History, current status, and prospects of soybean production and research in sub-Saharan Africa. *Crop J.* 6 (3), 226–235. <https://doi.org/10.1016/j.cj.2018.03.006>. Crop Science Society of China/Institute of Crop Sciences.
- Kursa, Miron B., Rudnicki, W.R., 2010. Feature selection with the boruta package. *J. Stat. Software* 36 (11), 1–13. <https://doi.org/10.18637/jss.v036.i11>.
- Kursa, Bartosz, Miron, Rudnicki, W.R., 2021. Package ‘boruta’-wrapper algorithm for all relevant feature selection. <https://gitlab.com/mqb/Boruta>.
- Kwak, G.H., Park, N.W., 2019. Impact of texture information on crop classification with machine learning and UAV images. *Appl. Sci.* 9 (4) <https://doi.org/10.3390/app9040643>.
- Leutner, B., Horning, N., Schwalb-Willmann, J., Hijmans, R.J., 2019. Tools for remote sensing data analysis-package ‘RStoolbox’. <https://github.com/bleutner/RStoolbox>.
- Ma, B.L., Dwyer, L.M., Costa, C., Cober, E.R., Morrison, M.J., 2001. Early prediction of soybean yield from canopy reflectance measurements. *Agron. J.* 93 (6), 1227–1234. <https://doi.org/10.2134/AGRONJ2001.1227>.
- Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F.B., 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of Environment* 237 (December 2019). <https://doi.org/10.1016/j.rse.2019.111599>.
- Makanza, R., Zaman-Allah, M., Cairns, J.E., Magorokosho, C., Tarekegne, A., Olsen, M., Prasanna, B.M., 2018. High-throughput phenotyping of canopy cover and senescence in maize field trials using aerial digital canopy imaging. *Rem. Sens.* 10 (2), 330. <https://doi.org/10.3390/RS10020330>, 2018, Vol. 10, Page 330.
- Malambo, L., Popescu, S.C., Murray, S.C., Putman, E., Pugh, N.A., Horne, D.W., Richardson, G., Sheridan, R., Rooney, W.L., Avanti, M., Vidrine, M., McCutchen, B., Baltensperger, D., Bishop, M., 2018. Multitemporal field-based plant height estimation using 3D point clouds generated from small unmanned aerial systems high-resolution imagery. *Int. J. Appl. Earth Obs. Geoinf.* 64, 31–42. <https://doi.org/10.1016/J.JAG.2017.08.014>.
- McFEETERS, S.K., 2007. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. <https://doi.org/10.1080/01431169608948714>, 17,7,1425-1432.
- Nguyen, Q.H., Ly, H.B., Ho, L.S., Al-Ansari, N., Van Le, H., Tran, V.Q., Prakash, I., Pham, B.T., 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Math. Probl Eng.* <https://doi.org/10.1155/2021/4832864>, 2021.
- Noi, P.T., Degener, J., Kappas, M., 2017. Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data. *Rem. Sens.* 9 (5), 398. <https://doi.org/10.3390/RS9050398>, 2017, Vol. 9, Page 398.
- Oladoye, A., 2015. Physicochemical properties of soil under two different depths in a tropical forest of international institute of tropical agriculture, Abeokuta, Ibadan, Nigeria. *J. Res. Forest. Wildlife Environ.* 7 (1), 40–54. <https://www.ajol.info/index.php/jrfwe/article/view/116910>.
- Perry, C.R., Lautenschlager, L.F., 1984. Functional equivalence of spectral vegetation indices [Species, leaf area, stress, biomass, multispectral scanner measurements, Landsat, remote sensing]. *Remote Sensing of Environment.* <https://agris.fao.org/agris-search/search.do?recordID=US19850043085>.



- Pinty, B., Verstraete, M.M., 1992. GEMI: a non-linear index to monitor global vegetation from satellites. *Vegetatio* 101 (1), 15–20. <https://doi.org/10.1007/BF00031911>, 1992 101:1.
- Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., Sorooshian, S., 1994. A modified soil adjusted vegetation index. *Remote Sensing of Environment* 48 (2), 119–126. [https://doi.org/10.1016/0034-4257\(94\)90134-1](https://doi.org/10.1016/0034-4257(94)90134-1).
- Quinlan, J.R., Quinlan, J.R., 1992. Learning with continuous classes. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.885>, 343–348.
- Randelović, P., Dordević, V., Milić, S., Bale ević-Tubić, S., Petrović, K., Miladinović, J., Đukić, V., 2020. Prediction of soybean plant density using a machine learning model and vegetation indices extracted from RGB images taken with a UAV. *Agronomy* 10 (8). <https://doi.org/10.3390/agronomy10081108>.
- Räsänen, A., Virtanen, T., 2019. Data and resolution requirements in mapping vegetation in spatially heterogeneous landscapes. *Remote Sensing of Environment* 230 (December 2018), 111207. <https://doi.org/10.1016/j.rse.2019.05.026>.
- Richardson, A.J., Weigand, C., 1977. Distinguishing vegetation from soil background information. *Photogramm. Eng. Rem. Sens.* [http://www.asprs.org/wp-content/uploads/pers/1977journal/dec/1977\\_dec\\_1541-1552.pdf](http://www.asprs.org/wp-content/uploads/pers/1977journal/dec/1977_dec_1541-1552.pdf).
- Rischbeck, P., Elsayed, S., Mistele, B., Barmerier, G., Heil, K., Schmidhalter, U., 2016. Data fusion of spectral, thermal and canopy height parameters for improved yield prediction of drought stressed spring barley. *Eur. J. Agron.* 78, 44–59. <https://doi.org/10.1016/J.EJA.2016.04.013>.
- Roth, L., Streit, B., 2018. Predicting cover crop biomass by lightweight UAS-based RGB and NIR photography: an applied photogrammetric approach. *Precis. Agric.* 19 (1), 93–114. <https://doi.org/10.1007/S11119-017-9501-1>.
- Rouse, J.W., Hass, R., Schell, J.A., Deering, D.W., 1974. Monitoring vegetation systems in the great plains with ERTS. In: Freden, S.C., Mercanti, E.P., Becker, M.A. (Eds.), *Third Earth Resources Technology Satellite-1 Symposium*, vol. 1. Technical Presentations, NASA. <https://www.scopus.com/record/display.uri?eid=2-s2.0-24344476424&origin=inward&txGid=efc9a15464a8e9860966c08a43caf7b>. Washington, D.C.
- Sagan, V., Maimaitijiang, M., Sidike, P., Maimaitiyiming, M., Erkol, H., Hartling, S., Peterson, K.T., Peterson, J., Burken, J., Fritschi, F., 2019. Uav/satellite multiscale data fusion for crop monitoring and early stress detection. *Int. Arch. Photogrammet. Rem. Sens. Spat. Inform. Sci. ISPRS Arch.* 42 (2/W13), 715–722. <https://doi.org/10.5194/isprs-archives-XLII-2-W13-715-2019>.
- Sanchez-Pinto, L.N., Venable, L.R., Fahrenbach, J., Churpek, M.M., 2018. Comparison of variable selection methods for clinical predictive modeling. *Int. J. Med. Inf.* 116 (October 2017), 10–17. <https://doi.org/10.1016/j.ijmedinf.2018.05.006>.
- Sankaran, S., Zhou, J., Khot, L.R., Trapp, J.J., Mndolwa, E., Miklas, P.N., 2018. High-throughput field phenotyping in dry bean using small unmanned aerial vehicle based multispectral imagery. *Comput. Electron. Agron.* 151, 84–92. <https://doi.org/10.1016/J.COMPAG.2018.05.034>.
- Santos, M., 2019. Soybean varieties in sub-Saharan Africa. *Afr. J. Food Nutr. Sci.* 19 (5), 15136–15139. <https://doi.org/10.18697/ajfand.88.SILFarmDoc06>.
- Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., Singh, A.K., 2021. Crop yield prediction integrating genotype and weather variables using deep learning. *PLoS One* 16 (6 June 2021), 1–19. <https://doi.org/10.1371/journal.pone.0252402>.
- Sidike, P., Sagan, V., Qumsiyeh, M., Maimaitijiang, M., Essa, A., Asari, V., 2018. Adaptive trigonometric transformation function with image contrast and color enhancement: application to unmanned aerial system imagery. *Geosci. Rem. Sens. Lett. IEEE* 15 (3), 404–408. <https://doi.org/10.1109/LGRS.2018.2790899>.
- Sinclair, T.R., Marrou, H., Soltani, A., Vadez, V., Chandolu, K.C., 2014. Soybean production potential in Africa. *Global Food Secur.* 3 (1), 31–40. <https://doi.org/10.1016/j.gfs.2013.12.001>.
- Singh, A., Ganapathysubramanian, B., Singh, A.K., Sarkar, S., 2016. Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci.* 21 (2), 110–124. <https://doi.org/10.1016/J.TPLANTS.2015.10.015>.
- Speiser, J.L., Miller, M.E., Toozee, J., Ip, E., 2019. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* 134, 93–101. <https://doi.org/10.1016/j.eswa.2019.05.028>.
- Stanton, C., Starek, M.J., Elliott, N., Brewer, M., Maeda, M.M., Chu, T., 2017. Unmanned aircraft system-derived crop height and normalized difference vegetation index metrics for sorghum yield and aphid stress assessment. *J. Appl. Remote Sens.* 11 (2), 026035. <https://doi.org/10.1117/1.JRS.11.026035>.
- Suab, S.A., Avtar, R., 2020. Unmanned aerial vehicle system (UAVS) applications in forestry and plantation operations: experiences in sabah and sarawak, Malaysian borneo. In: *Unmanned Aerial Vehicle: Applications in Agriculture and Environment*. [https://doi.org/10.1007/978-3-030-27157-2\\_8](https://doi.org/10.1007/978-3-030-27157-2_8).
- Tao, H., Feng, H., Xu, L., Miao, M., Yang, G., Yang, X., Fan, L., 2020. Estimation of the yield and plant height of winter wheat using UAV-based hyperspectral images. *Sensors* 20 (4), 1231. <https://doi.org/10.3390/S20041231>, 2020, Vol. 20, Page 1231.
- Thenkabail, P.S., Smith, R.B., De Pauw, E., 2000. Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote Sensing of Environment* 71 (2), 158–182. [https://doi.org/10.1016/S0034-4257\(99\)00067-X](https://doi.org/10.1016/S0034-4257(99)00067-X).
- Thiam, A.K., 1997. *Geographic Information Systems and Remote Sensing Methods for Assessing and Monitoring Land Degradation in the Sahel: The Case of Southern Mauritania*. Clark University, Worcester Massachusetts.
- Toda, Y., Kaga, A., Kajiya-Kanegae, H., Hattori, T., Yamaoka, S., Okamoto, M., Tsujimoto, H., Iwata, H., 2021. Genomic prediction modeling of soybean biomass using UAV-based remote sensing and longitudinal model parameters. *Plant Genome* 14 (3), e20157. <https://doi.org/10.1002/TPG2.20157>.
- Wang, B., Oldham, C., Hipsey, M.R., 2016. Comparison of machine learning techniques and variables for groundwater dissolved organic nitrogen prediction in an urban area. *Procedia Eng.* 154, 1176–1184. <https://doi.org/10.1016/J.PROENG.2016.07.527>.
- Watanabe, K., Guo, W., Arai, K., Takanashi, H., Kajiya-Kanegae, H., Kobayashi, M., Yano, K., Tokunaga, T., Fujiwara, T., Tsutsumi, N., Iwat, H., 2017. High-throughput phenotyping of sorghum plant height using an unmanned aerial vehicle and its application to genomic prediction modeling. *Front. Plant Sci.* 8. <https://doi.org/10.3389/FPLS.2017.00421>.
- Wu, Q., Qi, B., Zhao, T.-J., Yao, X.-F., Zhu, Y., Gai, J.-Y., 2013. A tentative study on utilization of canopy hyperspectral reflectance to estimate canopy growth and seed yield in soybean. *Acta Agron. Sin.* 39 (2), 309. <https://doi.org/10.3724/SP.J.1006.2013.00309>.
- Yang, G., Liu, J., Zhao, C., Li, Z., Huang, Y., Yu, H., Xu, B., Yang, X., Zhu, D., Zhang, X., Zhang, R., Feng, H., Zhao, X., Li, Z., Li, H., Yang, H., 2017. Unmanned aerial vehicle remote sensing for field-based crop phenotyping: current status and perspectives. *Front. Plant Sci.* 8, 1111. <https://doi.org/10.3389/FPLS.2017.01111/BIBTEX>.
- Yeom, J., Jung, J., Chang, A., Ashapure, A., Maeda, M., Maeda, A., Landivar, J., 2019. Comparison of vegetation indices derived from UAV data for differentiation of tillage effects in agriculture. *Rem. Sens.* 11 (13). <https://doi.org/10.3390/rs11131548>.
- Yoosofzadeh-Najafabadi, M., Earl, H.J., Tulpan, D., Sulik, J., Eskandari, M., 2021. Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. *Front. Plant Sci.* 11 (January), 1–14. <https://doi.org/10.3389/fpls.2020.624273>.
- Yu, N., Li, L., Schmitz, N., Tian, L.F., Greenberg, J.A., Diers, B.W., 2016. Development of methods to improve soybean yield estimation and predict plant maturity with an unmanned aerial vehicle based platform. *Remote Sensing of Environment* 187, 91–101. <https://doi.org/10.1016/J.RSE.2016.10.005>.
- Zhang, S., Liu, L., 2014. The potential of the MERIS Terrestrial Chlorophyll Index for crop yield prediction. <http://Dx.Doi.Org/10.1080/2150704X.2014.963734>. <https://doi.org/10.1080/2150704X.2014.963734>, 5, 8, 733–742.
- Zheng, H., Cheng, T., Zhou, M., Li, D., Yao, X., Tian, Y., Cao, W., Zhu, Y., 2019. Improved estimation of rice aboveground biomass combining textural and spectral analysis of UAV imagery. *Precis. Agric.* 20 (3), 611–629. <https://doi.org/10.1007/s11119-018-9600-7>.
- Zhou, J., Li, E., Wei, H., Li, C., Qiao, Q., Armaghani, D.J., 2019. Random forests and cubist algorithms for predicting shear strengths of rockfill materials. *Appl. Sci.* 9 (8), 1621. <https://doi.org/10.3390/APP9081621>, 2019, Vol. 9, Page 1621.
- Zhou, X., Zheng, H.B., Xu, X.Q., He, J.Y., Ge, X.K., Yao, X., Cheng, T., Zhu, Y., Cao, W.X., Tian, Y.C., 2017. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. *ISPRS J. Photogrammetry Remote Sens.* 130, 246–255. <https://doi.org/10.1016/J.ISPRSJPRS.2017.05.003>.
- Zvoleff, Alex, 2020. Glcm: calculate textures from grey-level Co-occurrence matrices (GLCMs) version 1.6.5 from CRAN. CRAN Package ‘GlcM’. <https://rdr.io/cran/glcM/>.