Machine learning-based prediction of sorghum biomass from UAV multispectral imagery data

Conference Paper · May 2023

DOI: 10.1109/I3CS58314.2023.10127516

3 authors, including:

Boubacar Gano
Donald Danforth Plant Science Center
15 PUBLICATIONS **122** CITATIONS
SEE PROFILE

Nurzaman Ahmed
Donald Danforth Plant Science Center
84 PUBLICATIONS **1,196** CITATIONS
SEE PROFILE

2023 4th International Conference on Computing and Communication Systems (I3CS) | 979-8-3503-2377-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/I3CS58314.2023.10127516

Machine learning-based prediction of sorghum biomass from UAV multispectral imagery data

Boubacar Gano[1], Nurzaman Ahmed[2], and Nadia Shakoor[3]

[1,2,3]Donald Danforth Plant Science Center
Saint Louis, MO, USA, 63132
{[1]bgano, [2]nahmed, [3]nshakoor}@danforthcenter.org

***Abstract*—Unmanned aerial vehicle (UAV)-based remote sensing applications in plant phenotyping have received attention in modern plant breeding programs that increasingly have the need to automate time-consuming manual measurements of agronomic traits. This paper focuses on the prediction of sorghum biomass using machine learning algorithms such as Linear Regression, K-Neighbors Regressor, and the XGBoost Regressor. Results from a field experiment of 344 sorghum genotypes conducted at the Donald Danforth Plant Science Center (Saint Louis, MO, USA) showed accurate prediction models. The K-Neighbors Regression model performed better than the other two models (R^2 = 0.65, RMSE = 4968.60 kg/ha). The developed approach in this study could be used as a decision support tool for sorghum biomass phenotyping in breeding programs.**

***Index Terms*—UAV, Remote sensing, Plant phenotyping, Machine learning, Sorghum**

I. Introduction

Biomass is an important trait governing the biofuel production capacity of sorghum genotypes, as it is indicative of plant growth and the ability to produce ethanol [1]. Novel remote sensing technologies involving multispectral imaging sensors on board a UAV provide a compelling alternative to the more challenging, labor-intensive, and time-consuming traditional phenotyping methods. Also, the existing methods for biomass measurement involve destructive sampling, which may be logistically challenging for large-scale breeding programs [2], [3]. Recently, remote sensing (RS) data have been used as inputs of machine learning (ML) models to develop accurate predictions of phenotypic traits such as plant biomass. Deep neural networks are also widely explored by many researchers and offer more accuracy compared to classical ML [4].

A major challenge of ML is the requirement of ground truth data, where quantity and quality underpin the ML model's capacity to consider hypotheses and distributions from the training datasets. The prediction accuracy will depend on ground truth data size, type, and prediction methods. Although ML and UAV-sensor data have been utilized to estimate biomass in many crops, including maize [5], wheat [6], pea [7], and sorghum [1], [8], [9], the prediction accuracy varies amongst models, genotypes and environment for the same crop species. Therefore, the quest for appropriate models for a defined number of genotypes under specific environmental conditions is a relevant subject and, to the best of our knowledge, no such studies have been done using the ML algorithms used in this study with multispectral sensor data for estimating end-of-season sorghum above-ground biomass.

To solve the above-mentioned issues, this work develops ML-based prediction models of final sorghum biomass using UAV multispectral imagery data, with the intention to aid field phenotyping operations for an efficient large-scale breeding program. The described work shows the potential of spectral data from UAV imaging to capture plant biomass traits. We also used ML models to obtain prediction accuracy with limited data sizes. Finally, we compare the ML models.

The rest of the paper is organized as follows. Section II presents the related works of our proposed solution. We discuss the proposed ML methods in Section III. Section IV discusses the performance analysis process and results. Finally, Section V summarizes the conclusions of the paper.

II. Related Works

We discuss the existing literature based on various aspects of data collection and ML methods in the case of biomass prediction.

Masjedi et al. [10] used a recurrent neural networks model to predict sorghum biomass from time series UAV data. Zhang et al. [11] used RGB and hyperspectral UAV-based image data as input features to explore multiple layer perception (MLP) neural networks and support vector regression (SVR) for predicting sorghum biomass. They found the MLP method to be more accurate when the number of samples in the training dataset was limited, while the SVR models performed better than MLP when the number of samples increased. Among these studies, ML algorithms achieved high-accuracy prediction of sorghum biomass.

Although the classical ML methods are powerful tools in modeling crop biomass, the model accuracy drops when time series data from multiple sensors, collected from different locations are all used as input features [10], [12]. In the last decade, deep neural networks have been widely explored by many researchers and offer more accuracy compared to classical ML [4]. When using ML regression models, the prediction accuracy will depend on many factors like the ground truth data size, type, and prediction method [12]. However, biomass prediction based on UAV data and ML models

979-8-3503-2377-1/23/$31.00 ©2023 IEEE

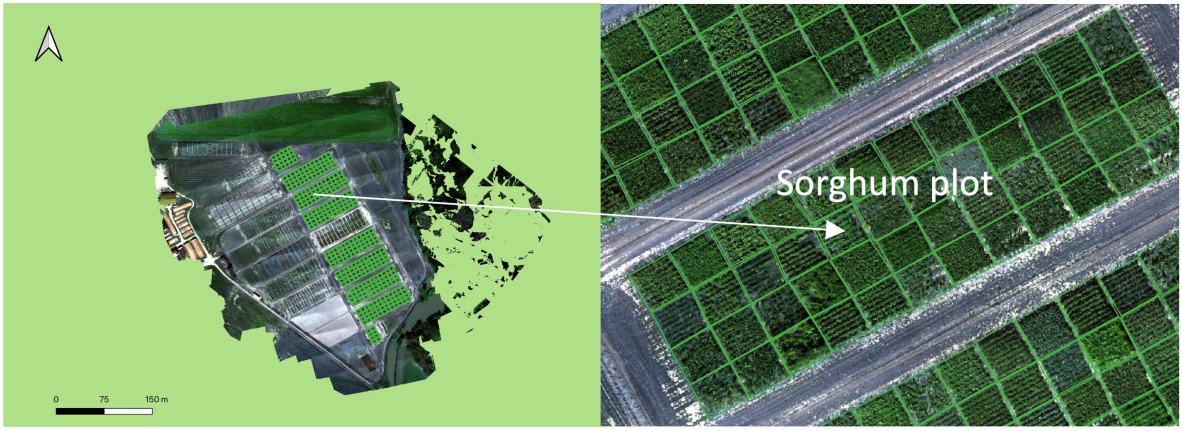

Fig. 1: Orthoimage of the field experiment with plot delimitation

remains challenging because of the complexity of the biomass trait [13], the lack of ground truth data in terms of quantity for model validation [14], and high phenotypic variability observed in field experiments [15]. The main characteristic of ML is the requirement of ground truth information that underlies the model's capacity to consider hypotheses and distributions directly from the training dataset [16]. For ML models to be able to properly predict traits, another common challenge is the requirement of a similar distribution between training and testing datasets, even for extensive training data [17], [18].

Previous studies used different ML algorithms (e.g., RNN, SVR, MLP, RF, PLSR, CART) with hyperspectral, RGB and LiDAR data to develop prediction models for sorghum biomass [1], [10], [11]. However, to our knowledge, there are no such studies that applied the ML algorithms used in this study to multispectral sensor data for estimating end-of-season sorghum above-ground biomass.

III. ML-BASED PREDICTION OF SORGHUM BIOMASS

We discuss the proposed ML-based prediction of sorghum biomass in the following subsections. The initial phase of the proposed method presents the data collection and processing. Thereafter, three different ML-based regression models – Linear Regression, K-Neighbors Regressor, and XGBoost Regressor are used to study the Root Mean Squared Error (RMSE) and R-squared (R^2).

A. Ground truth and UAV data Collection

The field experiment was conducted at the Donald Danforth Plant Science Center Field Research Site in O'Fallon, MO. An augmented design with 10 repeated genotypes (checks) and 344 non-repeated genotypes were distributed in 3 blocks and 9 sub-blocks (Fig. 1). UAV data collection was carried out using a multicopter drone equipped with a Micasense Altum multispectral camera (Micasense, inc) with five spectral bands (blue, green, red, rededge, and near-infrared). The drone made a round trip over the entire field, allowing a side and forward overlapping fraction of 0.8 between raw images. At crop maturity, plants were sampled and weighed for above-ground biomass measurements.

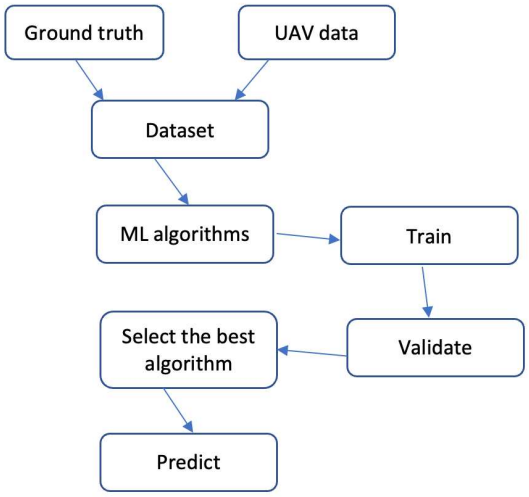

Fig. 2: Flow diagram of biomass prediction

B. Image processing

UAV raw images were processed using Pix4D [24] software to generate calibrated and corrected orthomosaics. Real reflectances were calculated using a control panel with known reflectance, and multiband co-registration was done to adjust and correct the images' coordinate system and decrease geometric distortions. Plot delimitation was done in QGIS software and the generated shapefile was used, along with the multiband orthoimage in R software (Raster, RStoolbox, tidyverse packages) to extract vegetation indices (Table 1) and near-infrared bands for input into the ML models. Features were selected based on their high relationship with biomass, and Modified Soil-Adjusted Vegetation Index, MSAVI, was the most related feature.

TABLE I: Vegetation indices used as inputs for ML algorithms

Vegetation indices	Formulas	Ref.
Normalized Difference Vegetation Index	NDVI = $\frac{\rho NIR-\rho Red}{\rho NIR+\rho Red}$	**[19]**
The Corrected Transformed Vegetation Index	CTVI = $\frac{NDVI+0.5}{\lvert NDVI+0.5\rvert} \times \sqrt{\lvert\, NDVI+0.5\,\rvert}$	**[20]**
Green Normalized Difference Vegetation Index	GNDVI = $\frac{\rho NIR-\rho Green}{\rho NIR+\rho Green}$	**[21]**
Modified Soil-Adjusted Vegetation Index	MSAVI = $\frac{2*\rho NIR+1-\sqrt{(2*\rho NIR+1)2-8*(\rho NIR-\rho Red))}}{2}$	**[22]**
Normalized difference red edge Index	NDRE = $\frac{\rho NIR-\rho RE}{\rho NIR+\rho RE}$	**[23]**

Note: Near-Infrared band (NIR), Red Edge band (RE)

C. ML-based evaluation process

Fig. 2 shows the proposed ML-based biomass prediction method. After collecting data in the form of ground truth and UAV, the combined data were passed through a set of pre-processing steps including data cleaning and feature engineering. We used feature ranking with recursive feature elimination for fitting. The final dataset was passed through the considered ML-based regression models for both the training and validation datasets. Finally, the best algorithm was chosen for predicting biomass. We discuss the best-performing set of ML models in the following subsections.

D. Machine Learning algorithms

1) Linear Regression (LR): Recently, Multiple Linear Regression (LR) model-based supervised learning has proven to be suitable and reliable for predictions. Linear regression is generally used in research studies to evaluate the predicted effects and model them against multiple input variables. This method usually analyzes and learns initial training data from which it models relationships between dependent and independent variables. The longitudinal regression of LR has high precision in long-term trait prediction with a slight variance [25]. The properties of this model, such as being well-understood, fast, and minimizing 'lack of fit,' motivate us to use it.

2) K-Neighbors Regressor (KNN): The input of this algorithm allows choosing the k-closest training examples in a dataset, which helps identify and remove outliers. We use dynamic k values in the considered use case scenarios. KNN can predict more accurately within a long-history database, bringing more similar neighboring patterns. One of the key advantages of the KNN model is that the prediction accuracy is not affected by increased data size after a certain threshold level [26].

3) XGBoost Regressor (XGB): Another supervised learning, XGBoost Regressor (XGB), is used for the proposed prediction. The objective function of XGB contains a loss function and a regularization term, which finds the difference between actual and predicted values. XGB is promised to be salable with a productive improvement of gradient-boosting decision tree implementation. It allows building a new weak learner that is highly correlated with the loss function negative gradient linked to the whole assembly for each iteration [27], [28]. XGB offers a novel distributed algorithm that expedites the boosted tree searching and construction. The contribution score of each feature to the training model is considered for evaluating and selecting the appropriate features for efficient prediction [17].

The choice of the three ML methods in this study was based on previous tests using many algorithms. The final set of models were found to be best suited to the considered features and scenarios. Further, tree learning algorithms like XGB are better suited for this type of dataset as they do not imply linear interactions between features. The KNN method is more tolerant to low data size, and LR has interesting properties like minimizing lack of fit, slight variance, and long-term prediction capability.

IV. Performance Evaluation

For the performance evaluation, we used three ML-based regression models, including Linear Regression (LR), K-Neighbors Regressor (KNN), and XGBoost Regressor (XGB). The regression model outputs were assessed using the R^2 coefficient of determination and the $RMSE$. The RMSE indicates the distances of predicted values from the observed values in the dataset. A ML model will fit better in the dataset when the RMSE value is less. The RMSE value is calculated as follows:

$$RMSE = \sqrt{\frac{\sum(P_i - O_i)^2}{N}} \tag{1}$$

where, P_i, and O_i are the predicted and observed value for the i^{th} observation, respectively, and N is the sample size. On the other hand, the value of R^2 ($0 \leq R^2 \leq 1$) indicates how closely data are associated with the fitted regression line. A higher R^2 value indicates a better fit of a model. The R^2 value can be calculated as:

$$R^2 = 1 - \frac{SSR}{TSS} \tag{2}$$

where, SSR is the sum of squares of residuals and TSS is the total sum of squares.

TABLE II: Performance comparison of ML algorithms

Algorithms	RMSE	R^2
K-Neighbors Regressor	4968.60	0.65
Linear Regression	5082.52	0.64
XGB Regressor	7896.11	0.13

A. Model evaluation (R^2 and RMSE)

For regression modeling, three ML-based models were used in this study, including Linear Regression (LR), K-Neighbors Regressor (KNN), and XGBoost Regressor (XGB). The regression model outputs were assessed using the R^2 coefficient of determination and the root-mean-square error (RMSE). The R^2 is the proportion of variance in the dependent variable that's explained by the original model using prediction test data, and it is highly related to the overall accuracy of the model. In each model, we used a data partitioning of 80% training and 20% testing data. The regression analyses were conducted using python language. Results presented in Table 2 showed good model accuracy for LR and KNN with R^2 values of 0.64 and 0.65, respectively. The XGB model, however, recorded low prediction accuracies, likely due to a non-suitable model and limited data size. This model performed better in soybean grain yield prediction with an R^2 of 0.41 when 60 features (radiometric and geometric) obtained from both RGB and multispectral cameras were used [17]. The results shown in Fig. 3 represent a single linear regression between MSAVI and biomass, highlighting the potential of the different vegetation indices used to estimate plant biomass in sorghum crops.

B. Features

Low data size remains the most problematic in the ML regression approach. Despite many studies showing an increase of model performance when the number of genotypes and plots is higher, the number of features also can have an impact on model performance. Here, we evaluate model performance under an increasing number of features (Fig. 4 & 5). The results showed different responses with three ML models. While the LR model exhibited the highest increase in performance when the number of features increased from 2 to 6, the XGB model did not show feature-dependent variation. However, the KNN model showed a slight increase with the increasing number of features (i.e., up to 3 features), then maintains an optimal R^2 (i.e., 0.65). This suggests that the KNN model fits better when fewer data features are used for ML studies, while the LR model accuracy should provide the best accuracy when the data feature size increases.

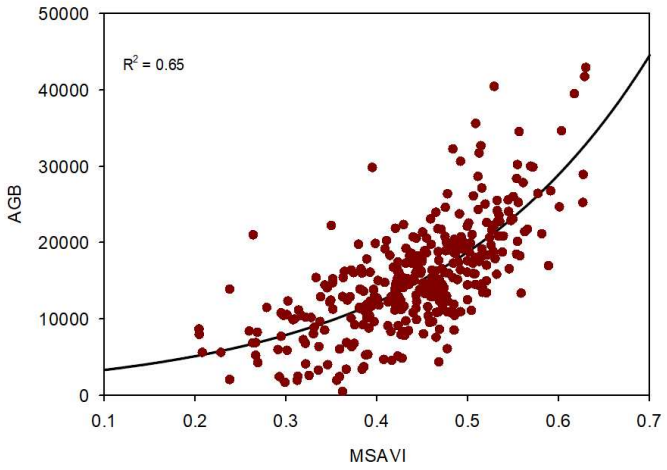

Fig. 3: Relationship between MSAVI and AGB (above ground biomass kg/ha)

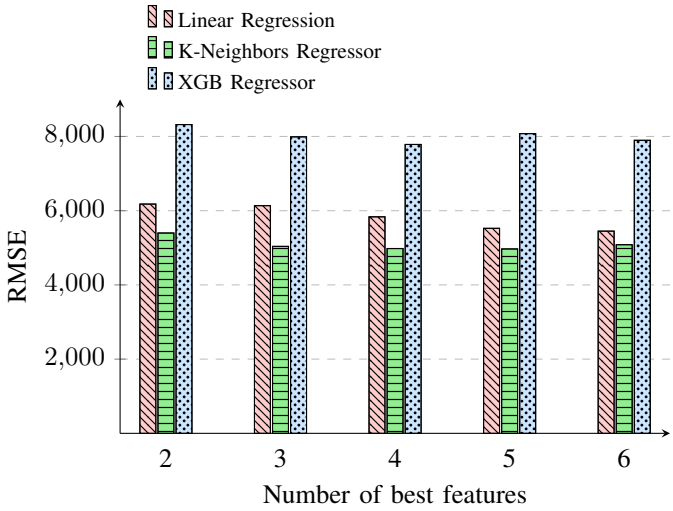

Fig. 4: Comparison of RMSE values for different ML models

V. Conclusion

The results found in this paper substantiate the potential of high-resolution images acquired with UAVs to effectively estimate end-of-season above-ground sorghum biomass. The relationship between vegetation indices and biomass proved to be robust for estimation. The ML algorithms, including LR and KNN, gave good accuracy in predicting biomass despite the small data size. As ground truth data collection will continue to be a bottleneck in plant phenomics, this work opens and highlights interesting key research regarding the choice of the appropriate model when data quantity is limited. In the future, it would be interesting to test other models with limited UAV data. This will be helpful in selecting the appropriate model for low data size, a common issue with in-field phenomic and breeding datasets.

Acknowledgment

We acknowledge the collaboration of investigators from Saint Louis University's Taylor Geospatial Institute (STL, MO, USA) for UAV data collection and the Shakoor Lab team that helped with the collection of field measurements. This work was supported by the USDA (grant no. 2020-67021-31530) and the Salk Institute's Harnessing Plants Initiative (HPI).

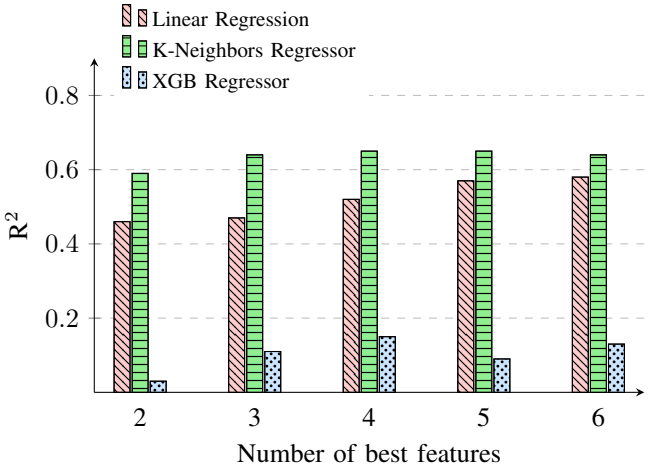

Fig. 5: Comparison of R^2 values for different ML models

REFERENCES

[1] A. Masjedi, M. M. Crawford, N. R. Carpenter, and M. R. Tuinstra, "Multi-temporal predictive modelling of sorghum biomass using uav-based hyperspectral and lidar data," *Remote Sensing*, vol. 12, no. 21, p. 3587, 2020.

[2] B. Gano, J. S. B. Dembele, A. Ndour, D. Luquet, G. Beurier, D. Diouf, and A. Audebert, "Using uav borne, multi-spectral imaging for the field phenotyping of shoot biomass, leaf area index and height of west african sorghum varieties under two contrasted water conditions," *Agronomy*, vol. 11, no. 5, p. 850, 2021.

[3] M. Mbaye, A. Ndour, B. Gano, J. S. B. Dembele, D. Luquet, G. Beurrier, and A. Audebert, "UAV Method Based on Multispectral Imaging for Field Phenotyping," in *Crop Adaptation and Improvement for Drought-Prone Environments* (N. A. Kane, D. Foncéka, and T. J. Dalton, eds.), ch. 7, pp. 171–187, Kansas State University Libraries, Manhattan, Kansas: New Prairie Press, 2022.

[4] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[5] M. Shu, M. Shen, J. Zuo, P. Yin, M. Wang, Z. Xie, J. Tang, R. Wang, B. Li, X. Yang, *et al.*, "The application of uav-based hyperspectral imaging to estimate crop traits in maize inbred lines," *Plant Phenomics*, 2021.

[6] F. J. Ostos-Garrido, A. I. De Castro, J. Torres-Sánchez, F. Pistón, and J. M. Peña, "High-throughput phenotyping of bioethanol potential in cereals using uav-based multi-spectral imagery," *Frontiers in plant science*, vol. 10, p. 948, 2019.

[7] A. T. Tefera, B. P. Banerjee, B. R. Pandey, L. James, R. R. Puri, O. O. Cooray, J. Marsh, M. Richards, S. Kant, G. Fitzgerald, *et al.*, "Estimating early season growth and biomass of field pea for selection of divergent ideotypes using proximal sensing," *Field Crops Research*, vol. 277, p. 108407, 2022.

[8] J. Li, D. P. Schachtman, C. F. Creech, L. Wang, Y. Ge, and Y. Shi, "Evaluation of uav-derived multimodal remote sensing data for biomass prediction and drought tolerance assessment in bioenergy sorghum," *The Crop Journal*, vol. 10, no. 5, pp. 1363–1375, 2022.

[9] E. Habyarimana and F. S. Baloch, "Machine learning models based on remote and proximal sensing as potential methods for in-season biomass yields prediction in commercial sorghum fields," *Plos one*, vol. 16, no. 3, p. e0249136, 2021.

[10] A. Masjedi, N. R. Carpenter, M. M. Crawford, and M. R. Tuinstra, "Prediction of sorghum biomass using uav time series data and recurrent neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[11] Z. Zhang, A. Masjedi, J. Zhao, and M. M. Crawford, "Prediction of sorghum biomass based on image based features derived from time series of uav images," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 6154–6157, IEEE, 2017.

[12] F. Fassnacht, F. Hartig, H. Latifi, C. Berger, J. Hernández, P. Corvalán, and B. Koch, "Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass," *Remote Sensing of Environment*, vol. 154, pp. 102–114, 2014.

[13] K. Prabhakara, W. D. Hively, and G. W. McCarty, "Evaluating the relationship between biomass, percent groundcover and remote sensing indices across six winter cover crop fields in maryland, united states," *International journal of applied earth observation and geoinformation*, vol. 39, pp. 88–102, 2015.

[14] A. Moghimi, C. Yang, and J. A. Anderson, "Aerial hyperspectral imagery and deep neural networks for high-throughput yield phenotyping in wheat," *Computers and Electronics in Agriculture*, vol. 172, p. 105299, 2020.

[15] J. Zhao, M. Karimzadeh, A. Masjedi, T. Wang, X. Zhang, M. M. Crawford, and D. S. Ebert, "Featureexplorer: Interactive feature selection and exploration of regression models for hyperspectral images," in *2019 IEEE Visualization Conference (VIS)*, pp. 161–165, IEEE, 2019.

[16] H. Buxton, "Learning and understanding dynamic scene activity: a review," *Image and vision computing*, vol. 21, no. 1, pp. 125–136, 2003.

[17] M. Herrero-Huerta, P. Rodriguez-Gonzalvez, and K. M. Rainey, "Yield prediction by machine learning from uas-based multi-sensor data fusion in soybean," *Plant Methods*, vol. 16, no. 1, pp. 1–16, 2020.

[18] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[19] J. W. Rouse, R. H. Haas, J. A. Schell, D. W. Deering, *et al.*, "Monitoring vegetation systems in the great plains with erts," *NASA Spec. Publ*, vol. 351, no. 1, p. 309, 1974.

[20] C. R. Perry Jr and L. F. Lautenschlager, "Functional equivalence of spectral vegetation indices," *Remote sensing of environment*, vol. 14, no. 1-3, pp. 169–182, 1984.

[21] A. A. Gitelson and M. N. Merzlyak, "Remote sensing of chlorophyll concentration in higher plant leaves," *Advances in Space Research*, vol. 22, no. 5, pp. 689–692, 1998.

[22] J. Qi, A. Chehbouni, A. R. Huete, Y. H. Kerr, and S. Sorooshian, "A modified soil adjusted vegetation index," *Remote sensing of environment*, vol. 48, no. 2, pp. 119–126, 1994.

[23] E. Barnes, T. Clarke, S. Richards, P. Colaizzi, J. Haberland, M. Kostrzewski, P. Waller, C. Choi, E. Riley, T. Thompson, *et al.*, "Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data," in *Proceedings of the Fifth International Conference on Precision Agriculture, Bloomington, MN, USA*, vol. 1619, p. 6, 2000.

[24] "Pix4D: Professional photogrammetry and drone mapping." https://www.pix4d.com. Accessed: March 7, 2023.

[25] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, 2020.

[26] J.-t. Zhong and S. Ling, "Key factors of k-nearest neighbours nonparametric regression in short-time traffic flow forecasting," in *Proceedings of the 21st International Conference on Industrial Engineering and Engineering Management 2014*, pp. 9–12, Springer, 2015.

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[28] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.