Chiara Rossi[3], Thiago Orlando Costa Barboza[1], Samira Luns Hatum de Almeida[4], Morgan Nicole Sysskind[3], Leticia Moreno[1], Adão Felipe dos Santos[1], Lorena Lacerda[2], George Vellidis[2], Cristiane Pilon[2]

[1]Department of Agriculture, School of Agriculture Sciences of Lavras, Federal University of Lavras (UFLA), Lavras, MG, 37200-900, BR

[2]Department of Crop and Soil Sciences, University of Georgia, 120 Carlton St , Athens, GA, 30602, USA

[3]Department of Crop and Soil Sciences, University of Georgia, 115 Coastal Way, Tifton, GA, 31794, USA

[4]Department of Engineering and Mathematical Sciences, School of Agricultural and Veterinarian Sciences, São Paulo State University (Unesp), Jaboticabal, SP, 14884-900, BR

COMBINING REMOTE SENSING AND MACHINE LEARNING TO ESTIMATE PEANUT PHOTOSYNTHETIC PARAMETERS

A paper from the Proceedings of the
16th International Conference on Precision Agriculture
21-24 July 2024
Manhattan, Kansas, United States

Abstract.

The accurate and efficient assessment of crop health and nutrient status is a critical aspect of modern agriculture, directly influencing management practices and crop yield optimization. This study aims integrate remote sensing data with machine learning techniques to develop robust models for estimating chlorophyll A and B in peanuts. The studied was developed in Southern Georgia, USA, during three years in four fields irrigated and one field non-irrigated. For all the fields the peanut cultivar GEORGIA-06G was planting in different dates. 75 days after planting the second open leaves of mainstem apex of two plants were collected weekly until the plants were inverted. The leaves were replaced in a amber vials containing 5 mL of 95% ethyl alcohol denaturated solution. After this all the leaves were measure and the chlorophyll A and B were determinate. The satellite images from PlanetScope were downloaded weekly with a difference of 2 days due the clouds in the images. Tweny one vegetation index were calculate to predicted the chlorophyll A and B using fur machine learning models: Random Forest, K-Nearest Neighbors, Multilayer Perceptron and Support Vector Machine. The macine learning models were implemented using K-fold cross validation with 5 folds and the GridsearchCV was used to choose

the best hyper parameters for each model. The models were evaluated using Determination Coefficient (R^2), the Root Mean Square Error (RMSE), and all the models and metrics were implemented using Python 3.10 in Google Collab interface. The results revealed that CVI, MSR and CRI were the best vegetation index to detect the chlorophyll in irrigated and dryland fields. The use of the green band presented a high performance of the vegetation index in detected the change. The multilayer perceptron model and K-nearest neighbors were the best machine learning model to predicted the chlorophyll A and B in both fields. However, the low number of the data for dryland fields decrease the quality and performance of the models to predict the chlorophyll content. Different results were found for irrigated fields the high variability in the fields and the highest number of data increase the quality of the machine learning models revealed its potential in predict chlorophyll A and B. The machine learning models when combine remote sensing have a high potential to be apply during the monitoring and for decision making. For this physiology parameters increase the number of data and explore more satellite bands can be a solution to machine learning models present an increase in the performance.

Keywords.
Random Forest, Support Vector Machine, Multilayer perceptron, K-nearest Neighbors, Chlorophyll.

Introduction

The accurate and efficient assessment of crop health and nutrient status is a critical aspect of modern agriculture, directly influencing management practices and crop yield optimization. Among various indicators of plant health, chlorophyll content plays a pivotal role, reflecting the photosynthetic capacity and overall vigor of the crop. Traditional methods of measuring chlorophyll content, such as destructive and non-destructive sampling, and laboratory analysis, are labor-intensive, time-consuming (Qi et al., 2021; Yuan et al., 2022), and often limited in spatial coverage. These limitations necessitate the development of more efficient, non-destructive techniques for large-scale and real-time monitoring of chlorophyll levels.

Remote sensing, coupled with advanced machine learning algorithms, offers a promising solution to this challenge. By utilizing spectral data captured from satellite, aerial, or proximal sensors, it is possible to estimate chlorophyll content over vast agricultural areas with high spatial and temporal resolution. Remote sensing technologies provide information into crop conditions, enabling timely and precise interventions. Remote sensing provides a non-destructive, large-scale, and real-time method for assessing plant health, allowing for frequent monitoring without the need for extensive fieldwork. The high spatial and temporal resolution of remote sensing data enables detailed mapping of chlorophyll distribution across entire fields, facilitating early detection of stress conditions and informing targeted management practices. When coupled with machine learning algorithms, the potential of remote sensing data was further enhanced. Machine learning models can capture complex, non-linear relationships between spectral data and chlorophyll content, outperforming traditional linear regression methods that may not adequately account for such intricacies (Qi et al., 2021). By leveraging vast amounts of multi-dimensional data, machine learning algorithms can improve prediction accuracy and adaptability to varying field conditions. This integration of remote sensing and machine learning thus represents a powerful tool for advancing precision agriculture, optimizing crop management and ultimately increasing yield and sustainability.

In peanut (A*rachis hypogaea* L.) cultivation monitoring chlorophyll content is particularly important due to its direct correlation with the plant's photosynthetic efficiency and nitrogen status (Rao et al., 2001). The chlorophyll content of plants can vary significantly between irrigated and non-irrigated fields that affecting the accuracy of remote sensing-based predictions. Irrigated fields typically exhibit more uniform chlorophyll levels due to consistent water availability (Pilon et al., 2008), which supports optimal photosynthesis and nutrient uptake. In contrast, non-irrigated fields are subject to greater variability in chlorophyll content (Arunyanark et al., 2008), as plants experience water stress, leading to fluctuating photosynthetic activity and nutrient status.

Accurate chlorophyll estimation can assist in optimizing fertilization strategies, thereby enhancing yield and reducing environmental impacts. Despite its significance, there has been limited

research focusing specifically on remote sensing-based chlorophyll assessment in peanut crops. Thus, this study aims to bridge this gap by integrating remote sensing data with machine learning techniques to develop robust models for estimating chlorophyll A and B in peanuts. By leveraging various vegetation indices and spectral features, the proposed approach seeks to enhance the accuracy and reliability of chlorophyll predictions.

Material and methods

The experiment was conducted in Southern Georgia of the United States during a three-year period in five different commercial fields. In 2019, an irrigated field (Field A) and a rainfed field (Field B) were selected. In 2020, an irrigated field (Field C) was chosen, whereas in 2021, two irrigated fields (Field D) and (Field E) were selected. Tift and Berrien Counties are among the largest peanut producing counties in Georgia (USDA NASS, 2022). All fields were planted to the runner-type peanut cultivar Georgia-06G (Branch, 2007). Seeds were treated prior to planting with a fungicide containing azoxystrobin, fludioxonil, and mefenoxam (Dynasty PD®, Syngenta Crop Protection, Greensboro, NC) in 2019 and a fungicide containing ipconazole, carboxin, and metalaxyl (Rancona® V PD, Arysta Life Science, Cary, NC) for the 2020 and 2021 seasons. The seedling rate was 20 seeds per linear meter and a depth of 5.1 centimeters for all fields. Irrigation and chemical applications during the season followed the University of Georgia's Extension recommendations (Monfort et al., 2022).

Table 1. Peanut's fields information.

Field	Year	Planting date	Area (hectares)	Irrigation
A	2019	April 29	10	Yes
B	2019	April 28	4	Not
C	2020	May 19	12	Yes
D	2021	May 1	25	Yes
E	2021	May 10	9	Yes

To analyze field variability, each commercial field was divided according the area. Then, Field A was divided into 24 plots, Field B into 12 plots with 6 subplots each, Field C into 24 plots with 3 subplots each, and fields D and E into 14 plots with 3 subplots each. Subplots were used for collecting the physiological data, and data was combined within each plot.

Photosynthetic pigment content

Leaf samples for chlorophyll a and chlorophyll b content were collected weekly in each field starting 75 days after planting (DAP) until plants were inverted. For each plot, four 6-mm in diameter leaf disks were collected from the second open leaves from the mainstem apex of two different plants and placed in amber vials containing 5 ml of a 95% ethyl alcohol denaturated solution. The vials were stored at a constant temperature of 4 °C for 14 days followed by absorbance readings at 470, 649, and 665 nm wavelengths using a multi-mode microplate reader (Synergy HTX, BioTek, Winooski, VT). Absorbance readings were then used to calculate the contents of chlorophyll A and chlorophyll B (μg cm^{-2}) according to the equations given by Lichtenthaler and Wellburn (1983).

Satellite platform and image processing

Data from the PlanetScope CubeSat platform sensor for the satellite images, product level 3B were used. Characterized with mass approximately 5 kg and with 3U CubeSats form (10 x 10 x 30 cm) (Mukono and Eliot 2022) and temporal resolution of 1 day. The sensors are divided in three, PS2, PS2.SD and PSB.SD, capture different wavelength. To standardized the calculation of vegetation index, the sensor PS2.SD was used, with space resolution 3 m (pixel size). This sensor was improved with a different filter to capture the spectral bands Blue: 464 – 517 nm, green: 547 – 585 nm, red: 650 – 682 nm, NIR: 846 – 888 nm (Planet Team, 2023; Mukono and Eliot, 2022).

The corrections of images, PlanetScope provide the Surface Reflectance product. This product ensures consistency in localized atmospheric conditions, which minimizes uncertainty in temporal

and spatial-spectral response. Surface Reflectance is available for all orthorectified scenes. The Surface Reflectance product was provided as a 16-bit GeoTIFF image with reflectance values scaled to 10,000. The correction of the images is determined from the reflectance of the top of the atmosphere (TOA), and was calculated from the coefficients provided with the Planet Radiance product (Planet Team, 2023).

Vegetation indexes

Twenty-one vegetation indices were calculated (Table 2) for each date and for each studied area. The selection of these indices was based on those that showed the highest correlation with vegetation biomass and chlorophyll content (TAHIR et al. 2018; KLEM et al. 2018; SHANG et al. 2015; WU et al. 2008; CHENG et al. 2022). This correlation was explained by the fact that electromagnetic radiation interacts with vegetation through leaf pigments and in the spongy parenchyma cells with water and gases (JENSEN, 2009; XUE, SU et al. 2017). The amount of energy reflected from this interaction can indirectly indicate whether the plant is stressed or not, allowing the estimation of physiological parameters.

Table 2. Vegetation indexes calculated for chlorophyll A and B analysis in irrigated and dryland fields.

VI	Equation	Reference
Green normalized difference vegetation index (GNDVI)	$\frac{(NIR - Green)}{(NIR + Green)}$	Gitelson e Merzlyak (1996)
Normalized vegetation difference index (NDVI)	$\frac{(NIR - Red)}{(NIR + Red)}$	Rouse et al. (1973)
Soil adjusted vegetation index (SAVI)	$\frac{(NIR - Red)}{(NIR + Red + L)} \times (1 + L)$	Huete (1988)
Chlorophyll vegetation index (CVI)	$\frac{R_{nir}}{R_{green}} \times \frac{R_{red}}{R_{green}}$	Vincini et al. (2008)
Normalized difference green index (NDGI)	$\frac{(R_{780} - R_{550})}{(R_{780} + R_{670})}$	Klem et al. (2018)
Ratio vegetation index (RVI)	$\frac{NIR}{Red}$	Wang e Huang (2010)
Chlorophyll green index (CIg)	$\frac{(NIR)}{(Green)} - 1$	Gitelson et al. 2005
Green Optimal Soil Adjusted Vegetation Index (GOSAVI)	$\frac{(1 + 0.16) \times (NIR - Green)}{(NIR + Green + 0.16)}$	Marin et al. 2022
Visible Atmospherically Resistant Index (VARI)	$\frac{Green - Red)}{(Green + Red - Blue)}$	Gitelson et al. (2022)
Enhanced vegetation index (EVI)	$2.5 \times \frac{(NIR - Red)}{(Nir + 6 \times Red - 7.5 \times Blue + 1)}$	Liu e Huete (1995)
Excess green index (ExG)	$2Green - red - blue$	Wang et al. (2019)
Modified simple ratio (MSR)	$\frac{\left(\frac{NIR}{Red}\right) - 1}{(\left(\frac{NIR}{Red}\right))^{0.5} + 1)}$	Chen (1996)
Renormalized difference vegetation index (RDVI)	$\left(\frac{(NIR - Red)}{\sqrt{Nir + Red}}\right)$	Roujean, Breon (1995)
Green Red Vegetation Index (GRVI)	$\left(\frac{Green - Red}{Green + Red}\right)$	Tucker (1979)
Triangular vegetation index (TVI)	$0.5[120(NIR - green) - 200(NIR + green)]$	Broge; Leblanc (2001)
Modified non-linear index (MNLI)	$\left(\frac{(NIR^2 - Red) \times (1 + 0.5)}{NIR^2 + Red + 0.5}\right)$	Gong et al. (2003)
Excess red index (ExR)	$(1.4 \times Red) - Green$	Meyer; Neto (2008)
Green leaf index (GLI)	$\left(\frac{(2 \times Green) - Red - Blue}{(2 \times Green) + Red + Blue}\right)$	Mathieu et al. (1998)
Modified green-red vegetation index (MGRVI)	$\left(\frac{Green^2 - Red^2}{Green^2 + Red^2}\right)$	Bending et al. (2015)
Modified triangular vegetation index (MTVI)	$\frac{1.5 \times [1.2 \times (NIR - Green) - 2.5 \times (Red - Green)]}{\sqrt{(2 \times NIR + 1)^2 - (6 \times NIR - 5\sqrt{Red}) - 0.5}}$	Haboudane et al. (2004)
Carotenoid reflectance index (CRI)	$\left(\frac{1}{Green}\right) - \left(\frac{1}{NIR}\right)$	Gitelson et al. (2005)

In this context, after calculating the vegetation indices, buffers were created to divide the area into 1 hectare polygons for extracting the average reflectance values of the vegetation indices. These buffers were created within the QGIS 3.28.2 platform using the buffer tool in the vector tab. After creating the buffers for each area, the zonal statistics tool was used to extract the average values

of each vegetation index within each polygon of the created buffer.

Data pre-processing

The final database used to train the models was divided in irrigated and dryland. The number of observation for each area (N) were 120 and 362 data for irrigated and dryland. All analysis were performed in each separated database. The initial step the data were analyzed using boxplot and the outliers values were corrected using the median. The outliers correction were done due the low number of points in each database. The interquartile method calculate the quartiles Q1, Q2, median, Q3 and the interquartile difference (IQR) represented by the box in the boxplot analysis. The upper and lower limited were calculated and the values that are above these limits are outliers and were corrected using the median of the values (Filho et al. 2023).

Feature selection using Random Forest

The feature selection is a technique in select best variables to decrease the number of input in the models. The high number of VI can decreased the generalization process of the models, thus find the best values can accelerate this process presenting the same accuracy and precision. Despite it many methods can be applied to study the relation between two variables and choose the best (Kursa and Rudnicki, 2011). One technique that showed good performance is the Random Forest classification. This technique analysis the input variables and for each node showed the value and the variable that more influence in the model. Thus, the Random Forest model were trained with the input variables (VI) and using a split of 70% and 30%. The feature importance value represent the values obtained in each node and showed the importance of these variables for dependents variables.

Spearman coefficient correlation

The spearman correlation was performed due the non-linear and non-parametric method adopted for analyze the relationship between two variables. The spearman correlation different Pearson's correlation does not assume any distribution frequency. This fact avoid that the coefficient can be used for analyze the correlation between variables that not had a linear behavior (Xiao et al. 2015).

$$r_s = 1 - \frac{6 \sum d_i^2}{N(N^2-1)} \quad (1)$$

Where $d_i^2 = X'_i - Y'_i$ represent the difference between each variables classify and N represent total number of samples.

Machine learning models

Four machine learning algorithms were used to predicted the physiology paramerts in each field (irrigated and dryland). The machine learning models used were Random Forest (RF), Support vector Machine (SVM), K-Nearest Neighboors (KNN) and Multlayer Perceptron (MLP). For all models the input variables were normalized using the minimum and maximum values (0 – 1). After this the hyper-paramerters of each model were tested and choose use the GridSearchCV algorithm and all the models were trained using K-fold cross-validation (CV = 5).

Hyper-parameters tunning using GridSearchCV

The GridSearchCV combine the hyper-parameters showed and choose the best parameters for the models according with the best results of determinations coefficient (R^2) (Pedregosa et al. 2011). The parameters used for the GridSearchCV were estimator (machine learning algoritm), parameters grid the machine learning parameters that will be test, scoring (evaluation metric) in this case R^2 and cross-validation (CV) with 5 folds (Pedregosa et al. 2011).

Random Forest (RF)

The Random Forest (RF) is a collection of decision tree classifiers created using two random

sources: bootstrap and sampling (Bentéjac et al., 2021). In bootstrap, the trees are trained with random samples where the original data are replaced by samples of the same size as the training set. In sampling, each split node is trained with a small random subset of the input data to determine the best split (Bentéjac et al. 2021). RF proposed by Breiman (2001) can be used for classification tasks with categorical variables and regression tasks with continuous responses (Cutler et al. 2012). In regression tasks a p-dimensional random vector $X = (X1, ..., X_p)^T$ represents the real input values and Y represents the real output values. The goal is to find an estimator function *f*(x) that predicts the values of Y. This function is determined by the cost function L(Y, f(X)), which is defined to minimize the expected loss $E_{X,Y}(L(Y, f(X)))$ (Cutler et al., 2012). The cost function measures how close *f*(X) is to Y, penalizing values that are far from *f*(X) using as penalization the mean squared error (Cutler et al. 2012).

Support Vector Machine (SVM)

The Support Vector Regression (SVR) is a type of supervised learning applied to regression studies to analyze the relationship between the dependent variable and the predictor variables (Zhang and O'donnell, 2020). The basic idea of this algorithm is similar to that of SVC where the machine with spaced kernels uses a hyperplane defined by the support vectors to classify the data (Zhang and O'DONNELL, 2020). However, in regression, the support vectors perform the optimization which does not depend on the dimension of the data but rather on the number of support vectors (Zhang and O'donnell, 2020). In summary, instead of fitting the widest margin possible between the two classes, as in classification, SVR fits as many instances as possible within the margins (support vectors and decision) defined by the hyper parameter ε that regulates the width between the support vectors and the decision (Geron, 2021).

Multilayer Perceptron (MLP)

MLP networks are part of a large group known as feedforward neural networks, which are applied in various models and optimization problems. The architecture is mainly divided into input, hidden, and output layers. The hidden layer contains the neurons (Ghorbani et al. 2016). Each neuron receives an activation function that processes the input data and passes it to the neurons in the next layer (if there is one), until it reaches the output layer, which estimates the values (Ghorbani et al. 2016).

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) regression algorithm is a non-parametric method that uses observed data to estimate the predictor variable in real time without defining any parametric relationship (Modaresi et al. 2018). To estimate values, the algorithm analyzes the neighborhood and calculates the relationship between neighbors using the Euclidean distance function. The weights were adjusted manually, and the estimated values are calculated using a probabilistic function. The kernel function applied to the K observed data points closest to the evaluated point are the main factors that affect the model's adjustment. This parameter need be adjust during the train data and can be test using GridsearchCV.

Evaluation metrics

The statistical measure of accuracy applied was the root mean square error (RMSE) (Equation 1). As an assessment of precision, the values of the coefficient of determination (R^2) (Equation 2) were calculated for each combination of vegetation indices.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{obs}-y_{est})^2}{n}} \quad (2)$$

Where RMSE is the Root Mean Square Error, y_{obs} is the observed values, y_{est} the estimated values and n the number of samples.

$$R^2 = \frac{\sum_{i=l}^{N}(Yest_i-\bar{Y})^2}{(Yobs_i-\bar{Y})^2} \quad (3)$$

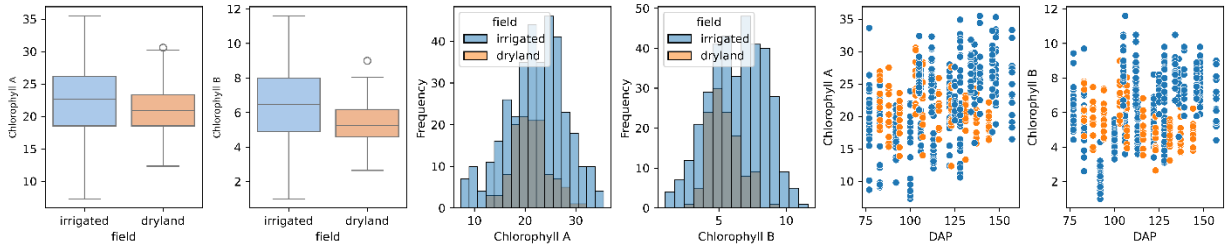
Where R^2 represent the determination coefficient and Y the mean values of the dataset collect in the field. All the models were ran in Python 3.10 using the Google Collab interface. The mainly libraries using for this present work were Scikit-Learn 1.5.0 for machine learning models, Seaborn 0.13.2 and Matplotlib 3.9 for data visualization.

Results and discussions

The descriptive analysis using boxplot, histogram and scatterplot graphs are presented in the figure 1. In the first moment the chlorophyll A and B for both fields showed outliers values. The outlier's values represent high values observed in the fields and due the low number of evaluations in both dataset the outlier's values were corrected using the median.

The chlorophyll pigments are important organelles for photosynthesis. The production by the plants depends on the environment and nutrition effects that control the high or low production in plants leaves. These effects on the production mainly in dryland fields revealed the importance of water availability which low values were reported. The low water availability (30% field capacity) also was reported in different peanut genotypes revealed that a decrease in total chlorophyll of 8 mg plant-1 to 4 mg plant-1(Arunyanark et al. 2008). This same effect was reported to Khatri and Rathore (2022) which the water availability combines with soil salinity limited the root length, biomass accumulation, efficiency photosynthetic and pigments production (Zeid and Shedeed, 2006; Fu and Huang, 2001).

Figure 1. Descriptive analysis for chlorophyll A and B in irrigated (blue labels) and dryland (orange labels) fields.

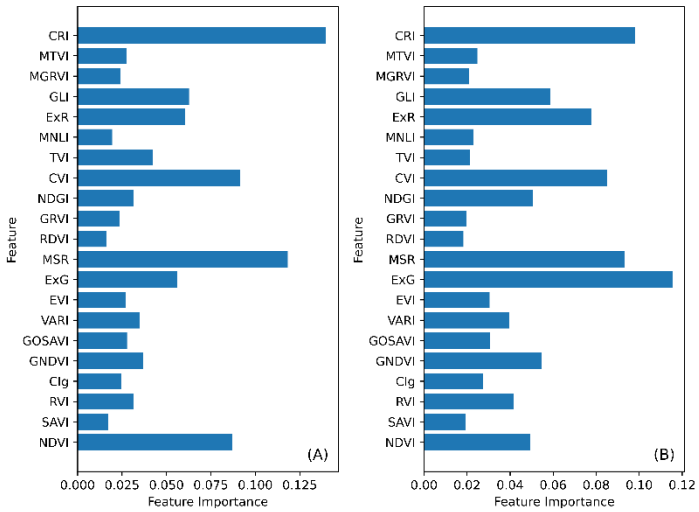


Despite the influence of the water availability in the pigments production measure these pigments and obtain accurate results involve a difficult process. The conventional methods analysis merely few leaves of the plants not considering the variability of the field production. The introduction of machine learning techniques and remote sensing authors reported good results in predict foliar pigments mainly using drone images (Qi et al. 2021; Balota et al. 2024; Monsef et al. 2019).

The remote sensing became a widely tool used in all field studies. The interaction of different wavelength of spectrum electromagnetic might help the researchers, farmers and extensions in monitoring the field conditions and predict the physiology parameters as well predict the yield. Especially in foliar pigments, the chlorophyll A and B presented a different behavior front the different wavelengths. The visible region the chlorophyll A and B presented a tendency to reflect in the green regions and a high absorbance at red region (Gitelson and Merzlyak, 2010; Gitelson et al. 2005; Gitelson et al. 2020, Gitelson, Gritz and Merzlyak, 2003). The other most use for analysis is the near infrared (NIR) (Gitelson et al. 2020; Gitelson et al. 2005).

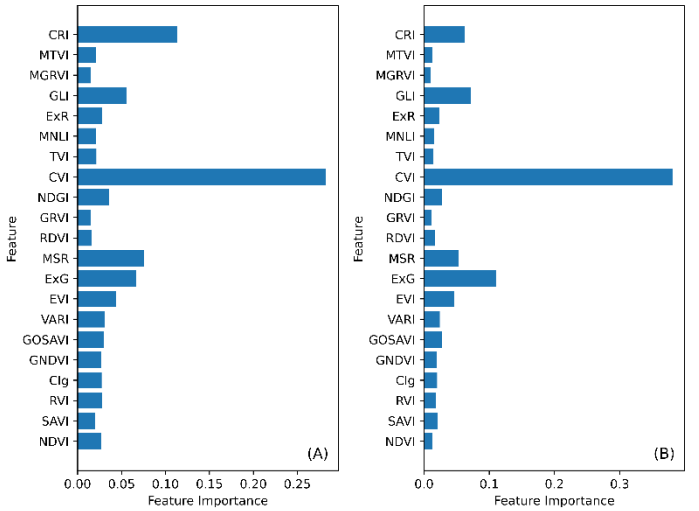
One of the most application of the remote sensing is the use of vegetation index (VI). This VI's combine different wavelengths and showed the photosynthetic status of the plants (Gitelson et al. 2020). The large number of VI revealed the importance of analysis and figure out the best VI for each situation. To select the VI the machine learning techniques showed a good performance mainly when bagging techniques are apply (Kursa and Rudnicki, 2011). In this study the feature selection using RF for dryland fields is present in the Figure 2.

Figure 2 . Feature importance using random forest algorithm apply for chlorophyll A graph (A) and B graph (B) in dryland fields



The four important VI obtained during the feature selection process for chlorophyll A (Figure 2A) were CRI, MSR, CVI and NDVI, while that the chlorophyll B (Figure 2B) the high values were observed for ExG, CRI, MSR and CVI. Amount this VI the CRI, MSR and CVI were the selected VI that appeared for both chlorophyll parameters in dryland fields. Therefore, when analysis in irrigated fields only the VI CVI presented a high feature importance value for both parameters' chlorophyll A and B (Figure 3). Despite this, for irrigated fields the four-best VI selected by RF algorithm to chlorophyll A (Figure 3A) were CVI, CRI, MSR and ExG, while the chlorophyll B (Figure 3B) the VI selected were CVI, EXG, GLI and CRI. These VI selected were like observe to dryland fields were similar mainly the CVI, CRI and MSR that showed in all feature importance analysis.

Figure 3. Feature importance using random forest algorithm apply for chlorophyll A graph (A) and B graph (B) in irrigated fields.

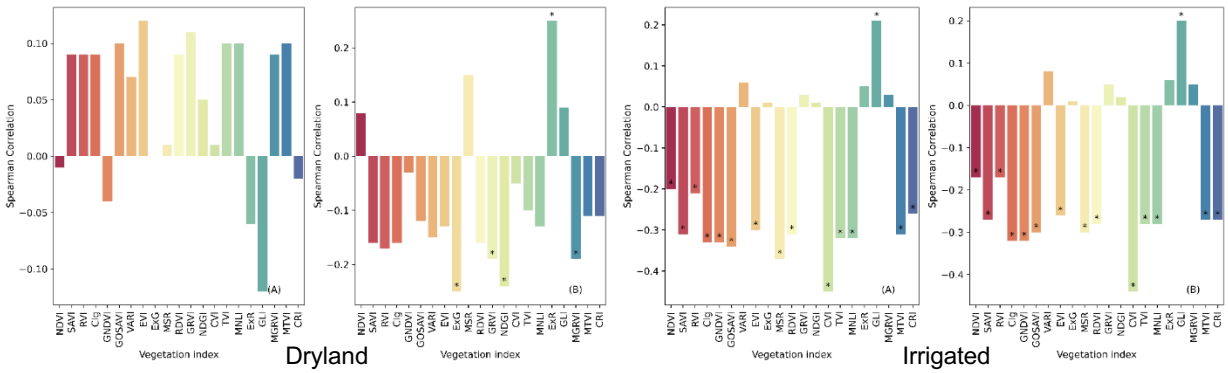


The CRI, ExG and CVI use the green band to calculate the VI and this band was described as fundamental wide studies using remote sensing (Balota et al. 2024; Monsef et al. 2019; Gitelson et al. 2020; Gitelson et al. 2005; Gitelson and Merzlyak, 2010). The green band is associated with chlorophyll content due the high reflectance sensitivity capture by this band mainly when the leaves change the colors for green to dark green or yellow (Kira, Linker and Gitelson, 2015; Merzlyak and Gitelson, 1995). The chlorophyll had a peak of absorbance in red and blue wavelengths, although, the blue regions showed an overlap with carotenoids pigments and the low concentrations of chlorophyll content saturated the red region (Sims and Gamon, 2002). Thus,

the 500 and 700 nanometers (nm) regions presented the best regions to estimate this pigments mainly to saturate this region is necessary higher chlorophyll concentration (Sims and Gamon, 2002). The PlanetScope green band showed the center wavelength of 531 nm regions next to describe by the other authors as the ideal for chlorophyll analysis (Sims and Gamon, 2002; Gitelson et al. 2005; Kira et al. 2015). The ExG despite using the green band does not use the NIR band as the CRI and CVI (Wang et al. 2019). The VI combine the visible wavelengths and enhance the values in the green bands to be capture in analysis (Wang et al. 2019). Similarly, the observed for the ExG was observed for GLI equation that the main difference between the VI is the GLI apply the normalization of the values increase the understand of the reflectance values (Mathieu et al. 1998). The other VI MSR and NDVI combine the red and NIR bands of the spectrum to understand the reflectance values in the fields. The MSR VI was developed based in RDVI to minimize the environment effects on the values (Chen, 1996). The equation values when modify increase the sensibility to biomass and chlorophyll become this VI more sensitive than RDVI (Chen, 1996; Wu et al. 2008). This vegetation index showed a non-linear relationship between the red and NIR bands that increase the efficiency in capture the differences in the fields (Chen, 1996). The NDVI showed similar behavior like MSR, the NDVI combine the red and NIR bands and apply the normalization of the reflectance values. These two bands represent importante regions of the spectrum to detect stress in plants (Gitelson et al. 2005). The red band showed a low reflectance value while the NIR presented a high reflectance values in health plants. In stress conditions the photosynthetic apparatus, gas exchange and water absorption decreased and the reflectance in red band increased and NIR band decreased (Gitelson et al. 2005; Sims and Gamon, 2002; Rouse et al. 1975).

After the feature selection using the RF the Spearman correlation was performed. In the dryland fields the correlation values observed for ExG, CVI, MSR, GLI, NDVI and CRI for chlorophyll A were 0, 0.01, 0.01, -0.12, -0.01 and –0.02 while for chlorophyll B the result increase with exception for GLI and the observed values were –0.25, -0.05, 0.15, 0.08, 0.09, -0.11.

Figure 4. Spearman correlation analysis applied for chlorophyll A graph (A) and B graph (B) in dryland and irrigated fields. *not significant (0.05).



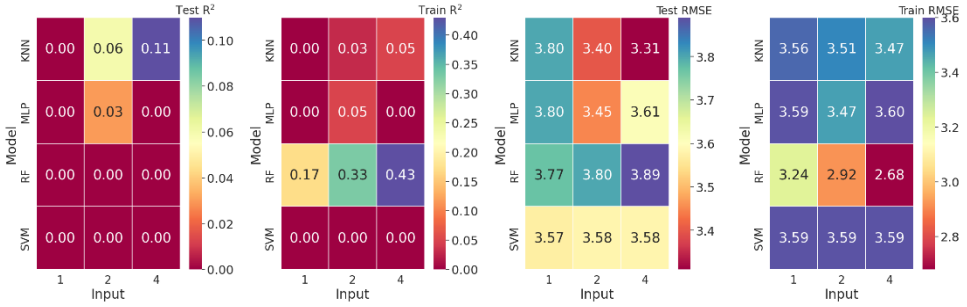
The correlation results for irrigated fields showed the highest values. For the same variables ExG, CVI, MSR, GLI, NDVI and CRI the chlorophyll A (Figure 4A) presented the values 0.01, -0.45, -0.37, 0.21, -0.20 and -0.26. The chlorophyll B (Figure 4B) irrigated the analysis showed similar results with the values of 0.01, -0,44, -0.30, 0.20, -0.17 and -0.27.

The feature selection using RF and the spearman correlation presented different results for both chlorophyll A and B. The high values observe for dryland fields were GLI, EVI, GRVI and GOSAVI, while the chlorophyll B the best results were found for ExG, ExR, NDGI and GRVI. The irrigated fields the best results found for chlorophyll A were CVI, MSR, GOSAVI, CIg and for chlorophyll B CVI, CIg, GNDVI and GOSAVI. The spearman correlation is a nonparametric function that does not assume any frequency. This coefficient is characterized as an arbitrary monotic function that try to describe the relationship between two variables (Hauke and Kossowski, 2011).

The results obtained for dryland fields were worse than irrigated fields. The chlorophyll A (Figure 5) presented low RMSE and R^2 values for all machine learning models in different input

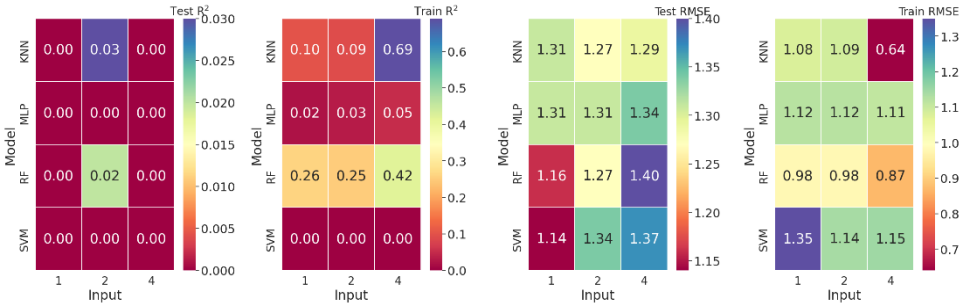
combinations. Despite the train showed 0.43 R^2 for RF model the test values were 0 with high RMSE values. When used two input variables (CRI and MSR) the KNN and MLP showed a low result. Similar results were observed for chlorophyll B in dryland fields (Figure 6) where the two inputs (ExG and CRI) presented low values of R^2 0.02 and 0.03 for RF and KNN respectively.

Figure 5. Determination Coefficient (R^2) and Root Mean Square Error (RMSE) of four machine learning models to predict chlorophyll A in dryland fields.



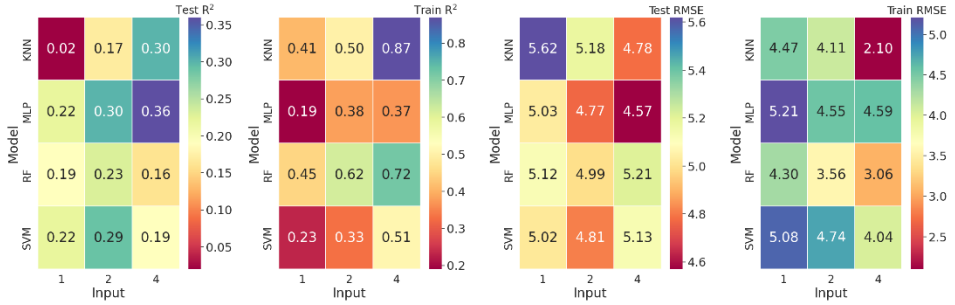
The results in the train for both variables in the dryland fields showed good results. However the model copied the data and the train model when applied for a new data (test) the results decreased drastically. This pattern in copied the data and not generalize the results in a new data frame describe the overfitting of the machine learning models. The overfitting describe the problem when low number of data was used to train the machine learning models. The dryland fields exhibited 120 data points and the division of the data in 70% and 30% (train and test) reduce the dataset for 84 data points for train and 36 for test. Despite the low values found in R^2 during the train and test for dryland areas, in other studies high values of R^2 0.82 and 0.95 and low values of RMSE were observed for RF and SVM models using hyperspectral cameras in UAV for predict the chlorophyll content (Ta, Chang and Zhang, 2021). Narmilan et al. (2022) predict sugarcane chlorophyll content using UAV obtained the best R^2 values 0.96, 0.95 and 0.94 for XGBoosting, RF and Decision Tree in the validation data. Despite the work showed low values for dryland fields, the present study used satellite images that have a high applicably in big fields increase the agility in monitor the chlorophyll. More specifically the above mention studies tried to predict the chlorophyll content different from this studied that tried to predict the chlorophyll A and B different pigments found in the plants.

Figure 6. Determination Coefficient (R^2) and Root Mean Square Error (RMSE) of four machine learning models to predict chlorophyll B in dryland fields.



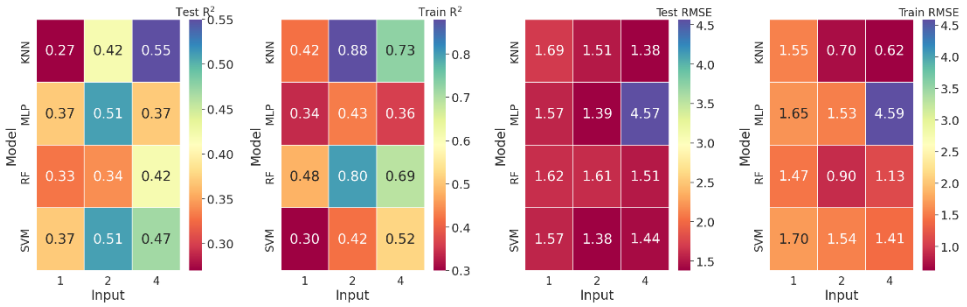
The results observed for irrigated fields were higher than dryland fields. The chlorophyll A (Figure 7) high R^2 values 0.87, 0.72 and 0.62 were observed for KNN and RF models when four (CVI, CRI, MSR and ExG) and two (CVI, CRI) inputs variables were used. The high values for train for these models showed a low values for test data presenting the overfitting problem. Except for MLP model that showed low values for train and test (0.37 and 0.36) the model not presented overfitting but the accuracy values (R^2) revealed the low generalization of the model. Change the number of inputs from four to one decrease the ability of the model in analysis and learned with the pattern of the data. The RMSE values for chlorophyll A (Figure 7) presented for test values high values 5.62 and low values 4.57 for KNN and MLP. These result revealed that MLP model has the best result in precision with an error of 13% between for RMSE.

Figure 7. Determination Coefficient (R^2) and Root Mean Square Error (RMSE) of four machine learning models to predict chlorophyll A in irrigated fields.



For irrigated fields the chlorophyll B (Figure 8) showed high R^2 values 0.88, 0.80 when combine two variables (CVI and ExG) and 0.73 when combine four input variables (CVI, ExG, GLI and CRI). Despite the best results for train the train model when applied for test dataset decrease the values. The best results for R^2 0.55 and 0.51 were found for KNN, MLP and SVM, however the decreased in 25% of the precision revealed that the model had overfitting and the SVM and MLP presented the best performance. The low results found in this work were different for obtained when the UAV multispectral and hyperspectral images were used (Narmilan et al. 2022; Wang et al. 2022). Detected the chlorophyll A and B showed that a specifically strip of the electromagnetic spectrum need be reach to find good results. Despite the chlorophyll B not showed a good performance in dryland fields its results were better than chlorophyll A. In eucalyptus leaves the 550 and 708 nm were describe as the highest wavelengths for chlorophyll A detection, while the chlorophyll B the best results were found better when combine R672/R550. This fact occur due the R550 be close the blue band (R550) and the use of red bands not showed any difference except for chlorophyll A that present a peak of absorption in R672 (Datt, 1998). In soybean the peak of absorbance were found in 460 and 650 nm for chlorophyll B and the chlorophyll A the best results were found in 440, 580, 630 and 670 nm (Chappelle, Kim and McMurtrey 1992). Found the optimum wavelength is a difficult process in multispectral image due the big length of the bands. Thus analysis the channel and change due the big length can be a solution to increase the capacity of the model in detect the less differences, specially because the NIR bands does not showed a high impact in the chlorophyll detection as reported by the authors (Datt, 1998; Chappelle, Kim and McMurtrey 1992)

Figure 8. Determination Coefficient (R^2) and Root Mean Square Error (RMSE) of four machine learning models to predict chlorophyll B in irrigated fields.



Machine learning models had a high performance in identify the pattern and generalize the results in an unknown dataset. The different methods showed different results what showed the importance in used more than one model to test the dataset. The KNN and MLP the best model found in this work are describe as non-parametric and parametric models. The non-parametric models not used any distribution frequency and it does not show any fix parametrization function that describe the model. The parametric model (MLP) present fixed parameters (number of neuron, layers and outputs) but these parameters are adjust before the train of the model. The MLP neural network architecture used in this work present a high performance using 50 neuron in one hidden layer. To improve the performance the best choose method for train the model were LBFGS and this method show a high performance in high dimension dataset that converge for a

minimum local faster. Similar results using this algorithm were report for Barman et al. 2021 where the MLP reach R^2 values of 0.802 using hyperspectral images and Spad sensor. The findings of this research were expected to contribute to the advancement of precision agriculture practices, offering a scalable and efficient tool for monitoring crop health. As agricultural challenges continue to evolve, the adoption of innovative technologies such as remote sensing and machine learning becomes imperative for sustainable and productive farming.

Conclusion

The study tried to find the best machine learning models combine different number of input variables and the MLP showed the best results for irrigated and dryland fields. The dryland fields presented low number of data decreased significantly the accuracy of the models. Despite this the CVI, CRI and MSR were described as the main VI to predict the chlorophyll A and B. Furthermore the results of this work showed that increase the number of data for dryland fields the accuracy and precision increase. Increasing these values the satellite images can become a new potential tool to monitor the fields for chlorophyll A and B, improve the management practices an understand the effects of the variability in the plant physiology process.

References

Arunyanark, A., Jogloy, S., Akkasaeng, C., Vorasoot, N., Kesmala,T., Nageswara Rao, R.C., Wright,, G.C.A. (2008). *A. Patanothai* Chlorophyll stability is an indicator of drought tolerance in peanut. *Journal Agron. Crop Sci.,* 194, 113-125.

Balota, M., Sarkar, S., Bennett, R. S., Burow, M. D. (2024). Phenotyping peanut drought stress with aerial remote-sensing and crop index data. *Agriculture*, doi: https://doi.org/10.3390/agriculture14040565

Barman, U., Sarmah, A., Sahu, D., Barman, G. G. (2021). Estimation of tea leaf chlorophyll using mlr, ann, svr, and knn in natural light condition. *Proceedings of the International Conference on Computing and Communication Systems*, doi: https://doi.org/10.1007/978-981-33-4084-8_27

Bending, J. et al. (2015). Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *International Journal of Applied Earth Observation and Geoinformation*, doi: https://doi.org/10.1016/j.jag.2015.02.012

Bentéjac, C., Csorgo, A., Martinez-munoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, doi: https://doi.org/10.1007/s10462-020-09896-5

Branch, W. D. (2007). Registration of 'Georgia-06G' peanut. *Journal of Plant Registrations*, doi: https://doi.org/10.3198/jpr2006.12.0812crc

Broge, N. H., Leblanc, E. (2001). Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. *Remote Sensing of Environment*, doi: https://doi.org/10.1016/S0034-4257(00)00197-8

Chapelle, E. W., Kim, M. S., McMurtey, J. E. (1992). Ratio analysis of reflectance spectra (RARS): An algorithm for the remote estimation of the concentrations of chlorophyll A, chlorophyll B, and carotenoids in soybean leaves. *Remote Sensing of Environment,* doi: https://doi.org/10.1016/0034-4257(92)90089-3

Chen, J. M. (1996). Evaluation of vegetation indices and a modified simple ratio for boreal applications. Canadian Journal of Remote Sensing, doi: https://doi.org/10.1080/07038992.1996.10855178

Cheng, J. et al. (2022). Estimating canopy-scale chlorophyll content in apple orchards using a 3D radiative transfer model and UAV multispectral imagery. *Computers and Electronics in Agriculture*, doi: https://doi.org/10.1016/j.compag.2022.107401

Cutler, A., Cutler, D. R., Stevens, J. R. (2012). Random forests. *Ensemble Machine Learning*, doi:

https://doi.org/10.1007/978-1-4419-9326-7_5

Datt, B. (1998). Remote sensing of chlorophyll a, chlorophyll b, chlorophyll a+b, and total carotenoid content in eucalyptus leaves. *Remote Sensing of Environment,* doi: https://doi.org/10.1016/S0034-4257(98)00046-7

Filho, D. F., Silva, L., Pires, A., Malaquias, C. (2023). Living with outliers: How to detect extreme observations in data analysis. *ANPOCS*, 99, 1-24.

Fu, J., Huang, B. (2001). Involvement of antioxidants and lipid peroxidation in the adaptation of two cool-season grasses to localized drought stress. Environmental and Experimental Botany, doi: https://doi.org/10.1016/S0098-8472(00)00084-8

Géron, A. Hands on: Machine Learning with Scikit-Learn, Keras & TensorFlow. O'REILY Media Inc. 2nd edition.

Ghorbani, M. A., Zadeh, H. A., Isazadeh, M., Terzi, O. (2016). A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction. *Environ Earth Sci*, doi: 10.1007/s12665-015-5096-x

Gitelson, A. A. (2020). Towards a generic approach to remote non-invasive estimation of foliar carotenoid-to-chlorophyll ratio. *Journal of Plant Physiology*, doi: https://doi.org/10.1016/j.jplph.2020.153227

Gitelson, A. A. et al. (2005). Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters*, doi: https://doi.org/10.1029/2005GL022688

Gitelson, A. A., Kaufman, Y. J., Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, doi: https://doi.org/10.1016/S0034-4257(96)00072-7

Gitelson, A. A., Merzlyak, M. N. (2010). Remote estimation of chlorophyll content in higher plant leaves. *International Journla of Remote Sensing*, doi: https://doi.org/10.1080/014311697217558

Gitelson, A. A., Stark, R., Grits, U., Rundquist, D., Kaufman, Y., Derry, D. (2022). Vegetation and soil lines in visible spectral space: A concept and technique for remote estimation of vegetation fraction. *International Journal of Remote Sensing*, doi: https://doi.org/10.1080/01431160110107806

Gitelson, A. A., Viña, A., Verónica, C., Rundquist, D. C., Arkebauer, T. J. (2005). Remote sensing of canopy chlorophyll content in crops. *Geophysical Research Letters*, doi: https://doi.org/10.1029/2005GL022688

Gong, J., Pu, R., Biging, G. S., Larrieu, M. R. Estimation of forest leaf area index using vegetation indices derived from Hyperion hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, doi: https://doi.org/10.1109/TGRS.2003.812910

Haboudane, D., Miller, J. R., Pattey, E., Tejada, P. Z., Stracha, I. B. (2004). Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environement*, doi: https://doi.org/10.1016/j.rse.2003.12.013

Hauke, J., Kossowski, T. (2011). comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quastiones Geographicae*, doi: 10.2478/v10117-011-0021-1

Khatri, L., Rathore, M. S. (2022). Salt and osmotic stress-induced changes in physio-chemical responses, PSII photochemistry and chlorophyll a fluorescence in peanut. *Plant Stress,* doi: https://doi.org/10.1016/j.stress.2022.100063

Kira, O., Linker, R., Gitelson, A. (2015). Non-destructive estimation of foliar chlorophyll and carotenoid contents: Focus on informative spectral bands. *International Journal of Applied Earth Observation and Geoinformation,* doi: https://doi.org/10.1016/j.jag.2015.01.003

Klem, K. et al. (2018). Interactive effects of water deficit and nitrogen nutrition on winter wheat. Remote sensing methods for their detection. *Agricultural Water Management*, doi: https://doi.org/10.1016/j.agwat.2018.08.004

Kursa, M. B., Rudnicki, W. R. (2011). The All Relevant Feature Selection using Random Forest. Cornell University, doi: https://doi.org/10.48550/arXiv.1106.5112

Lichtenthaler, H., Wellburn, A. (1983). Determinations of Total Carotenoids and Chlorophylls a and b of Leaf Extracts in Different Solvents. *Biochemical Society Transactions*, http://dx.doi.org/10.1042/bst0110591

Liu, H. Q., Huete, A. (1995). A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. IEEE Transactions on Geoscience and Remote Sensing, doi: https://doi.org/10.1109/TGRS.1995.8746027

Marin, D. B. et al. (2021). Detecting coffee leaf rust with UAV-based vegetation indices and decision tree machine learning models. *Computers and Electronics in Agriculture*, doi: https://doi.org/10.1016/j.compag.2021.106476

Mathieu, R., Pouget, M., Cervelle, B., Escadafal, R. (1998). Relationships between satellite-based radiometric indices simulated using laboratory reflectance data and typic soil color of an arid environment. *Remote Sensing of Environment*, doi: https://doi.org/10.1016/S0034-4257(98)00030-3

Merzlyak, M. K., Gitelson, A. (1995). Why and what for the leaves are yellow in autumn? on the interpretation of optical spectra of senescing leaves (*Acerplatanoides* l.). *Journal of Plant Physiology*, doi: https://doi.org/10.1016/S0176-1617(11)81896-1

Meyer, G. E., Neto, J. C. (2008). Verification of color vegetation indices for automated crop imaging applications. *Computers and Electronics in Agriculture*, doi: https://doi.org/10.1016/j.compag.2008.03.009

Modaresi, F.; Araghinejad, S. Ebrahimi, K. (2018). A Comparative Assessment of Artificial Neural Network, Generalized Regression Neural Network, Least-Square Support Vector Regression, and K-Nearest Neighbor Regression for Monthly Streamflow Forecasting in Linear and Nonlinear Conditions. *Water Resources Management*, doi: https://doi.org/10.1007/s11269-017-1807-2

Monfort, S. et al. (2022). Peanut production guide. Tifton, GA, University of Georgia Extension.

Monsef, H. A., Smith, S. E., Rowland, D. L., Rasol, N. A. (2019). Using multispectral imagery to extract a pure spectral canopy signature for predicting peanut maturity. *Computers and Electronics in Agriculture*, doi: https://doi.org/10.1016/j.compag.2019.04.028

Mukono, J., Eliot, K.(2022). Our constellation: PlanetScope. PlanetScope.

Narmilan, A. et al. (2022). Predicting canopy chlorophyll content in sugarcane crops using machine learning algorithms and spectral vegetation indices derived from uav multispectral imagery. *Remote Sensing*, doi: https://doi.org/10.3390/rs14051140

Pedregosa, F. et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning*, doi: https://scikit-learn.org/stable/index.html

Pilon, C., Snider, J. L., Sobolev, V., Chastain, D. R., Sorensen, R. B., Meeks, C. D., Massa, A. N., Walk, T., Singh, B., Earl, H. J. (2018). Assessing stomatal and non-stomatal limitations to carbon assimilation under progressive drought in peanut (Arachis hypogaea L.). Journal of Plant Physiology, 231, 124-134.

Planet Team. (2023). Planet imagery product specifications. Planet Labns, PBC.

Qi, H. et al. (2020). Hyperspectral inversion model of chlorophyll content in peanut leaves. *Applied Science*, doi: https://doi.org/10.3390/app10072259

Qi, H. et al. (2021). Monitoring of peanut leaves chlorophyll content based on drone-based multispectral image feature extraction. *Computers and Electronics in Agriculture*, doi:

https://doi.org/10.1016/j.compag.2021.106292

Rao, N.R., Talwar, H., Wright, G. (2001). Rapid assessment of specific leaf area and leaf nitrogen in peanut (*Arachis hypogaea* L.) using a chlorophyll meter. *J. Agron. Crop Sci.186*, 175–182.

Roujean, J. L., Breon, F. M. (1995). Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment*, doi: https://doi.org/10.1016/0034-4257(94)00114-3

Shang, J. et al. (2015). Mapping spatial variability of crop growth conditions using RapidEye data in Northern Ontario, Canada. *Remote Sensing of Environment*. Doi: http://dx.doi.org/10.1016/j.rse.2015.06.024

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

Ta, N., Chang, Q., Zhang, Y. (2021). Estimation of apple tree leaf chlorophyll content based on machine learning methods. *Remote Sensing*, doi: https://doi.org/10.3390/rs13193902

Tahir, M. N. et al. (2018). Real time estimation of chlorophyll content based on vegetation indices derived from multispectral UAV in the Kinnow orchard. Int. J. Precis. Agric. doi: 10.33440/j.ijpaa.20180101.0001

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, doi: https://doi.org/10.1016/0034-4257(79)90013-0

United States Department of Agriculture and National Agricultural Statistica Service (USDA NASS). (2022). 2022 census of agriculture. USDA NASS.

Vincini, M., Frazzi, E., D'alessio, P. (2008). A broad-band leaf chlorophyll vegetation index at the canopy scale. Precision Agriculture, doi: https://doi.org/10.1007/s11119-008-9075-z

Wang, A., Zhang, W., Wei, X. (2019). A review on weed detection using ground-based machine vision and image processing techniques. Computers and Electronics in Agriculture, doi: https://doi.org/10.1016/j.compag.2019.02.005

Wang, F., Hang, J. (2010). Development of a vegetation index for estimation of leaf area index based on simulation modeling. Journal of Planting Nutrition, doi: 10.1080/01904160903470380

Wang, T. et al. (2022). Winter wheat chlorophyll content retrieval based on machine learning using in situ hyperspectral data. *Computers and Electronics in Agriculture,* doi: https://doi.org/10.1016/j.compag.2022.106728

Wu, C., Niu, Z., Tang, Q., Huang, W. (2008). Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. *Agricultural and Forest Meteorology*, doi: https://doi.org/10.1016/j.agrformet.2008.03.005

Xiao, C., Ye, J., Esteves, R. M., Rong, C. Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation Practique and Experience*, doi: https://doi.org/10.1002/cpe.3745

Xue, J., Su, B. (2017). Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *Journal of Sensors*, doi: https://doi.org/10.1155/2017/1353691

Yuan Y., Wang X., Shi M., Wang P. (2022) Performance comparison of RGB and multispectral vegetation indices based on machine learning for estimating *Hopea hainanensis* SPAD values under different shade conditions. *Front. Plant Sci*. 13:928953.

Zeid, I. M., Shedeed, Z. A. (2006). Response of alfalfa to putrescine treatment under drought stress. *Biologia Plantarum*, doi: 10.1007/s10535-006-0099-9

Zhang, F.; O'donell, L. J. (2020). Chapter 7 - Support Vector Machine. *Machine Learning*, doi: https://doi.org/10.1016/B978-0-12-815739-8.00007-9