ICPA
17th INTERNATIONAL CONFERENCE ON PRECISION AGRICULTURE
ConBAP 2026
CONGRESSO BRASILEIRO DE AGRICULTURA DE PRECISÃO E DIGITAL
13-16 July 2026 | Porto Alegre, Brazil

Assessment of Machine Learning Models for Leaf Chlorophyll Estimation Using Visible-Range Reflectance

Thiago Orlando Costa Barboza[1], Wender Henrique Batista da Silva[1], Silvino Guimarães Moreira[1], Sérgio Henrique Godinho Silva[1], Lorena Nunes Lacerda[2], Adão Felipe dos Santos[1]

[1]Lavras School Agricultural Science (ESAL), Federal University of Lavras (UFLA)

[2]Integrative Precision Agriculture, Crop and Soil Science Department, University of Georgia

A paper from the Proceedings of the

17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture

13-15 July 2026

Porto Alegre, Brazil

Abstract.

Chlorophyll content plays a central role in the photosynthetic process directly influencing plant growth, development and yield. However, plant pigment dynamics arise from complex metabolic interactions that are not adequately captured by conventional statistical approaches or traditional laboratory analyses, which are time-consuming and impractical for large-scale field applications. In this context, remote sensing offers a non-destructive alternative for assessing foliar pigments in agricultural systems, although its effectiveness depends on robust modeling approaches capable of capturing field-scale variability. This study evaluated the performance of different machine learning models to estimate chlorophyll A, chlorophyll B, and total chlorophyll using visible-range reflectance acquired from a proximal optical sensor. Corn leaves were collected at 14, 32 and 68 days after planting (DAP) from an experimental field in Campo do Meio, Minas Gerais, Brazil. The experiment followed a split-plot design, with biofertilizer application (Azospirillum brasilense and control) as the main plot and nitrogen management strategy (conventional fertilization and variable-rate nitrogen) as the subplot factor. Reflectance data were obtained from five fully expanded leaves per plot using a NIX SpectraL® sensor across the 400–700 nm range at 10 nm intervals. Chlorophyll A, B and total chlorophyll were measured on the same leaves using a ClorofiLOG® sensor. Exploratory analysis indicated higher variability in the blue region (400–440 nm), while other spectral bands showed lower dispersion. Chlorophyll A exhibited the widest range of values, whereas chlorophyll B showed lower variability. Feature selection using the Minimum Redundancy Maximum Relevance (mRMR) method identified bands at 480, 410, and 670 nm for chlorophyll A, 670, 400, and 550 nm for chlorophyll B and 670, 550, and 400 nm for total chlorophyll. These selected features were used to train twenty machine learning models. The dataset was split into 70% for training and 30% for testing, with five-fold (K-

The authors are solely responsible for the content of this paper, which is not a refereed publication. Citation of this work should state that it is from the Proceedings of the 17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture. EXAMPLE: Last Name, A. B. & Coauthor, C. D. (2026). Title of paper. In Proceedings of the 17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture and Brazilian Congress on Precision and Digital Agriculture.

fold) cross-validation applied. Tree-based boosting models, including Extra Trees, XGBoost, and CatBoost, achieved high training performance (R^2 > 0.90) but showed a decline in test performance (R^2 < 0.80), indicating overfitting. In contrast, K-nearest neighbors (KNN), Gaussian Process, and support vector machine (SVM) models exhibited more consistent performance between training and testing datasets, with R^2 values around 0.75. These models also produced lower prediction errors for chlorophyll A and B (maximum RMSE = 4.01) and high agreement indices (Willmott's d > 0.90). This stability is consistent with the smooth and non-linear relationships between reflectance and chlorophyll content observed in the spectral scatter patterns. Overall, the visible spectral range proved effective for supporting machine learning-based estimation of chlorophyll A, B and total chlorophyll. While boosting models effectively captured complex local patterns, their high flexibility increased susceptibility to overfitting. In contrast, models based on similarity and distance metrics demonstrated greater robustness, capturing underlying physiological relationships rather than local spectral noise.

Keywords.
Nix, Corn, Proximal Sensor, Feature Selection.

Introduction

Chlorophylls a and b are the primary photosynthetic pigments, functioning as the light-harvesting and energy-transferring chromophores within the antenna pigment–protein complexes that enable photochemical charge separation in Photosystems I and II (P700 and P680, respectively) (Grimm, 2001; Mauzerall, 1976). Their biosynthesis depends on magnesium and nitrogen, so leaf chlorophyll content is closely linked to plant nitrogen status and, ultimately, to crop yield (Grimm, 2001). Because the chlorophyll a:b ratio and total chlorophyll respond to nitrogen supply and to abiotic stress before such effects are expressed as yield loss, their quantification is a valuable indicator for early, in-season crop diagnosis. The reference method, however, relies on solvent extraction followed by spectrophotometric or chromatographic determination, a destructive, time-consuming and labor-intensive procedure that is impractical for large-scale or real-time assessment in the field.

Reflectance measurements offer a rapid, non-destructive alternative to wet chemistry. Chlorophyll absorbs strongly in the blue (430–450 nm) and red (660–680 nm) regions, and reflectance near the red absorption minimum (675 nm) has long been used to estimate chlorophyll content (Datt, 1998). This relationship, however, saturates at medium-to-high chlorophyll contents, even when reformulated as a vegetation index (Gitelson et al., 2003). Reflectance in the green (500–550 nm) and in the red-edge region (700–750 nm) is less prone to saturation and relates to chlorophyll across a wide range of species and concentrations, although in a markedly non-linear way (Gitelson & Merzlyak, 1994; Gitelson et al., 1996). Capturing these features at high spectral detail requires hyperspectral sensors, which remain costly and demand complex calibration, whereas multispectral sensors lower the cost but sample only a few, relatively broad bands, constraining pigment retrieval (Singhal et al., 2019; Croft et al., 2015).

At the most accessible end of this cost gradient are handheld broadband color (RGB) sensors, which are inexpensive, provide instantaneous readings and require minimal operator training, making them attractive for routine in-field use. Their limitation is intrinsic by integrating reflectance over a few wide channels, they capture far less spectral information than narrow-band instruments, which would be expected to hinder pigment retrieval, particularly the separation of chlorophyll a from chlorophyll b, whose spectral contributions overlap and are dominated by chlorophyll a. Consequently, most proximal and UAV-based studies have targeted total chlorophyll rather than the individual a and b pools (Singhal et al., 2019; Croft et al., 2015; Wang et al., 2023; Chusnah et al., 2023).

The non-linear and collinear nature of the relationship between color features and leaf pigments limits the usefulness of simple linear models and motivates machine learning approaches able to exploit such structure. Support Vector Machines, K-Nearest Neighbors and Random Forest have been applied to estimate chlorophyll from UAV and satellite imagery across several crops (Singhal et al., 2019; Croft et al., 2015; Wang et al., 2023; Chusnah et al., 2023). These platforms, however, entail operational costs, depend on flight and illumination conditions, and seldom provide leaf-level estimates in real time. Combining a low-cost color sensor with non-linear, ensemble-based learners could therefore recover chlorophyll a, b and total content directly in the field, compensating through multivariate modelling for the limited spectral information of the device.

Accordingly, the objective of this study was to evaluate the performance of a range of machine-learning models in estimating chlorophyll a, chlorophyll b and total chlorophyll from visible reflectance acquired with a low-cost proximal color sensor. We hypothesized that (i) ensemble and non-linear models would achieve higher predictive accuracy on unseen data than linear models, owing to their capacity to capture non-linear interactions among the color features, and (ii) features in the green region (500–550 nm) would be the most informative, since green reflectance is less prone to saturation at medium-to-high chlorophyll contents than the red absorption region.

Material and Methods

Experimental design and location

The study was conducted in 2025 year in a commercial farm located in Campo do Meio, Minas Gerais, Brazil. The total area destined for the experiment was 4.5 hectares with plots of 24 meters wider by 35 meters long totally 54 plots. The dimensions were based on the machine applicator swath which was estimated to be 24 meters. The experimental design adopted was Split plot in blocks with the main factor the inoculation or not with *Azospirillum brasilense* called strip (3 repetition for each one) followed by the sub-plot treatments which represent the conventional farm production (CFP) and Variable Rate Nitrogen (VRN) for nitrogen strategy (9 repetitions for each one) and 3 blocks.

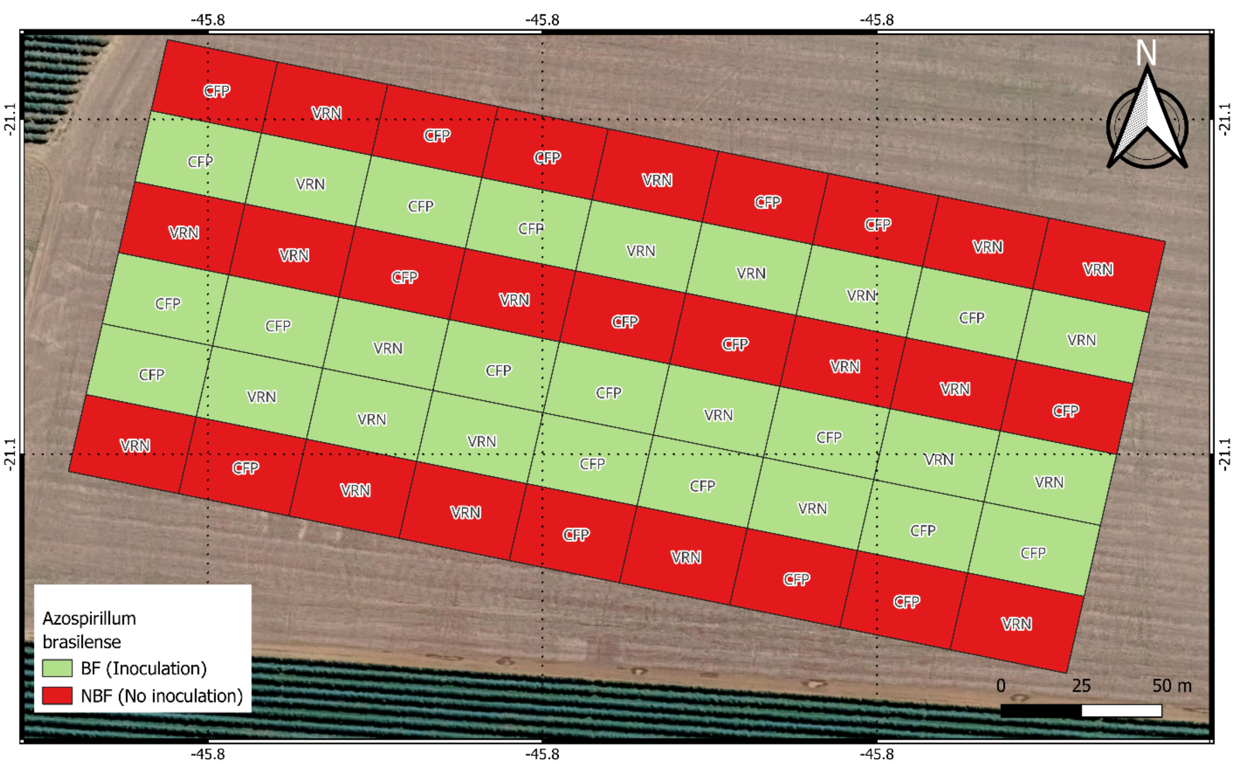

Fig 1. experimental design for 2025 year of data collection in Campo do Meio, Minas Gerais, BR. CFP – conventional farm production and VRN - variable rate nitrogen.

Field Management

Weed, insect and disease control followed the farmer's standard practices. The maize hybrid Pioneer P3440PWU was sown at a row spacing of 0.90 m and a target population of 74,000 plants ha^{-1}. The inoculant strips received *Azospirillum brasilense* (MBio Azo®, Mosaic, Tampa, FL, USA; strains AbV5 and AbV6), applied in-furrow at sowing at a dose of 0.2 L ha^{-1} and a carrier volume of 40 L ha^{-1}.

Two nitrogen (N) management strategies were compared a conventional fixed-rate practice (CFP) and variable-rate nitrogen (VRN). Both strategies received the same starter N at sowing. For the CFP, the full-season target was 144 kg N ha^{-1}, of which 27 kg ha^{-1} (≈19% of the season total) was applied at sowing and the remaining 117 kg ha^{-1} as a uniform side-dressing at the V5 growth stage. The starter N was supplied by 220 kg ha^{-1} of Mosaic NPK 12-16-13 (13% S, of which 10% as sulfate-S (S-SO_4) 0.12% B), and the side-dressing N was supplied as urea (46% N). For the VRN strategy, the at-sowing N was identical, but the side-dress rate was prescribed site-specifically following the algorithm of Holland and Schepers (2013) (Equation 1).

$$Napp = (N_{OPT} - N_{PF} - N_{OM}) \times \sqrt{\frac{(1-SI)}{\Delta SI \times (1 + 0.1 \times e^{m \times (SI_{threshold} - SI)})}} \quad (1)$$

Napp represents the nitrogen applied, Nopt the maximum nitrogen prescribed by producers, NPF means the pre-fertilizer nitrogen doses applied to sowing, NOM the nitrogen credit from organic matter, SI is the sufficient index calculated from equation 2, ΔSI is the sufficient index difference parameter, m is the back-off rate variable (0 < m < 100) and SIthreshold is the back-off cut-on point (Holland and Schepers, 2013).

$$SI = \frac{VI_{sensed\ crop}}{VI_{reference}} \quad (2)$$

where $VI_{sensed\ crop}$ is the NDRE of the VRN plots (Eq. 3) and $VI_{reference}$ is the Normalized Difference Red Edge (NDRE) of the reference strips. A separate reference was used for the non-inoculated and inoculated strips, for each strip, the NDRE histogram was built and its 95^{th} percentile (cumulative value) was taken as $VI_{reference}$. (Holland and Schepers, 2013).

$$NDRE = \frac{NIR - Red\ Edge}{NIR + Red\ Edge} \quad (3)$$

Where NIR and Red Edge are the near-infrared and red-edge reflectance bands, respectively (Sims and Gamon, 2002). Following Holland and Schepers (2013), the parameters were set to m = 20, SIthreshold = 0.7, ΔSI = 0.3, Nopt = 144 kg ha^{-1} and N_{PF} = 27 kg ha^{-1}. N_{OM} was set to zero, because soil organic matter was not measured.

Satellite image download

The vegetation index required for the variable-rate prescription (Equations 1–3) was derived from PlanetScope imagery acquired in 2025. Images were collected by the SuperDove sensor (instrument PSB.SD), which provides eight spectral bands at a 3 m spatial resolution and a daily revisit frequency. Besides that, only band 7 (Red Edge) and band 8 (NIR) were used to compute the NDRE.

The PlanetScope Surface Reflectance (SR) product was used, which is delivered already atmospherically corrected. In this product, the correction is performed with a radiative-transfer framework based on 6SV2.1 (Second Simulation of a Satellite Signal in the Solar Spectrum) precomputed lookup tables (LUTs) derived from 6SV2.1 simulations are combined with near-real-time atmospheric inputs (aerosol optical depth, water vapor and ozone) obtained from MODIS products, converting top-of-atmosphere (TOA) into bottom-of-atmosphere (BOA) reflectance and thereby minimizing atmospheric scattering and absorption effects.

A single cloud-free scene was acquired on 22 April 2025, one day before the side-dressing application was used. The image was clipped to the field boundary directly in the Planet web platform, so that only the area of interest was downloaded rather than the full scene. The NDRE was then calculated from the SR (BOA) Red-Edge and NIR bands using the *image_index* function of the *pliman* package (Olivoto, 2022).

Spectral Bands and Foliar Pigments

Leaf reflectance was measured with a NIX Spectro L (Nix Sensor Ltd., Hamilton, Ontario, Canada). This portable spectrometer records reflectance across the visible region from 400 to 700 nm at 10 nm intervals (Figure 2), yielding 31 reflectance bands interpolated from its nine physical channels (eight spectral channels and one clear channel). Measurements were taken at 14, 32 and 68 days after planting (DAP). On each date and within each plot, six maize plants located within a 5 m radius of the plot center were sampled, and one fully expanded leaf was detached from each plant. Because the sensor requires an internet connection to register each reading, the detached leaves were measured immediately at a nearby location with network coverage, giving one reflectance reading per plant. A small wooden block was placed behind each leaf during measurement to prevent background reflectance from influencing the readings.

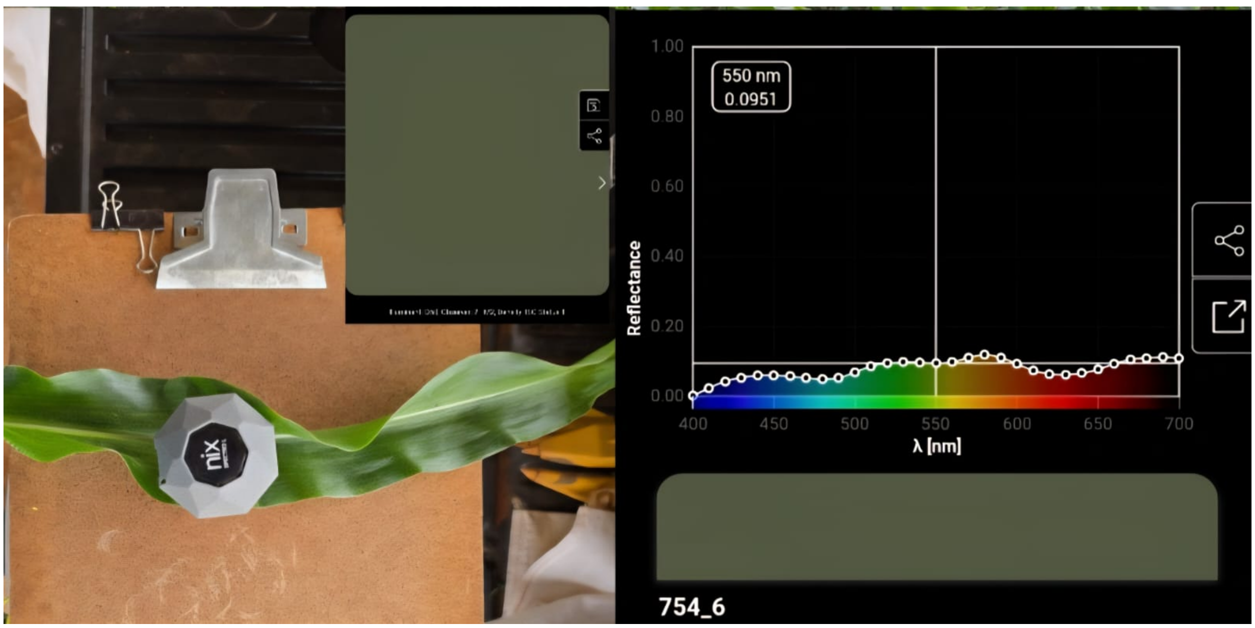

Fig 2. Measures using the NIX Spectra L and the reported reflectance values across 400 – 700 nanometers.

After read the leaves using the NIX Spectra L the chlorophyll a, b and total were estimate using the Clorofilog 1030, Falker®, Porto Alegre, Rio Grande do Sul, Brazil. The sensor measure the transmittance of three wavelengths, two close to the peak of absorbance of chlorophyll (Red) and one close to near infrared. Thus, the sensor generate a non-dimensional value called Falker Chlorophyll Index (ICF) which represents the chlorophyll a and b, and the total chlorophyll obtained by sum the chlorophyll a and b.

Feature Selection and Machine Learning Models

Spectral band selection was performed with the Minimum Redundancy Maximum Relevance implemented in the mRMRe R package (De Jay et al., 2013), using the package default settings for continuous variables. mRMR is a filter-based feature selection method that ranks predictors by balancing two competing criteria. Relevance quantifies the association between each reflectance band and the target variable, whereas redundancy quantifies the association among the candidate bands themselves. The algorithm proceeds incrementally, at each step retaining the band that maximizes relevance to the target while minimizing redundancy with the bands already selected, so that the resulting order favors informative and mutually complementary bands. For continuous variables, the package estimates relevance and redundancy from a correlation-based approximation of mutual information, so that the reported scores retain the sign of the underlying association. The algorithm was applied independently to chlorophyll a, chlorophyll b and total chlorophyll, and only to the training partition of the 70/30 split, to prevent information leakage into the test set. The full set of 31 bands was ranked, and the first three bands selected for each target were retained as model inputs. This parsimonious subset was chosen to limit redundancy among highly correlated adjacent bands and to reduce the risk of overfitting, consistent with the low-cost, field-oriented purpose of the sensor.

To provide a comprehensive and unbiased comparison across learning paradigms, twenty regression algorithms were evaluated under identical conditions, ranging from linear and regularized baselines (multiple linear regression, Ridge, Lasso, Elastic Net, Bayesian Ridge, Huber and linear SVR) to non-linear learners without ensembling (K-Nearest Nneighbors, a Support Vector Machine with a radial basis function kernel, a Gaussian Process, a Multilayer Perceptron and a single Decision Tree), and finally to ensemble methods based on bagging (Random Forest and Extra Trees) and boosting (AdaBoost, Gradient Boosting, Histogram-based Gradient Boosting, XGBoost, LightGBM and CatBoost). Spanning this gradient of model complexity was a deliberate choice, since it allowed the linear baselines to be contrasted directly with progressively more flexible learners and therefore to test whether the non-linear and collinear nature of the reflectance signal, ensemble and kernel-based approaches over simple linear ones. All algorithms were trained on the three bands retained by mRMR for each target, so that differences in performance could be attributed to the learning method rather than to the input features, and all shared the same 70/30 partition. Hyperparameters were tuned on the training set with grid search and five-fold cross-validation (GridSearchCV), while the test set was held out and used exclusively for the final assessment. Most models were implemented in Python with the scikit-learn library (Pedregosa et al., 2011), except XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018), which relied on their respective packages. In this context, the models were trained and tested for chlorophyll a, b and total chlorophyll.

Evaluation Metrics

Model performance was assessed with four metrics. The coefficient of determination (R^2, equation 4) measures how well the model explains the variation in the target variable and ranges from 0 to 1. The mean absolute error (MAE, equation 5) is the average absolute difference between predicted and observed values, treating all errors linearly (Terven et al., 2025), whereas the root mean square error (RMSE, equation 6) measures the average prediction deviation and, by squaring the residuals, is more sensitive to large errors and outliers. The Willmott index of agreement (d, equation 7) ranges from 0 to 1 and compares the magnitude of the residuals to the potential variability of observed and predicted values around the observed mean, providing a measure of agreement that is less affected by outliers than the correlation coefficient (Willmott, 1981).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4}$$

R^2 means coefficient of determination, y_i represent the observed values, $\hat{y}_i$ the predicted values and $\bar{y}_i$ the mean for observed values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

MAE represent the mean average error, y_i represent the observed values and $\hat{y}_i$ the predicted values and *n* represent the number of samples.

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

RMSE represent the root mean square error, yi represent the observed values and $\bar{y}_1$ the predicted values and n represent the number of samples.

$$d = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(|\hat{y}_i - \bar{y}_1| + |y_i - \bar{y}_1|)^2} \quad (7)$$

The models were running using the Python 3.10 version in VS code interface. The graphs were plotted using the Rstudio version 4.4.3 and the package GGplot2 (Wickham et al. 2016).

Results and Discussion

The exploratory analysis of the reflectance bands is presented as boxplots in Figure 3. In terms of magnitude, the lowest reflectance was observed in the blue region, from 400 to 440 nm, whereas the highest values occurred near 680, 690 and 700 nm. Beyond these overall levels, the bands differed markedly in their dispersion across samples. Reflectance varied little between 450 and 490 nm, indicating a relatively homogeneous response among samples, but the spread increased substantially from 550 to 590 nm and remained high up to 640 nm. This greater between-sample variability in the green to orange region is relevant because a band can only be informative for chlorophyll estimation if its reflectance varies across the sampled range, which is consistent with the bands subsequent prioritized by the mRMR algorithm.

The reference pigment measurements also differed in their distribution (Figure 3). Chlorophyll *a* ranged from approximately 30 to 58 ICF with a median near 48, and total chlorophyll followed a similar pattern, ranging from about 45 to 72 with a median near 58. This close correspondence indicates that total chlorophyll was governed mainly by chlorophyll a, consistent with the dominance of this pigment in healthy leaves. Chlorophyll b, in contrast, showed the narrowest range, from roughly 8 to 30, with a markedly lower median near 13 and a more asymmetric distribution. The resulting chlorophyll a to b ratio of about 3.7 to 1 is consistent with values reported for healthy green leaves (Datt, 1998).

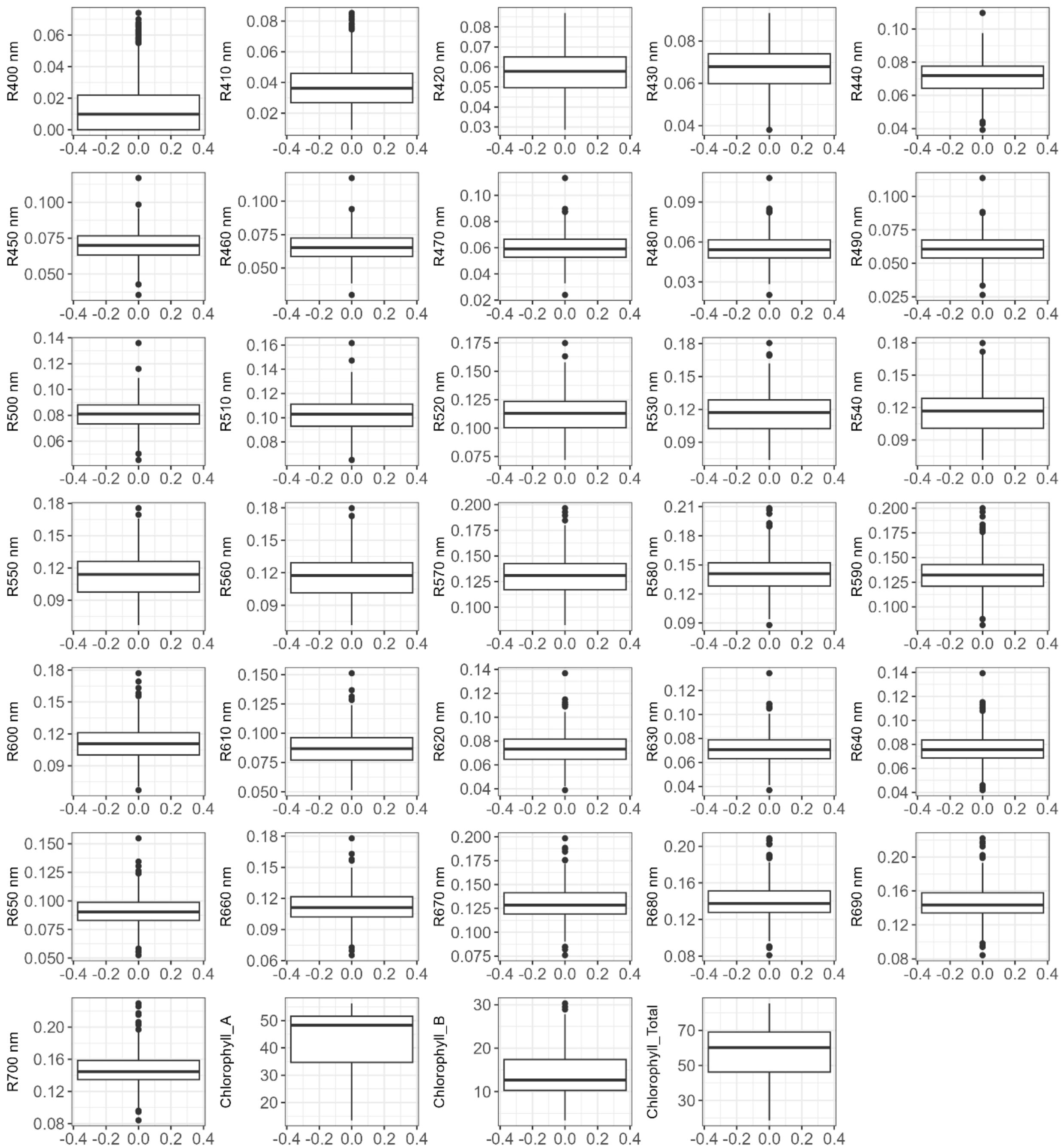

Fig 3. boxplot analysis for all wavelengths and chlorophyll a, b and total.

Following the band selection, the mRMR algorithm returned similar wavelengths for chlorophyll a, b and total (Figure 4A, B and C). For the three targets, the highest-magnitude scores were negative and concentrated in the green to orange region, with 610, 510 and 600 nm consistently ranked first, and 520 nm appearing among the top bands for chlorophyll b. The negative sign indicates an inverse relationship, since reflectance in this region decreases as chlorophyll content increases, which is consistent with the absorbing behavior of the pigment. This intermediate spectral region is informative because the absorptivity of chlorophyll in the green to orange wavelengths is high enough to respond to pigment content yet weaker than at the main absorption bands, so that reflectance varies appreciably with chlorophyll without saturating (Carter and Knapp, 2001). Earlier work supports this pattern, as the strongest changes in leaf reflectance with chlorophyll occur near the green peak, whereas the absolute differences near the red and blue absorption maxima are comparatively small (Yoder and Pettigrew-Crosby, 1995). In contrast, the bands near 670 and 400 nm received scores close to zero for the three pigments, indicating weak relevance. This is explained by the saturation of chlorophyll absorption in the red region around 670 to 680 nm, where large amounts of pigment must be lost before reflectance changes appreciably, and by the strong absorption of accessory pigments such as carotenoids in the blue region, which keeps reflectance there relatively insensitive to chlorophyll (Carter and Knapp, 2001). These results highlight the hypothesis II that features in the green region (500–550 nm) would be the most informative, since green reflectance is less prone to saturation at medium-to-high chlorophyll contents than the red absorption region. Red region is often associated with high absorbance for chlorophyll (Sims and Gamon, 2002; Gitelson, 2004), however medium to high concentration is enough to saturate the values and the bands did not capture the differences.

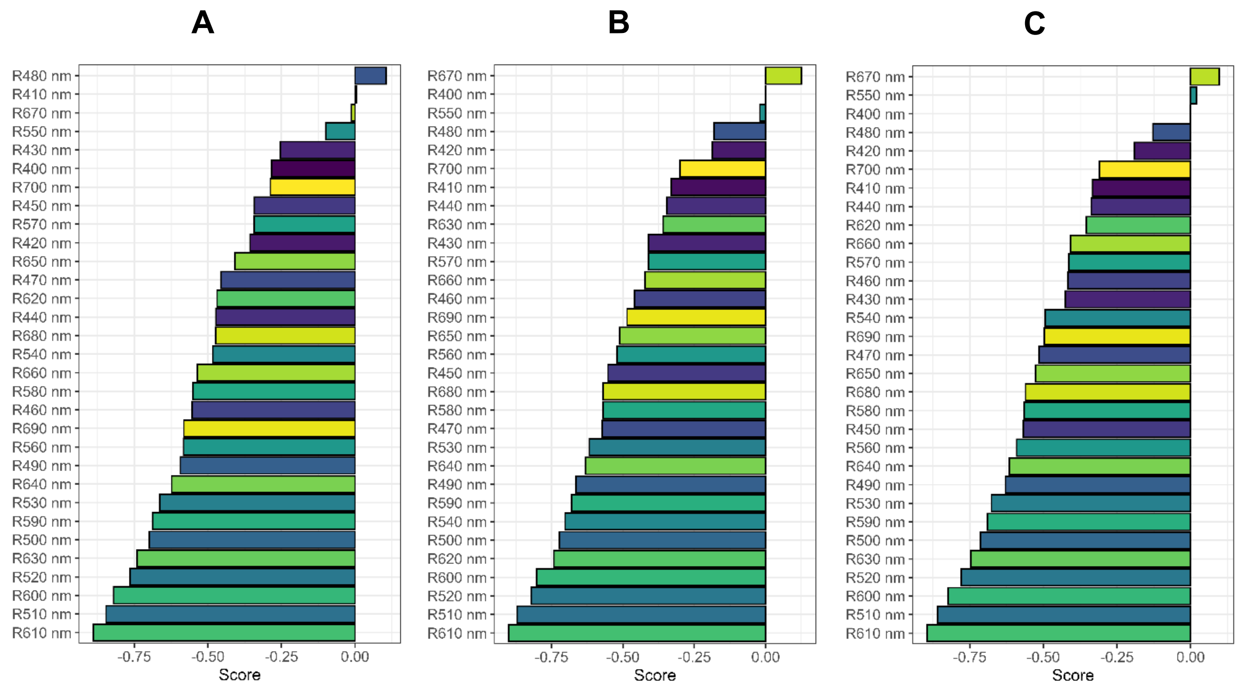

Fig 4. results for Minimum Redundancy Maximum Relevance (mRMR) algorithm for chlorophyll a (A), chlorohyll b (B) and total chlorophyll (C).

Based on the feature selection analysis, spectral bands at 510, 600, and 610 nm were identified as the most relevant for predicting chlorophyll a and total chlorophyll. For chlorophyll b, the selected bands were 510, 520, and 610 nm. Prior to training the machine learning models, we analyzed the reflectance distribution of these specific bands to understand their relationship with foliar pigment concentrations. The scatter plots reveal a consistent pattern across all pigments where the 610 nm band exhibits the lowest reflectance values. Conversely, the 600 nm and 520 nm bands demonstrate the highest reflectance. This behavior aligns with typical leaf optical properties where higher absorption occurs near the red edge region and higher reflectance occurs in the green region. A distinct bimodal distribution is particularly evident for chlorophyll a. We can observe two primary data clusters separated approximately between 30 to 40 ICF and 45 to 55 ICF. These distinct groupings correspond directly to the phenological stages of the corn crop. The cluster with lower chlorophyll levels represents the initial developmental stages at 14 and 32 days after planting. As the plant matures and expands its photosynthetic capacity, pigment accumulation increases significantly and forms the second cluster associated with advanced stages such as 68 days after planting.

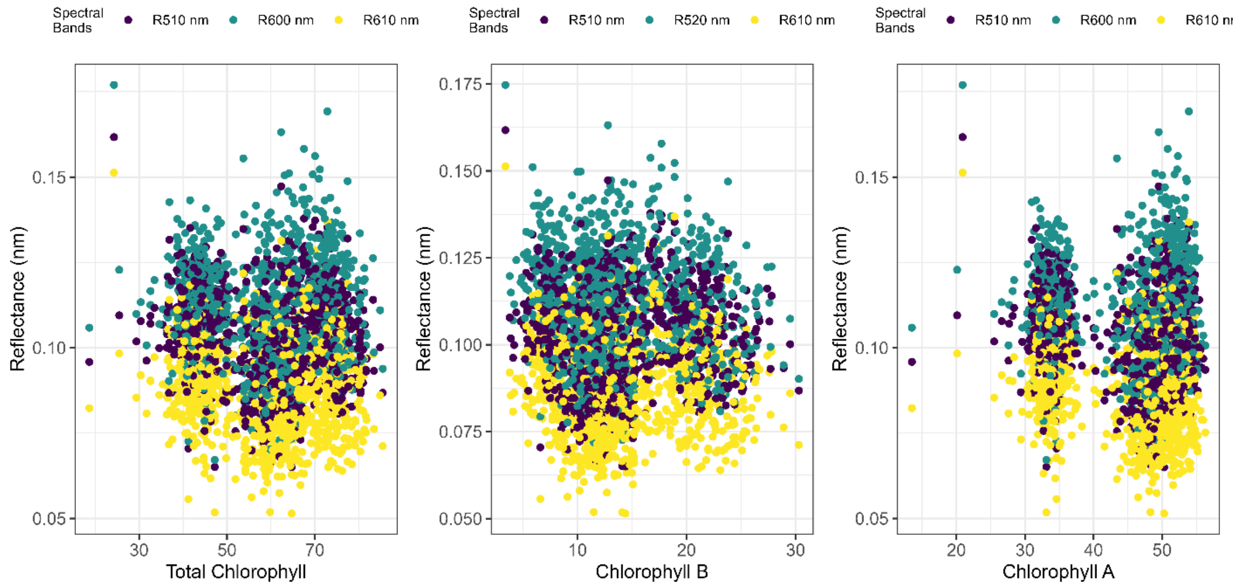

Fig 5. scatter plot analysis for 510, 520, 600 and 610 nanometers bands for Chlorophyll a, Chlorophyll b and Total Chlorophyll.

Chlorophyll a represents the most abundant plant pigment found in all light-harvesting complexes and reaction centers, functioning as the primary electron donor for photosystems I and II (Pareek et al., 2018). Chlorophyll b acts as an accessory pigment that assists chlorophyll a by absorbing light primarily in the blue region (with peaks at 453 nm and 625 nm). These pigments differ slightly in their molecular structures, with chlorophyll presenting the formula $C_{55}H_{72}MgN_4O_5$ and chlorophyll b possessing $C_{55}H_{70}MgN_4O_6$ (Pareek et al., 2018). During early developmental stages, such as V2 or V3, the total chlorophyll concentration remains relatively low as the vegetative structures are still expanding. However, as the plant progresses toward advanced vegetative stages and the canopy closes, the mutual shading of the lower leaves triggers a physiological adaptation. To maximize light capture under higher competition and diffuse light conditions within the lower canopy, plants alter their pigment production, specifically decreasing the chlorophyll a to b ratio, which results in a peak of total pigment accumulation (Taiz et al., 2015; Ciganda et al., 2009).

The predictive performance differed markedly across learning paradigms (Figure 6). On the test set, the linear and regularized models converged to the lowest accuracy, with R^2 close to 0.59 to 0.60 for chlorophyll a, and the ensemble and non-linear models reached the highest values, between 0.73 and 0.76. Extra Trees and the Gaussian Process achieved the best test R^2 for chlorophyll a (0.76), closely followed by the boosting methods and by the kernel-based SVM and KNN. This pattern supports the hypothesis that the non-linear and collinear nature of the reflectance signal favors flexible learners over simple linear models, although the advantage was shared by ensemble and non-ensemble non-linear methods rather than being exclusive to boosting. The four metrics were mutually consistent, as the Willmott index ranked the models in the same order as R^2 (around 0.92 to 0.93 for the best models against 0.88 to 0.89 for the linear ones), and RMSE and MAE followed the same trend.

The tree-based ensembles showed a pronounced gap between training and test performance, with R^2 for chlorophyll a falling from about 0.90 to 0.92 in training to 0.75 to 0.76 in testing and RMSE nearly doubling from training to test. The linear models, in contrast, showed almost no gap. This behavior is expected for high-capacity learners trained on three predictors and a moderate sample size, indicating some overfitting that nonetheless did not prevent the ensembles from generalizing better than the linear baselines. The single Decision Tree was the clearest example of instability, collapsing to an R^2 of 0.51 for total chlorophyll, which highlights the gain obtained by aggregating trees through bagging or boosting.

Among the three targets, chlorophyll b was the most difficult to predict, with the best test R^2 reaching only 0.66 to 0.67 and a narrower spread among models, whereas total chlorophyll behaved similarly to chlorophyll a. This result is consistent with the exploratory analysis, in which chlorophyll b showed the smallest range of variation, and with the band selection, in which chlorophyll b relied on an additional and weaker feature near 520 nm. The dominance of chlorophyll a in the total pigment pool explains the close correspondence between the chlorophyll a and total chlorophyll models.

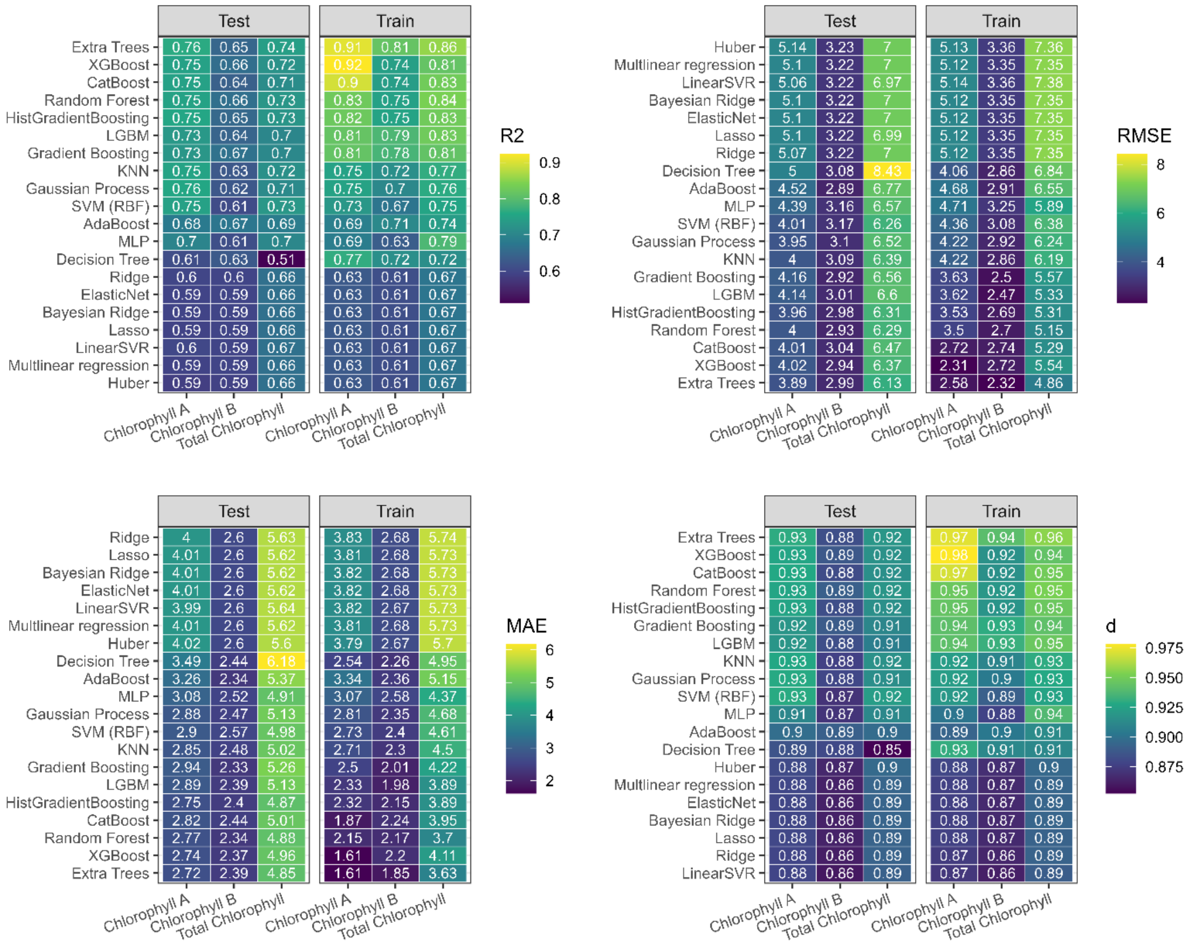

Fig 6. results for coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE) and Willmott Coefficient Index (d) for chlorophyll a, b and total chlorophyll.

The performance obtained in the present study is comparable to that reported in the recent literature on chlorophyll estimation. An et al. (2020) estimated SPAD values in rice from hyperspectral data and reported validation R^2 between 0.76 and 0.80, while Ta et al. (2021) estimated leaf chlorophyll content in apple trees with R^2 between 0.63 and 0.74. Gao et al. (2024) estimated chlorophyll in soybean leaves from hyperspectral bands and reported R^2 ranging from 0.458 to 0.846. Most of these studies rely on hyperspectral sensors, which are costly and require specialised processing. In the present study, a low-cost proximal sensor restricted to the visible region reached a comparable level of accuracy, which highlights the potential of combining small, field-deployable sensors with machine-learning models for rapid in-field chlorophyll diagnosis.

A clear direction for future work concerns the spectral range of the sensor. Because the device measures only up to 700 nm, it does not capture the red-edge region near 700 to 750 nm, which the literature identifies as the spectral interval most strongly correlated with chlorophyll content (Carter and Knapp, 2001; Yoder and Pettigrew-Crosby, 1995). The fact that the present models reached R^2 close to 0.76 without access to this region suggests that extending the sensing range into the red edge, together with validation across additional sites, crop stages and growing seasons, could further improve estimation accuracy, particularly for chlorophyll b, which remained the most difficult target to predict.

Conclusions

This study evaluated whether chlorophyll a, chlorophyll b and total chlorophyll could be estimated from visible reflectance acquired with a low-cost proximal sensor combined with machine-learning models. Ensemble and non-linear models clearly outperformed the linear baselines, confirming that the non-linear relationship between reflectance and leaf pigments is better captured by flexible learners. The most informative wavelengths were concentrated in the green to orange region, where chlorophyll absorptivity is high enough to respond to pigment content without saturating, while the red and blue absorption maxima contributed little. Total chlorophyll and chlorophyll a were estimated more accurately than chlorophyll b, which remained the most challenging target. Overall, the results show that a low-cost visible sensor coupled with machine-learning models is a promising tool for rapid, non-destructive chlorophyll diagnosis in the field, and that extending the spectral range into the red-edge represents a clear path to further improve estimation accuracy.

Acknowledgments

We would like to highlight the acknowledgements for Matheus Araujo from Fortaleza farm and his family for his effort in supporting the development of the project. The Digital Agriculture Extension and Research Group (GEPAD) members for the estimated support for data collection. We extend the acknowledgments for Brazilians agencies: Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), National Council for Scientific and Technological Development (CNPq) and The Research Support Foundation of the State of Minas Gerais (FAPEMIG) for the financial support to researchers, professor and students. Additionally, we would like to thank you for Federal University of Lavras (UFLA) in especial for Department of Agriculture (DAG) for the support with materials and structure to develop the project.

References

Gao, C. *et al.* (2024). Combined use of spectral resampling and machine learning algorithms to estimate soybean leaf chlorophyll. Computers and Electronics in Agriculture, doi: https://doi.org/10.1016/j.compag.2024.108675

An, G., Xing, M., He, B., Liao, C., Huang, X., Shang, J., & Kang, H. (2020). Using Machine Learning for Estimating Rice Chlorophyll Content from In Situ Hyperspectral Data. *Remote Sensing*, doi: https://doi.org/10.3390/rs12183104

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

Pareek, S., Sagar, N. A., Sharma, S., Kumar, V., Agarwal, T., González-Aguilar, G. A., & Yahia, E. M. (2018). Chlorophylls: Chemistry and Biological Functions. *Chemistry and Human Health*, doi: https://doi.org/10.1002/9781119158042.ch14

Carter, G. A, & Knapp, A. K. (2001). Leaf optical properties in higher plants: linking spectral characteristics to stress and chlorophyll concentration. *Physiology and Chemistry*, doi: https://doi.org/10.2307/2657068

Yoder, B. J., & Pettigrew-Crosby, R. E. (1995). Predicting nitrogen and chlorophyll content and concentrations from reflectance spectra (400–2500 nm) at leaf and canopy scales. *Remote Sensing of Environment*, doi: https://doi.org/10.1016/0034-4257(95)00135-N

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

Grimm, B. (2001). Chlorophyll: Structure and Function. Encyclopedia of Life Sciences, link: https://biomar.fciencias.unam.mx/Sobretiros/2019/Informacion%20algas/Algas/alfabetico/g/GRIMM%20clorofila.pdf accessed on 13 June of 2026

Mauzerall, D. (1976). Chlorophyll and photosynthesis. *Philos Trans R Soc Lond B Biol Sci*, doi: https://doi.org/10.1098/rstb.1976.0014

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. *Journal of Plant Physiology*, doi: https://doi.org/10.1078/0176-1617-00887

Gitelson, A., Merzlyak, M. N. (1994). Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology B: Biology*, doi: https://doi.org/10.1016/1011-1344(93)06963-4

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

Singhal, G., Bansod, B., Mathew, L., Goswami, J., Choudhury, B. U., & Raju, P. L. N. (2019). Chlorophyll estimation using multi-spectral unmanned aerial system based on machine learning techniques. *Remote Sensing Applications: Society and Environment*, doi: https://doi.org/10.1016/j.rsase.2019.100235

Croft, H., Chen, J. M., Zhang, Y., Simic, A., Noland, T. L., Nesbitt, N., & Arabian, J. (2015). Evaluating leaf chlorophyll content prediction from multispectral remote sensing data within a physically-based modelling framework. *ISPRS Journal of Photogrammetry and Remote Sensing*, doi: https://doi.org/10.1016/j.isprsjprs.2015.01.008

Wang, Y., Tan, S., Jia, X., Qi, L., Liu, S., Lu, H., Wang, C., Liu, W., Zhao, X., He, L., Chen, J., Yang, C., Wang, X., Chen, J., Qin, Y., Yu, J., & Ma, X. (2023). Estimating Relative Chlorophyll Content in Rice Leaves Using Unmanned Aerial Vehicle Multi-Spectral Images and Spectral–Textural Analysis. *Agronomy*, doi: https://doi.org/10.3390/agronomy13061541

Chusnah, W. N., Chu, H. J., & Tatas, L. M. J. (2023). Machine-learning-estimation of high-spatiotemporal-resolution chlorophyll-a concentration using multi-satellite imagery. Sustainable Environment Research, doi: https://doi.org/10.1186/s42834-023-00170-1

Holland, K. H., & Schepers, J. S. (2013). Use of a virtual-reference concept to interpret active crop canopy sensor data. *Precision Agriculture*, doi: https://doi.org/10.1007/s11119-012-9301-6

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

Olivoto, T. (2022). Lights, camera, pliman! An R package for plant image analysis. Methods in Ecology and Evolution, doi:10.1111/2041-210X.13803.

De Jay, N., Cavanagh, S. P., Olsen, C., El-hachem, N., Bontempi, G., & Haibe-Kains, B. (2013). mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics*, doi: https://doi.org/10.1093/bioinformatics/btt383

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). *Association for Computing Machinery*, doi: https://doi.org/10.1145/2939672.2939785

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 (pp. 3146–3154).

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesari-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 6638–6648).

Terven, J. R., Cordova-Esparza, D. M., Ramizrez-Pedraza, A., Chavez-Urbiola, E. A., & Romero-Gonzalez, J. A. (2025). A comprehensive survey of loss functions and metrics in deep learning. Artificial Intelligence Review, doi: https://doi.org/10.1007/s10462-025-11198-7

Willmott, C. J. (1981). On the Validation of Models. Physical Geography,, doi: http://dx.doi.org/10.1080/02723646.1981.10642213

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). *Springer-Verlag*, doi: https://doi.org/10.1007/978-3-319-24277-4

Gitelson, A. A. (2004). Wide Dynamic Range Vegetation Index for Remote Quantification of Biophysical Characteristics of Vegetation. Journal of Plant Physiology, doi: https://doi.org/10.1078/0176-1617-01176

Taiz, L., Zeiger, E., Møller, I. M., & Murphy, A. (2015). *Plant physiology and development* (6th ed.).

Ciganda, V., Gitelson, A., & Schepers, J. (2008). Vertical Profile and Temporal Variation of Chlorophyll in Maize Canopy: Quantitative “Crop Vigor” Indicator by Means of Reflectance-Based Techniques. *Agronomy Journal*, doi: https://doi.org/10.2134/agronj2007.0322