Evaluation of Machine Learning Methods for Estimation of Leaf Chlorophyll Content (LCC) Across 15 Soybean Cultivars During Early Reproductive Stage

Carli Kriek[1], Philemon Tsele[1], George Chirima[2], Adolph Nyamugama[2]

[1]Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria 0002, South Africa;
[2]Agriculture Research Council Natural Resource & Engineering (NRE), Pretoria, 0001, South Africa

Keywords: Soybean LCC estimation, Remote Sensing, Multispectral UAV, Machine learning regression, Vegetation indices

Abstract

South Africa is the leading soybean producer in Africa, contributing approximately 35% of the continent's total production. Soybean is important for national food security and agricultural sustainability, serving as a key nitrogen-fixing crop that supports soil fertility and economic growth. Monitoring biochemical parameters such as leaf chlorophyll content (LCC) is essential for assessing soybean health; however, cultivar-level variability can complicate the use of remote sensing-based approaches. This study evaluates the performance of four machine-learning algorithms, eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Partial Least Squares Regression (PLSR), and Artificial Neural Network (ANN), using unmanned aerial vehicle (UAV)-based data across 15 soybean cultivars during the early reproductive phase. Results show that model performance is strongly cultivar dependent. Tree-based models achieved the highest accuracy, with XGBoost and RF reaching Root Mean Square Error (RMSE) values as low as 2.9 μmol m^{-2} for PHIP62T16R and R^2 values up to 0.96 for RA655R. In contrast, ANN and PLSR performed substantially worse for cultivars with more complex spectral responses, such as PAN1555R. Residual analyses from generalised models revealed systematic over- and under-prediction in several cultivars, indicating that pooled models could not fully account for cultivar-specific spectral differences. Variable importance analyses identified red-edge, near-infrared (NIR), and greenness-enhancing indices as the most influential predictors of LCC. Overall, the study demonstrates that incorporating cultivar information and using stratified model calibration significantly improves the reliability of UAV-based chlorophyll monitoring in heterogeneous soybean canopies.

1. Introduction

South Africa accounts for 35% of the total soybean production in Africa (Sedibe et al., 2022). As the continent's leading soybean producer, production in South Africa is of great socioeconomic and ecological importance (Clua et al., 2018, Engelbrecht et al., 2020). Since the early 2000s, the South-African demand for soybean production has increased significantly (Engelbrecht et al., 2020). Today, soybean plays a crucial role in South Africa's agricultural landscape for several compelling reasons. As a protein-rich crop, soybean could impede food insecurity in Sub-Saharan Africa, accounting for the rapid population growth, which has almost doubled to 60m (Engelbrecht et al., 2020). Additionally, the crop also contributes significantly as livestock feed (Engelbrecht et al., 2020) and serves as a biological nitrogen source for sustainable farming practices (Clua et al., 2018). The rapid growth in South Africa's soybean production is increasingly placing demands on arable land (Engelbrecht et al., 2020). Furthermore, soybean is faced with a heightened vulnerability to several pests and diseases (Engelbrecht et al., 2020)– threatening the plant's potential and causing significant yield loss (Roth et al., 2019). Effective management, planning, and predictive strategies are essential to optimise soybean production and achieve the highest possible yields. Precision agriculture (PA) incorporates a range of advanced technologies to monitor and manage spatial and temporal variability within fields. By optimising inputs and improving production efficiency, PA helps address food security challenges, while enhancing the resilience and profitability of agricultural ecosystems (Kganyago et al., 2021).

One of the most established approaches to assess crop spatial variability and monitor agricultural systems' productivity is by estimating key crop biophysical parameters such as leaf chlorophyll content (LCC) (Kganyago et al., 2021, Brewer et al., 2022). Because chlorophyll is strongly associated with leaf nitrogen content, it provides a robust indicator of photosynthetic capacity (Kganyago et al., 2021) and a critical proxy for the biophysical and biochemical processes governing photosynthesis in crops (Brewer et al., 2022). Moreover, chlorophyll plays a central role in the exchange of energy and materials between plants and their environment (Shi et al., 2023), and therefore acts as a key indicator of both vegetation health and growth conditions (An et al., 2020). As a result, LCC is an important variable in optimising the use of fertiliser while maximising soybean yield, and is therefore widely recognised for assessing crop productivity (Brewer et al., 2022). Traditional laboratory chlorophyll measurements are accurate; however, these measurements are often destructive, costly, and restricted in spatial and temporal context (Pasqualotto et al., 2019, Kganyago et al., 2021).

Since 1983, technology has played a crucial role in developing the agricultural industry (Jha et al., 2019). Remote sensing (RS) refers to the process of acquiring information about an object or phenomenon, from a distance (Weiss et al., 2020). RS systems are a cost-effective and non-destructive tool, that delivers timely, systematic and essential agricultural data (Weiss et al., 2020), at multiple spatial and temporal scales (Kganyago et al., 2021). In particular, unmanned aerial vehicles (UAVs) could potentially reduce measurement errors that often arise from environmental conditions (Hu et al., 2023). Not only do UAVs facilitate the collection of comprehensive spatio-temporal data, they continuously prove to be a valuable area of exploration in crop health monitoring (Mouafik et al., 2024).

The integration of RS and machine learning holds significant potential for advancing PA, by promoting scalable and data-driven decisions. Machine learning (ML) is widely recognised in the agricultural sector, for its ability to model complex relationships between input and output data, with remarkable accuracy (Mouafik et al., 2024).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-731-2026

Soybean flowers when day length becomes shorter than its critical photoperiod– the length of daylight a plant is exposed to within a 24-hour cycle (Liu et al., 2017, Sun et al., 2025). Since photoperiod is controlled by latitude and season, soybean cultivars are specifically bred to match the day-length patterns for the regions and latitudes where they are grown. Soybean cultivars are therefore classified into early, medium, medium-late, and late growth classes. This classification is based on their total growth duration and sensitivity to photoperiod which together determine how long they take to progress from planting to physiological maturity (Liu et al., 2017, Fehr, 1977). Assessing soybean during the early reproductive phase (R2–R3) is essential, as this is a peak physiological window when chlorophyll concentration and nitrogen demand increase sharply to support flowering, pod initiation, and early seed development (Board and Kahlon, 2011, Fehr, 1977). During this period, spectral reflectance can vary substantially among cultivars due to differences in canopy structure, pigment composition, and developmental timing.

Understanding the extent to which cultivar variability affects model accuracy is essential for determining whether a universal model can be applied across multiple cultivars, or if cultivar-specific models yield more accurate estimates. This study therefore aims to evaluate and compare the performance of various machine learning algorithms for estimating and mapping LCC from UAV multispectral imagery across 15 soybean cultivars during the early reproductive phase.

2. Study Area

This study was conducted on a commercial dryland crop production farm in the Highveld, situated 20 km South of Heidelberg, in Gauteng, South Africa (Figure 1). The spatial extent of the study area ranges from 28.378° to 28.383° E and –26.688° to –26.684° S (WGS84).

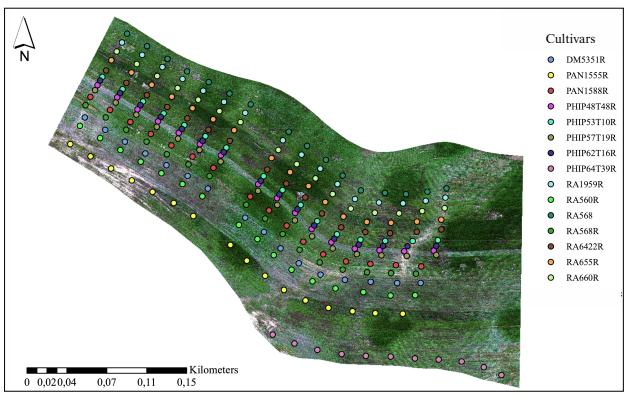

Figure 1. True-colour UAV composite image of the soybean study area captured during the early reproductive stage.

The study area belongs to the Highveld grassland biome, a notable, species rich, relatively flat and high-altitude, ecosystem with elevations ranging between 1400-1800 m above sea level (Swierk, 2024). This region naturally lends itself to agriculture and is well recognised for its extensive crop production. with a temperate climate, moderate rainfall, and fertile soils. The study area has a temperate climate with warm, wet summers and cool, dry winters. Summer temperatures range from 15–34°C, while winter temperatures vary between –7°C and 28°C. Average annual rainfall is 671 mm, with 725 mm recorded during the 2022/2023 soybean season. Soil depths in the area range from 30 to 50 cm, and soil types include, Westleigh, Cartref, Tukulu and Montagu. Typical to that of the Highveld eco-region, shallow soil profiles and seasonal water stress often influence crop health in the area. This makes the study area particularly well-suited to understanding spatial variability in crop health and assessing the effectiveness of crop health monitoring across different cultivars.

3. Data Sets Used

3.1 LCC field data sampling

LCC field measurements were collected on 24 and 25 January 2023. Samples were collected during the early reproductive phase of soybean (R2–R3), a stage crucial for determining final yield. At R2, the crop reaches peak flowering, and by R3 the pods start forming on the upper nodes (Fehr, 1977). During this phase, soybean experience a rapid increase in nitrogen demand to support pod initiation and early seed development. Given the strong relationship between chlorophyll content, nitrogen status, and canopy biophysics, this period represents a sensitive and informative window for remote-sensing-based chlorophyll estimation (Kganyago et al., 2021).

A systematic stratified sampling design was implemented where each cultivar served as a distinct stratum within the sampling framework, capturing both within- and between-cultivar variability in LCC. Within each stratum, 1 x 1 m plots were systematically sampled at 20 m intervals along the centre of the planted rows. Within each plot, six leaf LCC readings were taken at random, from leaves at mid-canopy level, and averaged to derive a plot-level average LCC value. LCC measurements were recorded using the SPAD 502 Plus chlorophyll meter.

The collected LCC measurements showed considerable variability across the sampled plots, ranging from 20.00 to 59.75 µmol m^{-2}. The mean (36.94 µmol m^{-2}) and the median (36.77 µmol m^{-2}) indicate a balanced distribution, while the standard deviation of 7.78 µmol m^{-2} suggests moderate within-field variation in chlorophyl (Table 1).

Statistic	Value
Sample Size	129
Min	20.00
Max	59.75
Mean	36.94
Median	36.77
Standard Deviation	7.78

Table 1. Summary statistics of field LCC (µmol m^{-2}) measurements for 24 and 25 January 2023.

Unitless SPAD unitless readings were converted to absolute chlorophyll content (µmol m^{-2}) by applying the calibration method proposed by Cerovic et al. (2012):

$$LCC = \frac{99 \times SPAD}{144 - SPAD} \quad (1)$$

where LCC = Absolute chlorophyll content (µmol m^{-2})
SPAD = Leaf chlorophyll reading.

The relationship between SPAD readings and derived chlorophyll content is illustrated in Figure 2 below.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-731-2026 | © Author(s) 2026. CC BY 4.0 License.

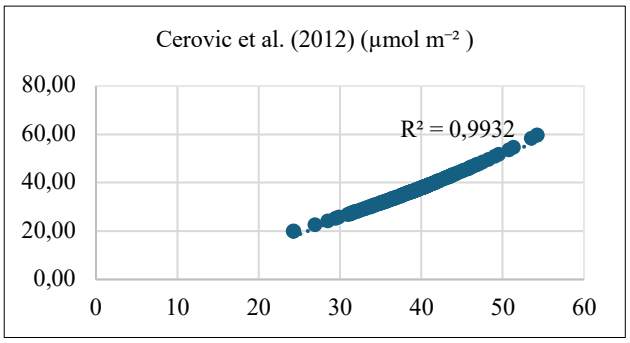

Figure 2. Relationship between SPAD readings and chlorophyll content (µmol m⁻²) using the Cerovic et al., (2012) conversion equation.

3.2 Selection of different soybean cultivars

This study monitored 15 different soybean cultivars obtained from four major commercial seed suppliers in South Africa, namely Don Mario, Pannar (Corteva), Pioneer, and Santa Rosa. Soybean is a short-day species that flowers when day length decreases below its critical photoperiod, therefore soybean cultivars are specifically bred for different regions according to latitude and photoperiod threshold (Sun et al., 2025, Liu et al., 2017).

The selected cultivars strategically capture variability in canopy structure, chlorophyll levels, and phenological development across different growth classes. These cultivars were selected by the host farmer, based on their proven performance and adaptability to local agro-ecological challenges in this region– such as shallow soil profiles and seasonal water stress. By including cultivars from various growth classes (Figure 3), the study allows for a more robust assessment of model behaviour and generalisability across cultivar-specific phenological and physiological variation.

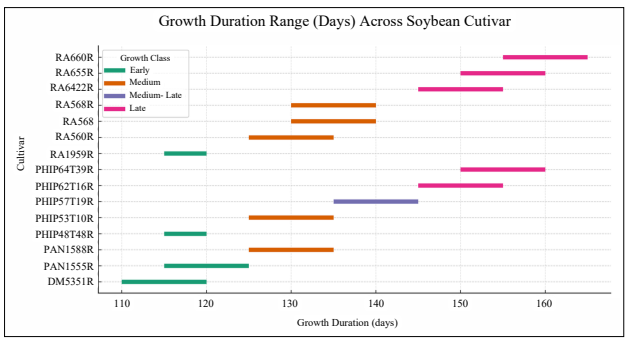

Figure 3. Growth duration distribution across selected soybean cultivars.

3.3 UAV imagery and pre-processing

Multispectral data for the study area were collected on 27 January 2023, using a DJI Matrice 300 UAV fitted with a Micasense RedEdge-P sensor and a downwelling light sensor (DLS-2). The RedEdge-P sensor recorded data in five multispectral bands namely: blue, green, red, red-edge and near-infrared. The sensor onboard the UAV was flown at a 110 m altitude, and yielded imagery with a spatial resolution of approximately 7 cm. In this study, four ground control points (white cross-boards) were positioned at the corners of the field. Following the UAV flight, the centre coordinates of these markers were surveyed using a handheld Carlson RTK GNSS instrument.

These high-accuracy coordinates were then used during pre-processing in Agisoft photogrammetry software to ensure precise georeferencing of the UAV orthomosaic. Calibration was supported by capturing images of the MicaSense Calibrated Reflectance Panel (CRP) before and after the flight, which were later processed in Agisoft to convert raw imagery to reflectance. The flight took place around midday (±12:00) under mostly clear conditions, with approximately 20% scattered cloud cover.

To improve data quality and minimise the influence of soil brightness and mixed pixels, the UAV image was processed in ArcGIS Pro to isolate only soybean canopy pixels. A soybean vegetation mask was generated by reclassifying pixels based on spectral thresholds, allowing for the removal of soil and other non-soybean features. This mask was then used to extract soybean pixels only, ensuring that subsequent analyses focus solely on canopy reflectance.

To obtain plot-level spectral information, 1 × 1 m polygons were generated to align with the field sampling design. Mean pixel values for each polygon were extracted in R software using the *raster* package. The resulting averaged values were then linked to the corresponding field chlorophyll measurements.

The UAV-derived multispectral reflectance values exhibit low reflectance in the visible bands (blue, green and red), a sharp increase in the red-edge region, and the highest reflectance in the near-infrared (NIR) region– characteristic of a typical spectral curve for green vegetation (Table 2). Reflectance ranged from 0.032 – 0.054 in the blue band, to 0.210 – 0.507 in the NIR band, with relatively low standard deviation suggesting limited within-plot variability.

	Blue (475 nm)	Green (560 nm)	Red (668 nm)	Red Edge (717 nm)	NIR (842 nm)
Min	0.032	0.066	0.047	0.158	0.210
Max	0.054	0.108	0.119	0.252	0.507
Mean	0.046	0.091	0.079	0.205	0.352
Median	0.047	0.092	0.080	0.204	0.349
StDev	0.004	0.008	0.013	0.021	0.054

Table 2. Descriptive statistics of plot-level reflectance values, extracted from the UAV-derived multispectral imagery for Blue, Green, Red, Red-edge, and NIR bands, on 27 January 2023.

The spectral signatures of cultivars (Figure 4) followed a similar trend, with all cultivars showing a rise in reflectance from the red to red-edge and NIR regions, typical of healthy soybean canopies during the early reproductive stage. Although the curves were tightly clustered, small separations were visible in the red-edge and NIR regions, suggesting subtle cultivar-level differences in canopy structure and chlorophyll-related reflectance. These differences are important because they may influence how LCC is detected from UAV-derived spectral data. This supports the need for robust modelling approaches that can distinguish between subtle spectral differences among cultivars.

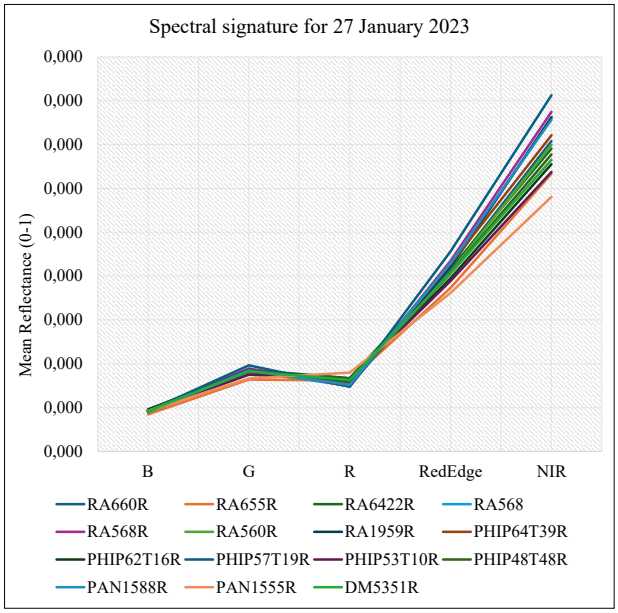

Figure 4. Graph showing the mean spectral reflectance for different soybean cultivars, during early reproductive phase (R2 – R3) as extracted from UAV imagery on 27 January 2023.

4. Methodology And Analysis

4.1 Overview of Methods

Figure 5 below summarises the methodological workflow followed in this study, including UAV image acquisition and preprocessing, field-based LCC measurements, vegetation index extraction, machine-learning model calibration and validation, and the generation of spatial LCC prediction maps.

Figure 5. Shows the workflow used to prepare the data, train the machine-learning models, and evaluate LCC prediction performance across soybean cultivars.

4.2 Vegetation Index Computation

Vegetation indices are created through linear or non-linear combinations of spectral bands. The combination of reflectance highlights various characteristics of vegetation (Kang et al., 2021). The study implemented a set of 16 vegetation indices, to capture different aspects of soybean reflectance. These included chlorophyll-sensitive such as Normalised Difference Vegetation Index (NDVI), Normalised Difference Red-Edge Index (NDRE), Modified Chlorophyll Absorption Ratio Index (MCARI), Transformed Chlorophyll Absorption Ratio Index (TCARI), CIgreen, and CIred-edge which use red, red-edge, and NIR wavelengths to approximate pigment concentration and photosynthetic capacity (Xue and Su, 2017). Greenness enhancing indices such as Green Normalised Difference Vegetation Index (GNDVI), Green Difference Vegetation Index (GDVI), Green Leaf Index (GLI), and Green Optimised Soil-Adjusted Vegetation Index (GOSAVI) are effective indicators of vegetation greenness, as they measure the degree to which vegetation reflects green or NIR light, relative to its absorption of visible wavelengths (Xue and Su, 2017). Simple ratio indices including Difference Vegetation Index (DVI), Simple Ratio Red-Edge Index (SRRE), and Ratio Vegetation Index (RVI), provide sensitivity to biomass and canopy density (Zhang et al., 2015), while soil-adjusted indices like Soil-Adjusted Vegetation Index (SAVI), Optimised Soil-Adjusted Vegetation Index (OSAVI), and GSAVI mitigate soil background effects in reflectance data (Zhen et al., 2021). Collectively, the combination of indices effectively summarises canopy biochemical and structural variability, promoting more accurate estimation of LCC content.

Vegetation indices were computed in R, through custom functions on the georeferenced raster stack, producing a series of index layers that highlight variations in canopy reflectance characteristics. These indices were then compiled into a single raster stack for extractions at plot level.

4.3 Machine Learning modelling, training and validation

Machine-learning algorithms offer a powerful approach for retrieving biophysical parameters, with strong potential to improve the accuracy and reliability of remote-sensing estimates (Kganyago et al., 2021, Mouafik et al., 2024). A key advantage of these methods is their ability to capture complex, non-linear relationships between target biophysical variables and spectral reflectance data and provide data-driven solutions to agricultural challenges (Jha et al., 2019, Kganyago et al., 2021).

Multispectral predictors and field-measured LCC were compiled into a modelling dataset in R, where cultivar was treated as a categorical variable and all spectral bands and vegetation indices were included as numeric predictors. After removing non-predictive variables and observations with missing values, the dataset was partitioned into training (70%) and testing (30%) subsets using stratified sampling to maintain a proportional representation of all cultivars. Feature selection was performed only on the training subset and included the removal of near-zero-variance predictors and variables exhibiting correlations above 0.95. Model training followed a consistent preprocessing pipeline comprising of median imputation, centring, and scaling of predictors, and model optimisation was carried out using repeated 10-fold cross-validation to ensure robust hyperparameter tuning and reduce overfitting.

This study evaluated both parametric (PLSR) and non-parametric (ANN, RF, XGBoost) modelling approaches to evaluate LCC prediction across soybean cultivars. Model performance was assessed on the independent test set using the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2), with additional cultivar-level evaluations to quantify variation in predictive accuracy across cultivar groups.

All models were implemented in R using the *caret* framework (Kuhn, 2008) with normal implementations of Random Forest

(Breiman, 2001) PLSR (Geladi and Kowalski, 1986) XGBoost (Chen and Guestrin, 2016) and artificial neural networks (Bishop, 1995).

5. Results And Discussion

5.1 Evaluation of cultivar-specific model performance

Cultivar-specific model evaluation highlights how predictive performance varies across soybean cultivars, reflecting differences in canopy structure and spectral responses. Assessing RMSE per cultivar provides a direct indication of prediction error and helps identify which cultivars are reliably modelled. As shown in Figure 6 below, the RMSE results demonstrate substantial variability in accuracy across cultivars and models. Tree-based approaches (RF and XGBoost) consistently achieve the lowest errors for several cultivars—for example, RF attains RMSE values of 3.44 µmol m^{-2} for PHIP62T16R, 4.66 µmol m^{-2} for PHIP57T19R and 4.18 µmol m^{-2} for RA560R, while XGBoost reaches 2.87 µmol m^{-2} for PHIP62T16R, 4.87 µmol m^{-2} for PHIP57T19R and 4.87 µmol m^{-2} for RA568. In contrast, ANN and PLSR show noticeably higher RMSE values for more challenging cultivars, including PAN1555R (12.55–13.71 µmol m^{-2}) and PAN1588R (9.42–9.65 µmol m^{-2}), reflecting more complex or less stable spectral–chlorophyll relationships. Overall, the RMSE patterns clearly demonstrate that model performance is strongly cultivar-dependent, with certain cultivars proving substantially easier to predict than others.

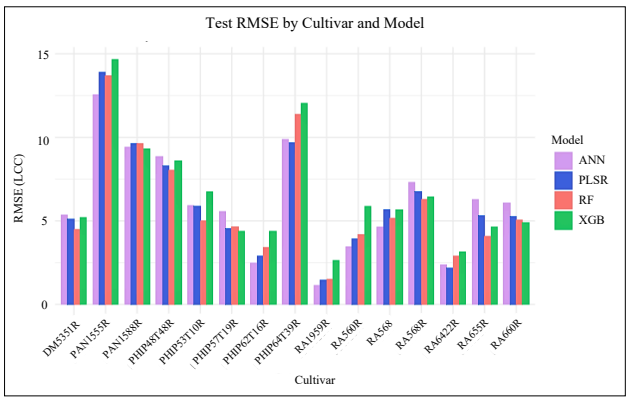

Figure 6. Test RMSE for each of the four models (ANN, PLSR, RF and XGB) across all cultivars, highlighting how prediction accuracy varies both by ML algorithm and cultivar.

The R² heatmap in Figure 7 below, further highlights strong cultivar-level differences in model performance. Across most cultivars, RF and XGB achieve significantly higher R² values—for example, RF reaches 0.79 for PHIP53T10R, 0.64 for RA568R, and 0.96 for RA655R, while XGB achieves similar or higher values, including 0.80 for PHIP53T10R and 0.94 for RA568. In contrast, PLSR and ANN show greater variability, with several cultivars dropping well below R² values of 0.3 and ANN returning no reliable R² estimates for multiple cultivars (e.g., DM5351R, PHIP48T48R, PHIP57T19R). Some cultivars, such as RA660, RA655R, and RA568, show a consistently strong agreement between observed and predicted LCC across models, whereas others (including PAN1555R, RA1959R, and RA6422R) show weaker relationships, reflecting greater within-cultivar variability and reduced spectral sensitivity. The R² results, suggest that non-linear models capture chlorophyll variation across cultivars more effectively than linear or parametric approaches.

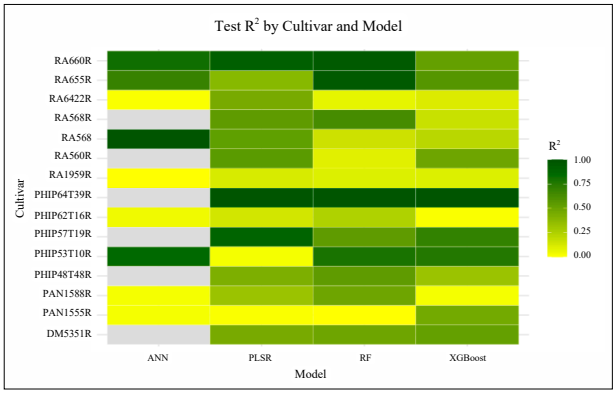

Figure 7. Test R² across cultivars and all four models (ANN, PLSR, RF and XGBoost), highlighting how model performance varies across cultivars and models.

Overall, the spread in RMSE and R² values across cultivars shows that the prediction accuracy is tightly linked to cultivar identity. Therefore, the results indicate that cultivar-specific spectral and physiological differences directly affect how consistently the models perform.

5.2 Evaluation of global and cultivar-level prediction error

The residual plots below (Figure 8 – Figure 11) were analysed using the full test dataset, with observations from all cultivars pooled together. As a result, the residuals reflect the overall behaviour of each model rather than cultivar specific error patterns. While this provides a useful indication of global model bias and the presence of systematic under- or over-estimation, it does not isolate which cultivars contribute most strongly to these patterns. The residual analysis is therefore interpreted alongside the cultivar-level results, which reveal the underlying variability in prediction performance across different soybean cultivars.

The resultant residual plots show clear differences in prediction behaviour across the four models. RF (Figure 8) and XGBoost (Figure 9) display residuals that are more tightly centred around zero, indicating relatively unbiased results across LCC. Although some scatter is present, these models do not show strong under- or over-prediction, suggesting more stable model behaviour.

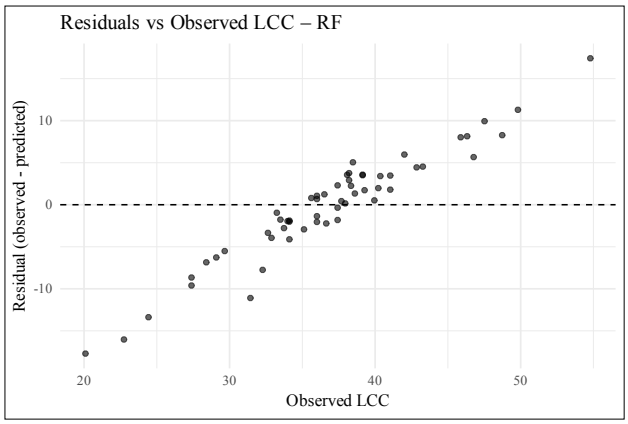

Figure 8. Residuals versus observed LCC values for RF model.

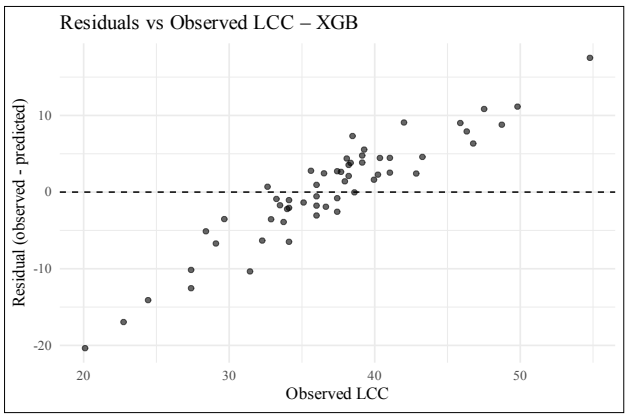

Figure 9. Residuals versus observed LCC values for XGBoost model.

In contrast, ANN (Figure 10) shows a clear residual pattern, with residuals increasing as observed LCC increases, suggesting that the model under-predicts at higher chlorophyll values. The ANN model also produces a narrower band of predictions clustered along a line, indicating reduced flexibility in capturing the full variability of LCC. The PLSR residual plot (Figure 11) shows a broader spread of residuals across the observed LCC range, indicating less consistent predictive performance compared with RF and XGBoost.

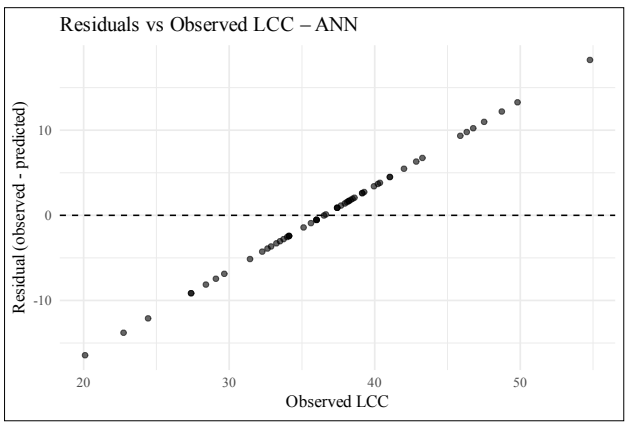

Figure 10. Residuals versus observed LCC values for ANN model.

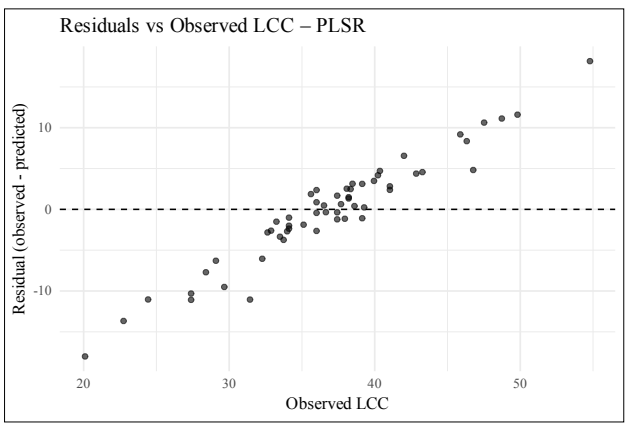

Figure 11. Residuals versus observed LCC values for PLSR model.

While the global residual plots provide an overview of general model behaviour, they do not reveal which cultivars contribute most strongly to the observed residual structure. The cultivar-specific residual distributions (Figure 12) show clear variation in error and bias among cultivars, with some exhibiting consistent under- or over-estimation across models.

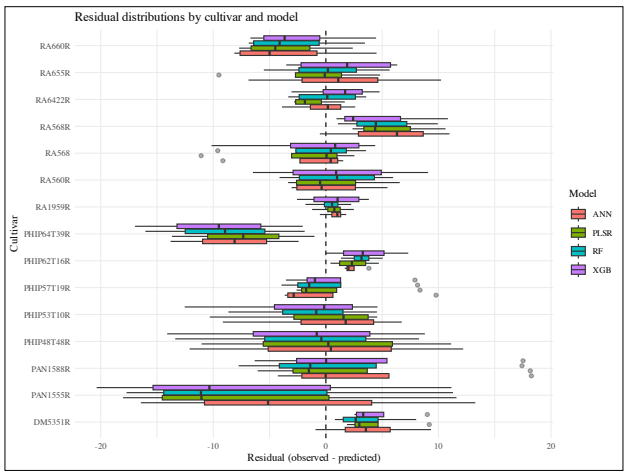

Figure 12. Residuals versus observed distribution by cultivar and model (ANN, PLSR, RF, XGBoost) for LCC.

The cultivar-specific residual distributions (Figure 11) reveal substantial heterogeneity in model performance across the 15 soybean cultivars. Several cultivars, such as PAN1555R, PHIP48T48R and PHIP57T19R, exhibit wide residual spreads and strong negative medians, indicating consistent overestimation of chlorophyll content and higher prediction uncertainty across all models. In contrast, cultivars such as DM5351R and PHIP62T16R show narrow, centred distributions, suggesting more stable model behaviour and fewer systematic errors. The presence of both positive and negative median shifts across cultivars demonstrates that prediction bias is not uniform, reflecting underlying differences in canopy structure, pigment composition and leaf internal anatomy.

These cultivar-level patterns align with the non-random residual structure observed in the global residual plots and indicate that a single pooled model struggles to fully capture cultivar-specific chlorophyll dynamics. Based on the results from the global residual plots, clear patterns in residuals reinforce that the tree-based models provide more reliable predictions and better capture the spectral–chlorophyll relationship. Additionally, the cultivar-level results, reinforces the observed differences in RMSE and R^2 at the cultivar level, confirming that spectral–chlorophyll relationships differ significantly across cultivars.

5.3 Variable Importance

To better understand which predictors, in terms of bands and VIs, contributed most strongly to LCC estimation, variable importance was derived for all four modelling approaches using a unified *caret*-based framework in R. For the tree-based models (RF and XGBoost), importance reflects each predictor’s contribution to reducing prediction error, while importance in PLSR and ANN was based on latent component loadings and network connection weights, respectively. Importance values were scaled to enable direct comparison across parametric and non-parametric algorithms.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-731-2026 | © Author(s) 2026. CC BY 4.0 License.

As seen in Figure 13 below, out of five spectral bands, and 16 VIs, the green and red-edge bands, together with TCARI and CIRedEdge indices, are repeatedly highlighted as the strongest predictors of LCC. These predictors show the highest scaled variable importance, especially under XGBoost and RF, indicating a strong nonlinear relationship between spectral features and LCC. In comparison, blue and red bands, together with indices like MCARI contribute far less. Overall, the results highlight the dominant role of red-edge and green spectral bands in chlorophyll prediction.

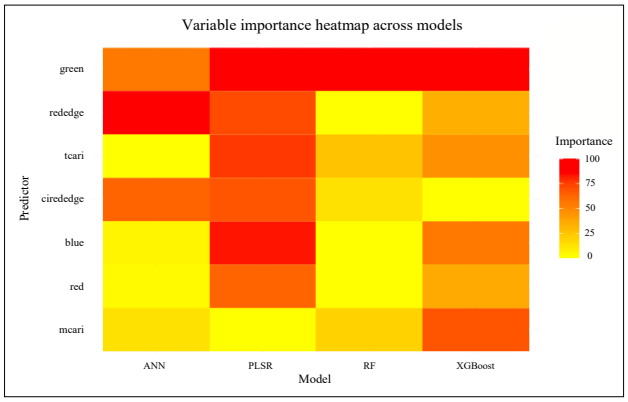

Figure 13. Heatmap comparing the relative importance of key spectral bands and VIs across the four models (ANN, PLSR, RF, XGBoost).

5.4 LCC Spatial Prediction Map

Among all four modelling approaches, XGBoost achieved the highest predictive accuracy and most stable performance across cultivars, consistently producing lower RMSE and higher R^2 values compared to the other machine-learning algorithms. Based on this performance, the XGBoost model was selected to generate the final spatial prediction map for soybean LCC (Figure 14) which illustrates clear within-field variability in chlorophyll content during the early reproductive stage (R2–R3). Higher LCC values (shown in green) cluster in the central and eastern portions of the field, while lower values (yellow to light brown) appear along field edges and zones likely associated with soil or moisture constraints.

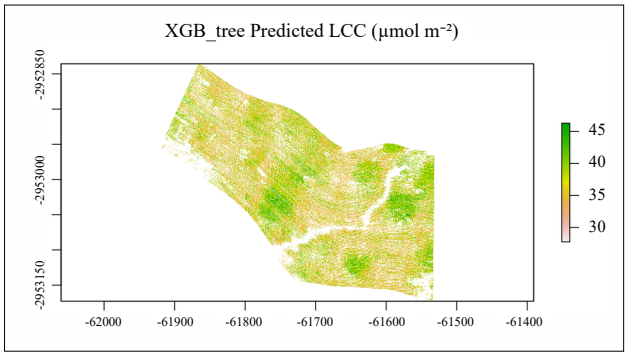

Figure 14. Spatial Distribution of XGBoost-Predicted LCC for the early reproductive (R2 – R3) stage.

The histogram (Figure 15) shows a unimodal distribution of predicted LCC, centred around ~35–40 µmol m^{-2}, indicating that most soybean plants fall within a moderate chlorophyll range at this stage. The slightly right-skewed tail reflects smaller patches of high-chlorophyll vegetation, whereas the thinner left tail represents localised low-chlorophyll stress areas. Together, the map and histogram highlight both the general crop condition and the spatial heterogeneity captured by the model.

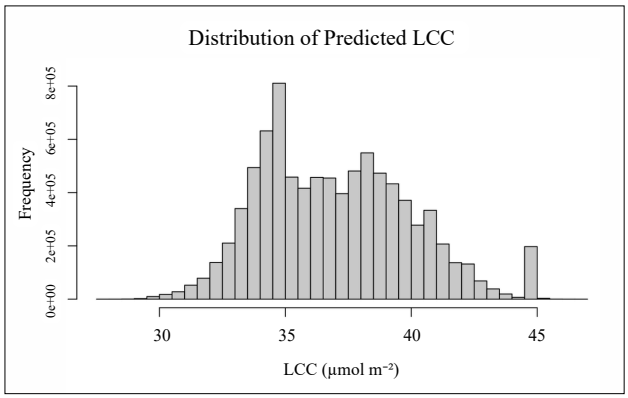

Figure 15. Distribution of XGBoost-Predicted LCC, across the study field for the early reproductive (R2 – R3) stage.

Overall, the spatial prediction map provides a coherent representation of canopy chlorophyll patterns within the field and demonstrates the model’s capacity to capture physiologically meaningful variation at high spatial resolution. These findings highlight the applicability of XGBoost for UAV-based chlorophyll estimation and provide a foundation for integrating chlorophyll maps into precision management decisions and future cultivar-specific analyses.

6. Conclusion

Across cultivars, model accuracy showed substantial variability, with the tree-based algorithms consistently providing the strongest predictive performance. XGBoost achieved the lowest RMSE and highest R^2 for several cultivars, for example PHIP62T16R (RMSE ≈ 2.9 µmol m^{-2}; R^2 = 0.16) and RA568 (RMSE ≈ 4.9 µmol m^{-2}; R^2 = 0.94). Random Forest performed similarly well for PHIP53T10R (RMSE ≈ 5.0 µmol m^{-2}; R^2 = 0.79) and RA655R (RMSE ≈ 4.1 µmol m^{-2}; R^2 = 0.96). In contrast, ANN and PLSR returned significantly higher errors for more cultivars such as PAN1555R (RMSEs in the range 12–14 µmol m^{-2}; R^2 < 0.10), demonstrating weaker modelling of their spectral–chlorophyll dynamics.

Residual results revealed non-random error patterns across the full dataset, with systematic over- and under-estimation for several cultivars, indicating that global models do not fully capture cultivar-specific physiological and spectral behaviour. Cultivars such as PAN1555R and PHIP48T48R demonstrated wide and biased residual distributions, while others (e.g., RA568, PHIP53T10R) showed tight, centred residuals, further highlighting that prediction reliability is strongly dependent on soybean cultivar.

Variable importance results showed that red-edge and NIR bands, alongside green enhancing indices, contributed most to prediction accuracy across models. These results are consistent

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-731-2026 | © Author(s) 2026. CC BY 4.0 License.

with the biochemical and structural controls on chlorophyll reflectance. These variables were especially influential in the ensemble models (RF and XGBoost), explaining their superior performance relative to linear and parametric approaches.

Together, the patterns seen in the RMSE and R^2 results, residual structure, and variable importance findings all demonstrate that cultivar-specific spectral characteristics strongly influence prediction performance, and that non-linear ensemble methods, particularly XGBoost and RF, offer the most robust and biologically meaningful estimates of chlorophyll across genetically diverse soybean cultivars, during the early reproductive phase.

7. Limitations And Future Recommendations

This study was constrained by a single growing season, a limited number of cultivars, and UAV data collected during one phenological window, which may restrict broader generalisation. Future research could incorporate multi-season datasets, additional growth stages, and a wider range of cultivars to improve model transferability. It may also be valuable to integrate Leaf Area Index (LAI) alongside chlorophyll to support multi-trait crop assessment. Further incorporating hyperspectral or multi-source satellite data could enhance robustness and operational relevance.

Acknowledgements

We would like to thank the University of Pretoria at the Department of Geography, Geoinformatics and Meteorology, and the Agricultural Research Council (ARC) (Natural Resources and Engineering) for their academic guidance and support throughout this study. We also gratefully acknowledge Farm Banabatau for granting access to the fields used for the experimental trials.

References

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,* 12.

BISHOP, C. M. 1995. *Neural Networks for Pattern Recognition,* Oxford, Claredon Press.

BOARD, J. & KAHLON, C. S. 2011. Soybean Yield Formation: What Controls It and How It Can Be Improved. *Soybean Physiology and Biochemistry.*

BREIMAN, L. 2001. Random Forests. University of California: Berkeley, CA 94720.

BREWER, K., CLULOW, A., SIBANDA, M., GOKOOL, S., NAIKEN, V. & MABHAUDHI, T. 2022. Predicting the Chlorophyll Content of Maize over Phenotyping as a Proxy for Crop Health in Smallholder Farming Systems. *Remote Sensing,* 14.

CEROVIC, Z. G., MASDOUMIER, G., GHOZLEN, N. B. & LATOUCHE, G. 2012. A new optical leaf-clip meter for simultaneous non-destructive assessment of leaf chlorophyll and epidermal flavonoids. *Physiol Plant,* 146**,** 251-60.

CHEN, T. & GUESTRIN, C. 2016. XGBoost: A Scalable Tree Boosting System.

CLUA, J., RODA, C., ZANETTI, M. E. & BLANCO, F. A. 2018. Compatibility between Legumes and Rhizobia for the Establishment of a Successful Nitrogen-Fixing Symbiosis. *Genes (Basel),* 9.

ENGELBRECHT, G., CLAASSENS, S., MIENIE, C. M. S. & FOURIE, H. 2020. South Africa: An Important Soybean Producer in Sub-Saharan Africa and the Quest for Managing Nematode Pests of the Crop. *Agriculture,* 10.

FEHR, W. R. 1977. Stages of Soybean Development. Iowa State University.

GELADI, P. & KOWALSKI, B. R. 1986. Partial least-squares regression: A Tutorial. *Analytica Chimica Acta,* Volume 185**,** 1-17.

HU, J., YUE, J., XU, X., HAN, S., SUN, T., LIU, Y., FENG, H. & QIAO, H. 2023. UAV-Based Remote Sensing for Soybean FVC, LCC, and Maturity Monitoring. *Agriculture,* 13.

JHA, K., DOSHI, A., PATEL, P. & SHAH, M. 2019. A comprehensive review on automation in agriculture using artificial intelligence. *Artificial Intelligence in Agriculture,* 2**,** 1-12.

KANG, Y., MENG, Q., LIU, M., ZOU, Y. & WANG, X. 2021. Crop Classification Based on Red Edge Features Analysis of GF-6 WFV Data. *Sensors (Basel),* 21.

KGANYAGO, M., MHANGARA, P. & ADJORLOLO, C. 2021. Estimating Crop Biophysical Parameters Using Machine Learning Algorithms and Sentinel-2 Imagery. *Remote Sensing,* 13.

KUHN, M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software,* 28.

LIU, X., WU, J. A., REN, H., QI, Y., LI, C., CAO, J., ZHANG, X., ZHANG, Z., CAI, Z. & GAI, J. 2017. Genetic variation of world soybean maturity date and geographic distribution of maturity groups. *Breed Sci,* 67**,** 221-232.

MOUAFIK, M., FOUAD, M., AUDET, F. A. & EL ABOUDI, A. 2024. Comparative analysis of multi-source data for machine learning-based LAI estimation in Argania spinosa. *Advances in Space Research,* 73**,** 4976-4987.

PASQUALOTTO, N., BOLOGNESI, S. F., BELFIORE, O. R., DELEGIDO, J., D'URSO, G. & MORENO, J. 2019. Canopy chlorophyll content and LAI estimation
from Sentinel-2: vegetation indices and Sentinel-2
Level-2A automatic products comparison. *IEEE International Workshop on Metrology for Agriculture and Forestry.* Portici, Italy: IEEE.

ROTH, M. G., NOEL, Z. A., WANG, J., WARNER, F., BYRNE, A. M. & CHILVERS, M. I. 2019. Predicting Soybean Yield and Sudden Death Syndrome Development Using At-Planting Risk Factors. *Phytopathology,* 109**,** 1710-1719.

SEDIBE, M. M., MOFOKENG, A. M. & MASVODZA, D. R. 2022. *Soybean - Recent Advances in Research and Application*, IntechOpen.

SHI, H., GUO, J., AN, J., TANG, Z., WANG, X., LI, W., ZHAO, X., JIN, L., XIANG, Y., LI, Z. & ZHANG, F. 2023. Estimation of Chlorophyll Content in Soybean Crop at Different Growth Stages Based on Optimal Spectral Index. *Agronomy,* 13.

SUN, H., SHEN, S., YANG, J., ZOU, J., HARRISON, M. T., WANG, Z., HU, J., GUO, H., UMBURANAS, R. C., ZHAI, Y., WEN, X., CHEN, F. & YIN, X. 2025. Soybean Cultivar Breeding Has Increased Yields Through Extended Reproductive Growth Periods and Elevated Photosynthesis. *Plants (Basel),* 14.

SWIERK, L. N. 2024. *Highveld grasslands* [Online]. EBSCO. Available: https://www.ebsco.com/research-starters/earth-and-atmospheric-sciences/highveld-grasslands [Accessed].

WEISS, M., JACOB, F. & DUVEILLER, G. 2020. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment,* 236.

XUE, J. & SU, B. 2017. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *Journal of Sensors,* 2017**,** 1-17.

ZHANG, L., SHAO, Z. & DIAO, C. 2015. Synergistic retrieval model of forest biomass using the integration of optical and microwave remote sensing. *Journal of Applied Remote Sensing,* 9.

ZHEN, Z., CHEN, S., YIN, T., CHAVANON, E., LAURET, N., GUILLEUX, J., HENKE, M., QIN, W., CAO, L., LI, J., LU, P. & GASTELLU-ETCHEGORRY, J. P. 2021. Using the Negative Soil Adjustment Factor of Soil Adjusted Vegetation Index (SAVI) to Resist Saturation Effects and Estimate Leaf Area Index (LAI) in Dense Vegetation Areas. *Sensors (Basel),* 21.