

Optimizing Sentinel-2 feature space for improved crop biophysical and biochemical variables retrieval using the novel spectral triad feature selection algorithm

Mahlatse Kganyago, Clement Adjorlolo & Paidamwoyo Mhangara

To cite this article: Mahlatse Kganyago, Clement Adjorlolo & Paidamwoyo Mhangara (2024) Optimizing Sentinel-2 feature space for improved crop biophysical and biochemical variables retrieval using the novel spectral triad feature selection algorithm, Geocarto International, 39:1, 2309174, DOI: [10.1080/10106049.2024.2309174](https://doi.org/10.1080/10106049.2024.2309174)

To link to this article: <https://doi.org/10.1080/10106049.2024.2309174>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 13 Feb 2024.



Submit your article to this journal [↗](#)



Article views: 113



View related articles [↗](#)



View Crossmark data [↗](#)



Optimizing Sentinel-2 feature space for improved crop biophysical and biochemical variables retrieval using the novel spectral triad feature selection algorithm

Mahlatse Kganyago^{a,b} , Clement Adjorlolo^{b,c} and Paidamwoyo Mhangara^b

^aDepartment of Geography, Environmental Management and Energy Studies, University of Johannesburg, Johannesburg, South Africa; ^bSchool of Geography, Archaeology and Environmental Studies, University of the Witwatersrand, Johannesburg, South Africa; ^cAfrican Union Development Agency (AUDA-NEPAD), Johannesburg, South Africa

ABSTRACT

This study presents a novel Spectral Triad feature selection (STfs) technique based on music theory and compares it to the entire Sentinel-2 feature space and Random Forest-Recursive Feature Elimination (RF-RFE). The optimal subsets were evaluated with Random Forest for retrieving Leaf Area Index (LAI), Leaf Chlorophyll Content (LC_{ab}), and Canopy Chlorophyll Content (CCC) in a semi-arid agricultural landscape. The results indicated that the proposed STfs algorithm obtained equivalent or better (i.e. by 1 – 3%) retrieval accuracies for LAI (R^2_{cv} of 66%, root mean squared error of cross-validation [$RMSE_{cv}$] of $0.53 \text{ m}^2 \text{ m}^{-2}$), LC_{ab} (R^2_{cv} : 74%, $RMSE_{cv}$: $7.09 \mu\text{g cm}^{-2}$) and CCC (R^2_{cv} : 77%, $RMSE_{cv}$: $33.69 \mu\text{g cm}^{-2}$), using only 5, 7 and 7 variables, respectively, when compared to RF-RFE and entire Sentinel-2 feature space. Overall, the proposed STfs algorithm has great potential to optimize the spectral feature space of quasi-hyperspectral sensors for rapid crop biophysical and biochemical parameter retrieval.

ARTICLE HISTORY

Received 3 November 2023
Accepted 18 January 2024

KEYWORDS

Remote sensing
; feature selection;
chlorophyll content; leaf
area index; sentinel-2;
random forest; biophysical
and biochemical variables

1. Introduction

The retrieval of biophysical and biochemical variables (BVs) over agricultural areas has become a critical task in recent years due to the need to optimise yield and farm inputs and promote sustainability of the agro-ecological landscapes. The BVs such as Leaf Area Index (LAI), Leaf Chlorophyll $a + b$ (LC_{ab}) and Canopy Chlorophyll Content (CCC) have been identified as the critical plant growth indicators, critical to supporting precision agriculture technologies for fertiliser and pesticides application and irrigation management. LAI, defined as one-sided leaf area per unit area (i.e. $\text{m}^2 \text{ m}^{-2}$), is a significant plant structural parameter indicative of the different growth conditions and phenology of the crops, while LC_{ab} and CCC are essential to detect crop stress due to biotic and abiotic stressors,

CONTACT Mahlatse Kganyago kganyagoml@gmail.com, mahlatsek@uj.ac.za

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

and gross primary productivity (Gitelson et al. 2014) and is highly correlated to the Nitrogen (N) content. N content is one of the most limiting factors in plants and it is critical for crop growth, health, and yield (Daughtry et al. 2000; Gitelson et al. 2003). Remotely sensed parameters of plant structural properties and N content such as LAI, LC_{ab} , and CCC, reduce the need for destructive, laborious, and costly direct estimation and lab-based methods, thus crucial for optimizing irrigation, N fertilization for site-specific management and yield (Vincini et al. 2016; Jia et al. 2013).

To date, the approaches for the retrieval of BVs can be divided into four broad categories, i.e. spectral vegetation indices (SVIs), Radiative Transfer Models (RTMs), Machine Learning Regression Algorithms (MLRAs) and hybrid techniques integrating the MLRAs and Look-up-Tables (LUTs) generated using RTMs (Frampton et al. 2013). Among these, hybrid techniques have the greatest potential for wide application as the RTM-simulated LUTs can represent varying acquisition conditions and crop structural, biochemical, and physiological conditions, while MLRAs can estimate the complex and non-linear relations between the response and predictor variables. However, the RTMs require substantial site-specific data (Verrelst et al. 2012), and their universality claim has not been yet fully explored, as studies assessing the transferability of RTMs are often in similar climatic environments (e.g. Mediterranean climates). Besides, operational techniques such as Artificial Neural Networks (ANN) and PROSAIL-generated LUTs were found to perform poorly in semi-arid regions (Kganyago et al. 2020). Recent studies (Verrelst et al. 2012; Verrelst et al. 2015; Kganyago et al. 2021) show that simpler tree-based and kernel-based MLRAs can perform equally well as complex algorithms (e.g. ANN). Therefore, the power of MLRAs can be combined with real data consisting of various informative variables.

The optimization of the feature space, i.e. selection of informative feature subset, for BV retrieval is critical to reducing errors and uncertainties, thus increasing trust in remotely sensed BVs for farm decisions and site-specific management. Feature subset selection is commonly applied in studies utilizing hyperspectral data, which is characterized by many (i.e. thousands of) narrow and adjacent spectral bands. Some of these studies (Verrelst et al. 2012; Verrelst et al. 2016) have shown that only four to nine bands, selected through feature selection techniques, are sufficient to achieve robust retrievals of BVs. For example, Verrelst et al. (2016) developed the Gaussian Process Regression Band Analysis Tool (GPR-BAT) based on the Sequential Backward Band Removal (SBBR) algorithm, where the GPR prediction error in response to a changing number of explanatory variables (i.e. from all to one) is evaluated. In estimating Crop Water Content, LAI and LC_{ab} using field spectroscopy and airborne hyperspectral sensor, i.e. HyMap, the authors found that the best-performing models for each parameter had 6, 4, and 9 spectral bands, respectively. Another study, Xu et al. (2022), used GPR feature optimization and RF-RFE (Recursive Feature Elimination) algorithm to select the most relevant and informative Sentinel-1 and -2 features for biomass modeling. Their study found that the accuracy of biomass estimation improved significantly when feature selection methods were employed. However, hyperspectral data are limited to small areas and are costly to acquire. In the current study, we argue that feature subset selection is also relevant for retrieving crop BVs from multispectral data due to the high dimensionality of various possible covariates that can be derived, such as spectral bands, SVIs, and textural measures. Moreover, the advent of quasi-hyperspectral sensors such as Sentinel-3, Worldview-2 and its successors, and Sentinel-2 Multi-Spectral Instrument (MSI)—characterized by many multispectral broad and narrow bands that are strategically positioned for vegetation characterization (Fitzgerald et al. 2010; Delegido et al. 2013; Dimitrov et al. 2019)—may introduce collinearity, which has been shown to affect MLRA performance (Yu & Liu, 2004). Moreover,

redundant variables in the dataset introduce additional computational burden to an already costly process of tuning hyperparameters, and training and testing the machine learning models.

Generally, the optimization of variables for retrieval of crop BVs has been limitedly explored, especially for quasi-hyperspectral sensors. In fact, previous studies (Kganyago et al. 2021; Kganyago et al. 2022; Kganyago et al. 2022) either applied the entire feature space (i.e. visible, near-infrared, and shortwave infrared) from such sensors or tested VNIR (i.e. S2–10m) and RE-SWIR (i.e. S2–20m) configurations of Sentinel-2. Others combined all spectral bands with derived co-variables such as SVIs and textural measures, which are often arbitrarily selected or based on prior knowledge (Sibanda et al. 2017). For example, in estimating grassland LAI, LI et al. (2017) integrated Landsat Operational Land Imager (OLI) bands with seven vegetation indices and found improvements in accuracy once the variables were optimized, while Ramoelo et al. (2015) utilized WorldView-2 spectral bands and vegetation indices as input to an RF model in estimating grass quality indicators, i.e. leaf N and above-ground biomass, to R^2 of 89% and 84%, respectively. However, selecting covariates based on prior knowledge is beset with subjectivity and uncertainty due to the complex and interacting environmental and climatic factors that affect biophysical and biochemical traits in vegetation canopies, thus the canopy reflectance. In agricultural environments, the selection of relevant covariates using prior knowledge is further complicated by the varying agricultural management practices, phenology, crop types, and limiting factors such as soil nutrients and water. Therefore, statistical, data-based feature selection techniques are highly sought due to their capability to associate each or a set of covariates to the BVs of interest and assign a rank (e.g. importance) based on linear and non-linear relationships among covariates or between covariates and BVs. The statistical, data-based approaches, commonly classified as filter-based (e.g. analysis of variance), wrapper-based (e.g. Recursive Feature elimination-Support Vector Machine), and embedded (e.g. sparse Partial Least Squares) algorithms, are attractive for optimizing crop BVs for site-specific management in various environments and field conditions and are more objective since they determine the few optimal variables based on the information content of the covariates, while also eliminating collinearity and reducing training time. It is, therefore, essential to evaluate existing and develop new approaches to optimize crop BV retrieval with quasi-hyperspectral sensors such as Sentinel-2 MSI.

The aim of this study was to optimize crop BV retrieval using feature selection techniques to identify and select relevant remotely sensed covariates from Sentinel-2 data. To achieve this, we propose a novel Spectral Triad feature selection (STfs) algorithm inspired by music theory. The STfs algorithm was compared to the entire MSI feature space (i.e. no variable selection) and a popular wrapper-based algorithm, i.e. Recursive Feature elimination, coupled with Random Forest (RF-RFE) (Gregorutti et al. 2017) in terms of their accuracy and training times in retrieving LAI, LC_{ab} , and CCC. Moreover, we assessed the consistency in the selected feature subsets between STfs and RF-RFE in a semi-arid agricultural landscape, i.e. Bothaville, South Africa. The characteristics of the study area were described in our previous works (Kganyago 2021; Kganyago et al. 2021; Kganyago et al. 2022; Kganyago et al. 2022).

2. Materials and methods

Figure 1 summarizes the methods of the study. Generally, *in situ* data consisting of observed BVs and GPS coordinates of the plot centroid (subsection 2.1) were used to

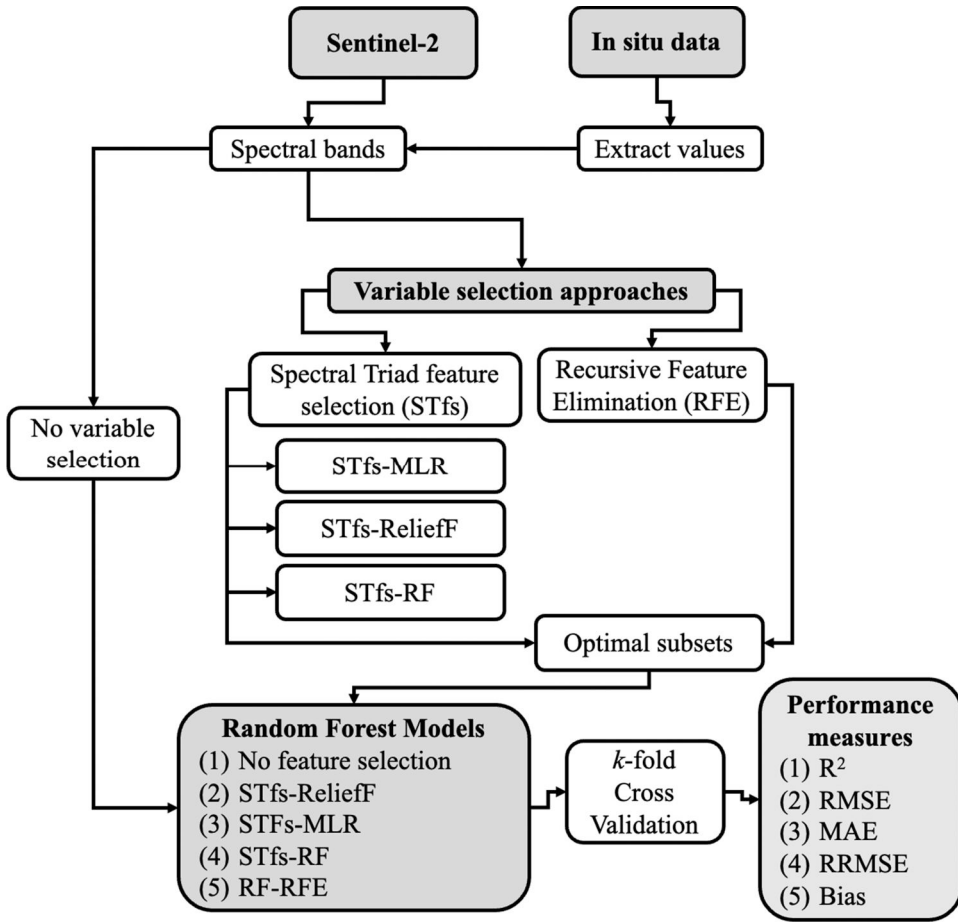


Figure 1. An overview of the data and methods used in this study.

extract spectral values using a $4m \times 4m$ pixel block from the Sentinel-2 image. The extracted tabular data were then subjected to the various variable selection techniques, including the Spectral Triads feature selection algorithm (subsection 2.2) and Recursive Feature Elimination (subsection 2.3), yielding optimal spectral bands for each BV. Next, the optimal spectral bands from variable selection techniques and the entire dataset (i.e. No feature selection) were used to build BV prediction models using the Random Forest algorithm and 10-fold cross-validation (subsection 2.4).

2.1. Data

The experimental data for this study was collected in Bothaville between April 11th and 23rd, 2021. The measured crop parameters consisted of LAI and LC_{ab} , using the Plant Canopy Analyzer (Li-Cor 2200c, LiCor Inc., Lincoln, NE, USA) and Chlorophyll Concentration Meter (MC-100 Apogee Instruments, Inc., Logan, UT, USA), respectively. The Plant Canopy Analyzer was equipped with a view restricting cap of 180° FOV to reduce the operator's influence and that of neighboring objects on the LAI measurements. On the contrary, LC_{ab} samples were acquired on sun-exposed leaves, where any sampling point within the plot comprised a mean of several (i.e. 3 to 5) leaf measurements. For

each 40 m × 40 m plot—selected on a random transect within various crop fields—an average of six to eight random sampling points were measured for both LC_{ab} and LAI. In total, 351 plots were sampled (Table 1). The third parameter studied here, i.e. CCC, was obtained as $LC_{ab} \times LAI$ (Jacquemoud et al. 1995). The plots were Geo-tagged with a GPS coordinate and a picture using TDC600 (Trimble Inc., Irvine, CA, USA) with an accuracy of 1.5 m. Table 1 shows the descriptive statistics of the measurements.

A Sentinel-2 image was collected on 14-04-2021 and was close to fieldwork dates. The Level-2A image (i.e. atmospherically corrected) was retrieved from Sentinel Hub Cloud API (Sinergise Laboratory for geographical information systems, Ltd., Ljubljana, Slovenia). We used 10 m spectral bands located at 490 nm (B2), 560 nm (B3), 665 nm (B4), and 842 nm (B8), and 20 m bands at 705 nm (B5), 740 nm (B6), 783 nm (B7), 865 nm (B8A), 1610 nm (B11), and 2190 nm (B12). To standardize the various resolutions, the 10 m bands were converted to 20 m using nearest neighbor resampling technique in SNAP Toolbox for further analysis.

The selected bands were masked using a crop mask, and non-vegetated pixels were masked using an NDVI threshold of 0.2 following Kganyago et al. (2021).

2.2. Spectral Triads feature selection algorithm

For decades, algorithm development has been inspired by nature with a range of algorithms developed based on the structures of the trees (e.g. Classification and Regression Trees, CART), tree ensembles (i.e. Random Forests, RF), neural activity (e.g. Artificial Neural Networks, ANN), and genetic mutations (i.e. Genetic Algorithm, GA). For example, the popular RF algorithm for the classification and regression of remotely sensed images is based on CART but builds many trees instead of one to form a forest. On the other hand, GA was designed using theories of natural selection and population genetics mechanisms (Ye 2018), while ANN uses the natural behavior of animals.

The Spectral Triads feature selection (STfs) algorithm, introduced here, aims to select the most harmonious (i.e. informative) subset of features (e.g. spectral bands) from multi-spectral data. The STfs algorithm is based on music theory, specifically, the major chord formula, which consists of intervals 1-3-5 on a diatonic scale. A chord consists of three or more pitches (or notes) played simultaneously; a diatonic scale is any scale with seven pitches in an octave and consists of five whole steps (or tones) and two half steps (or semitones) in each octave. The major chords, i.e. 1-3-5, consist of the root note (i.e. starting note), a major third (3rd), and the perfect fifth (5th). Therefore, the intervals of the major chords (on a diatonic scale) inspire the formulation of the STfs algorithm because their formula remains the same regardless of the root note and ensures a pattern where there is a maximal separation between the half steps (i.e. adjacent variables) in multiple octaves. In the context of the STfs algorithm, this interval ensures that each spectral triad contains variables that are adequately separated (i.e. spectrally), thus avoiding excessive collinearity in evaluating the triad. The STfs algorithm uses the major triad chord formula, where 1-3-5 represents the starting variable (root note, R_n) and the third and fifth

Table 1. Descriptive statistics of measured LAI (m² m⁻²), LC_{ab} (μg cm⁻²), and CCC (μg cm⁻²). n represents the total number of plots.

	n	Min	Mean	Max	SD	Range	Coef. var
LAI	351	1.16	3.46	6.17	0.89	5.01	0.26
LC_{ab}	351	3.32	28.38	63.62	12.79	60.30	0.45
CCC	351	7.87	100.46	339.09	58.85	331.23	0.59

variables, respectively. The R_n refers to the starting variable (an index of the explanatory variables in the dataset), which is 1 by default. The octaves, in the context of STfs, are the search iterations or sequences (progressions) that stop when there are no more unique spectral triads to evaluate or all the explanatory variables in the dataset are evaluated (i.e. equivalent to when all notes on the diatonic scale are played).

For each triad, an evaluation criterion is used to rank the explanatory variables where the best single variable in a triad is selected. The evaluation criteria in the filter mode of STfs (Algorithm 1) can be an Entropy-based measure (e.g. Information gain) or ReliefF while in wrapper mode, it uses an explainable machine learning algorithm such as Multiple Linear Regression (MLR) or RF Variable Importance measure (Percent Increase in Mean Squared Error, %IncMSE). The overview of the algorithm is outlined in Figure 2.

The STfs algorithm then returns an optimal subset of features $m_{optimal}$ i.e. optimal features selected from each triad, which is fewer than the original input features. Indeed, depending on the dimensionality of the dataset, the number of evaluated spectral triads and selected variables in a triad may still retain high dimensionality, and cause a loss in

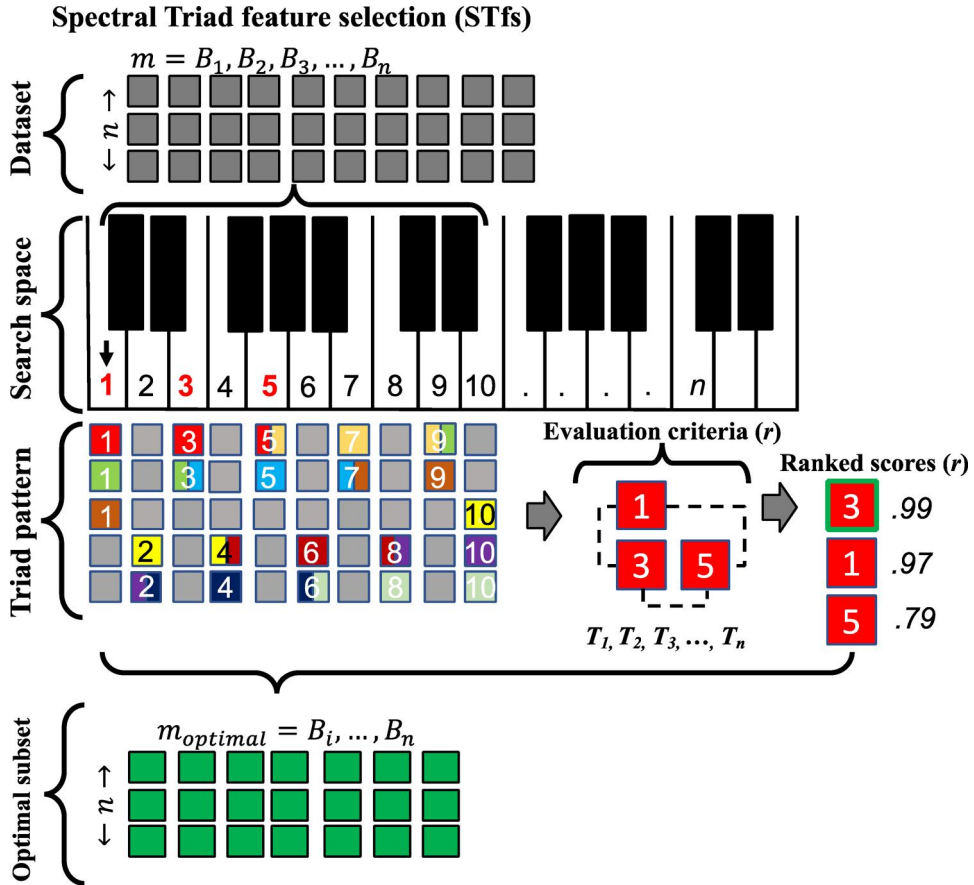


Figure 2. An overview of the spectral triads feature selection (STfs) algorithm, demonstrating a case of variable subset selection with Sentinel-2's ten spectral bands as input data matrix ($m \times n$, where m is the Sentinel-2 bands $[B_i]$, and n is the number of observations) superimposed on ten keys to form a search space for the algorithm and correlation analysis r as an evaluation criterion. All possible spectral triads (T) are evaluated using the search space, with the stopping criterion reached when all triads are evaluated. A final subset consists of all best bands (i.e. with the highest rank scores) from each triad (i.e. $m_{optimal}$). The colors indicate the pattern of the different triads across the feature space and overlaps of bands per triad."

computational efficiency. Therefore, in such cases, an optimal desired number of variables or subset size (Pd) can be defined. If this value is different from the default $m_{optimal}$ (i.e. all optimal features selected from each triad), the algorithm will be triggered to form and evaluate new triads from the default $m_{optimal}$ until Pd is achieved. If features in the triad cannot be distinguished using the evaluation metric or the evaluation criteria (see subsubsection 2.2.2), e.g. r , R^2 , %IncMSE, or the p -value, the entire triad is discarded. The proposed algorithm steps are given in [Algorithm 1](#).

Given the training data with the number of covariates n , the STfs routine in the filter or wrapper modes (i.e. combined with an explainable learning algorithm such as RF) is outlined below.

Algorithm 1. Spectral Triad feature selection

1. **for** Number of covariates n **do**
 2. Choose the Root node R_n (from 1 to n of variables),
 3. Based on the R_n , select the third R_3 and fifth R_5 notes (i.e., variables) to form a triad T_1 ,
 4. Evaluate the relationship of T_1 to the response variable y using ReliefF as an evaluation criterion OR Train the MLR algorithm or Random Forest Model using T_1 to T_n and the relevant response variable y ,
 5. Rank and select the most significant (i.e., influential) variable R_i , i.e., one with the lowest p -value of F-Statistic or highest importance score,
 6. Repeat steps 3 to 5, until all triads T_n are evaluated,
 7. Return all significant or important variables from T_1 to T_n with their triad of origin T_0 and p -value or importance scores,
 8. **End for**
-

2.2.1. Evaluation criteria

The STfs algorithm can accommodate various evaluation criteria, determining if the algorithm is used in a filter or wrapper mode. For example, Pearson's correlation coefficient r (or its non-parametric alternative, Spearman's rank coefficient ρ), Entropy-based algorithms (such as Information gain and ReliefF) can be used in the filter mode, while an explainable model such as MLR (or its non-parametric alternative, Generalized Additive Models, GAM) and RF can be used in the wrapper mode. To demonstrate the performance of STfs in filter mode, we used the ReliefF algorithm, while in wrapper mode, we used MLR and RF variable importance as evaluation criteria for each triad. Below is a short description of the evaluation criteria considered here.

2.2.2. ReliefF

The ReliefF algorithm was developed to overcome the limitations of the original Relief algorithm (Robnik-Šikonja and Kononenko 2003), which was limited to classification problems and could not deal with missing values in the data. Relief and its adaptations aim to estimate the quality of variables according to how well they can distinguish close observations. To achieve this, the original Relief algorithm first randomly selects an observation R_i , then determines two nearest observations (i.e. neighbors), where one is from the same class as R_i , i.e. nearest hit H , and the other is from a different class, i.e. nearest miss M . Then, it estimates and updates the quality $W[A]$ of all variables A depending on their values for R_i , H , and M . This process is repeated for m times, defined by the user.

Instead of only two nearest neighbors (i.e. H and M), the extension of Relief, i.e. ReliefF, uses k nearest H and M, and the final $W[A]$ of A is determined by averaging the contribution of k . ReliefF was later adapted for regression problems (called RReliefF), where a probability that the predicted values of two observations are different is used instead of whether they belong to the same class for classification problems (Robnik-Šikonja and Kononenko 2003). Further details can be found in relevant publications (Kononenko et al. 1996; Robnik-Šikonja and Kononenko 2003). RReliefF was performed in R-Statistic software using the ‘FSelectorRcpp’ package (Zawadzki and Kosinski 2020; Romanski 2021), which is a reimplementation of the ‘FSelector’ package (Romanski 2021), but eliminates dependence on JAVA/WEKA and allows parallel processing and sparse matrix.

2.2.3. Multiple linear regression analysis

Multiple Linear Regression (MLR, Equation 1) determines the most influential predictor variables (x) according to the amount that each predictor variable (y) reduces the residual sum of squares.

$$y = m_1x_1 + m_2x_2 + \dots + m_nx_n + b \quad (1)$$

Where y is the response variable (i.e. LAI, LC_{ab} , CCC), and $x_{1\dots n}$ are predictor variables, m_1 represent regression coefficients, while b represents the intercept. The regression coefficient, i.e. the partial derivative of the response variable for each predictor variable, measures the linear sensitivity of y to inputs $x_{i\dots n}$ (Mohanty and Codell 2002). The coefficient of determination, adjusted for the degree of freedom (adjusted R^2), was used to determine the relationship between the response variable and predictor variables in each triad, while the F-statistic p -value was used to determine the significance of the relationships at $\alpha = 0.05$ and rank the most significant explanatory variable.

2.3. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a greedy optimization wrapper-based algorithm, which sequentially removes the least important variables while retaining a few and an optimal subset of variables. Because it uses a machine learning regressor to evaluate each predictor variable for its predictive power in the modeling of a response variable, it falls under the category of wrapper-based variable selection approaches and can be combined with various MLRAs. In this way, it starts with a full set of variables to iteratively build regression models, while progressively eliminating the variables that do not improve the estimation errors of the model at each iteration until all features are exhausted. Consequently, RFE reduces dimensionality and eliminates dependencies between predictor variables. The performance of each model is assessed using RMSE and the cross-validation resampling technique (Demarchi et al. 2020). Previous studies (Demarchi et al. 2020; Georganos et al. 2018; Kganyago et al. 2017; Rajasheker R Pullanagari et al. 2018) have combined the RFE algorithm with Support Vector Machine (SVM-RFE) and Random Forest (RF-RFE) in classification and regression problems. In this study, we used the RF-RFE, where: at each iteration, the algorithm eliminates the variables with the least importance in the RF models based on a 10-fold cross-validation strategy.

2.4. Modelling and prediction accuracy assessment

2.4.1. Random Forest

Utilizing R-statistics software and the 'randomForest' package, the retrieval of BVs under consideration here, LAI, LC_{ab} , and CCC, was carried out (Breiman et al. 2018). RF uses bagging to create several decision trees repeatedly and independently from a set of random training samples produced by resampling with replacement from the original sample (Fawagreh et al. 2014; Breiman 2001). During RF modelling, approximately 64% of the training data are kept for model building (i.e. in-bag samples), while the remaining 36% (i.e. out-of-bag or OOB samples) are used for internal model performance evaluation and variable importance calculations (Gislason et al. 2006). When permuting OOB samples while holding all the other factors constant, the %IncMSE (Percent Increase in Mean Squared Error) is calculated to determine the importance of each explanatory variable. %IncMSE, a ranking metric for the various variables, is used to evaluate the variables' predictive ability. For instance, a variable might be considered significant if its exclusion from the model significantly lowers prediction accuracy. In essence, a high importance score for an explanatory variable implies a strong association with the response variable or predictive power (Mutanga et al. 2012; Ramoelo et al. 2015). The RF variable importance measure was used in the STfs algorithm.

For modelling, the optimal tuning parameters *ntree* (i.e. the number of trees) and *mtry* (i.e. the number of variables), were determined using a grid-search strategy by testing all combinations of *ntree* values from 100 to 500 with an interval of 10 and *mtry* values from 1 to *p* (i.e. number of explanatory variables) with a single interval. Five models per BV (LAI, LC_{ab} and CCC) were built, i.e. STfs-ReliefF, STfs-MLR, STfs-RF, RF-RFE and No feature selection, totalling 15 models.

2.4.2. Model performance assessment

The prediction errors of each subset of features or the entire feature space were assessed with the Coefficient of determination (R^2_{cv}), Root Mean Squared Error ($RMSE_{cv}$), and Mean Absolute Error (MAE_{cv}), Relative RMSE ($RRMSE_{cv}$) and Bias ($BIAS_{cv}$) of a 10-fold cross-validation. In 10-fold cross-validation, the dataset is randomly divided into 10 equal data subsets, where nine data subsets serve as the training set and one data subset as a validation set. The regression model is trained 10 times. For each training instance, one of the subsets is omitted from training and used only to assess the regression accuracy using the described metrics. The final reported metrics are an average of validation sets omitted from each training instance. In contrast to the traditional 70/30 (i.e. training/validation), the 10-fold cross-validation ensures that all data are used for training and validation (Verrelst et al. 2015; Shah et al. 2019). The R^2 is a correlation-based metric that determines how much of the variance of the response variable, such as LAI, is explained by the explanatory variables (e.g. spectral bands). The RMSE and MAE, on the other hand, measure the amount of error between the predictions and observations in the units of the BV, i.e. $m^2 m^{-2}$ and $\mu g cm^{-2}$ for LAI and LC_{ab} , and CCC, respectively. The RRMSE is a dimensionless index for comparing different variables or ranges, where values below 10% are considered Excellent, and 10% to 20% are considered Good (Richter et al. 2012). Lastly, a model's propensity to under- or overestimate a BV is measured by Bias, where 0 is the optimum Bias, and values near 0 Bias denote a correct model (Gara et al. 2019). These accuracy metrics were computed using R-statistics version 4.1.2, packages "Metrics" (<https://cran.r-project.org/web/packages/Metrics/index.html>, accessed 18 July 2021) and "Fgmutils" (<https://rdrr.io/cran/Fgmutils/>, accessed 18 July 2021).

3. Results

As shown by the results in Table 2, the number and location of spectral bands selected by the proposed STfs algorithms varied by biophysical and biochemical parameters, i.e. LAI, LC_{ab} , and CCC, and ranged from five to eight. For LAI, STfs-ReliefF had the best performance with R^2_{cv} of 66% and $RMSE_{cv}$ of $0.53 \text{ m}^2 \text{ m}^{-2}$ using only five variables (i.e. B3 [560 nm], B5 [705 nm], B6 [740 nm], B8 [842 nm], and B11 [1610 nm]). In contrast, the STfs-MLR and STfs-RF had an equivalent performance with R^2_{cv} of 63% and $RMSE_{cv}$ of $0.55 \text{ m}^2 \text{ m}^{-2}$. However, the former used relatively more spectral bands, i.e. seven (B3 [560 nm], B4 [665 nm], B5 [705 nm], B7 [783 nm], B8 [842 nm], B11 [1610 nm], and B12 [2190 nm]), while the later used only six spectral bands (i.e. B3 [560 nm], B4 [665 nm], B5 [705 nm], B6 [740 nm], B11 [1610 nm], and B12 [2190 nm]). On the other hand, STfs-MLR was optimal in the retrieval of both leaf-level and canopy-level chlorophyll content, i.e. LC_{ab} (R^2_{cv} : 74%, $RMSE_{cv}$: $7.09 \mu\text{g cm}^{-2}$) and CCC (R^2_{cv} : 77%, $RMSE_{cv}$: $33.69 \mu\text{g cm}^{-2}$), using only seven spectral bands, i.e. B2 (490 nm), B3 (560 nm), B4 (665 nm), B5 (705 nm), B6 (740 nm), B8A (865 nm), and B11 (1610 nm) and B3 (560 nm), B4 (665 nm), B5 (705 nm), B6 (740 nm), B7 (783 nm), B8A (865 nm), and B11 (1610 nm). The spectral bands selected by RF-RFE are B3 (560 nm), B5 (705 nm), B6 (740 nm), B8A (865 nm), and B11 (1610 nm) for LAI, B4 (665 nm), B6 (740 nm), B7 (783 nm), B11 (1610 nm), B12 (2190 nm) for LCC and B3 (560 nm), B5 (705 nm), B6 (740 nm), B8A (865 nm), and B12 (2190 nm) for CCC.

Figure 3 shows the Similarities and differences in selected features between STfs and RF-RFE. None of the evaluated feature selection algorithms selected B2 (490 nm) for LAI, while STfs-MLR and STfs-ReliefF selected it for LC_{ab} and only STfs-ReliefF for CCC. Sentinel-2 B3 (560 nm), B5 (705 nm), and B11 (1610 nm) for LAI, B4 (665 nm) for LC_{ab} , and B5 (705 nm) and B6 (740 nm) for CCC were mutually selected by all algorithms. Generally, the selected bands by STfs were consistent with those selected by a well-established algorithm, i.e. RF-RFE.

When benchmarked against the entire feature space (i.e. no feature selection), results show that the retrieval accuracy of crop biophysical and biochemical parameters using the optimized MSI feature space by STfs algorithms was consistently slightly higher R^2_{cv} by 3%, 1%, and 2% and $RMSE_{cv}$ by $0.03 \text{ m}^2 \text{ m}^{-2}$, $0.12 \mu\text{g cm}^{-2}$, $1.61 \mu\text{g cm}^{-2}$, for LAI, LC_{ab} , and CCC, respectively. Similarly, the comparison of the optimized MSI feature space by

Table 2. The performance of optimal variables selected by spectral triad feature selection (STfs), Random Forest-Recursive Feature Elimination (RF-RFE), and comparison to entire feature space (i.e. no feature selection).

	Feature selection approaches	n Selected features	R^2_{cv}	$RMSE_{cv}$	MAE_{cv}	$RRMSE_{cv}$	$BIAS_{cv}$	Training time (s)
LAI	STfs-MLR	7	0.63	0.55	0.37	0.01	0	26.39
	STfs-ReliefF	5	0.66	0.53	0.36	0.01	0	24.25
	STfs-RF	6	0.63	0.55	0.37	0.01	0	25.67
	RF-RFE	5	0.66	0.53	0.37	0.01	0	23.33
	No feature selection	10	0.64	0.55	0.37	0.02	−0.15	29.11
LC_{ab}	STfs-MLR	7	0.74	7.09	5.45	0.02	−0.1	25.44
	STfs-ReliefF	6	0.73	7.23	5.63	0.02	−0.15	24.09
	STfs-RF	8	0.74	7.13	5.54	0.02	−0.12	26.43
	RF-RFE	5	0.73	7.14	5.64	0.02	−0.09	22.63
	No feature selection	10	0.73	7.13	5.55	0.02	−0.15	27.34
CCC	STfs-MLR	7	0.77	33.69	24.01	0.02	0.16	27.28
	STfs-ReliefF	7	0.76	34.45	24.26	0.03	−0.13	26.80
	STfs-RF	7	0.74	35.13	25.50	0.03	0.45	25.89
	RF-RFE	5	0.78	32.82	23.24	0.02	−0.03	23.83
	No feature selection	10	0.75	34.76	24.71	0.03	0.35	29.58

The best-performing feature selection algorithms are highlighted in bold.

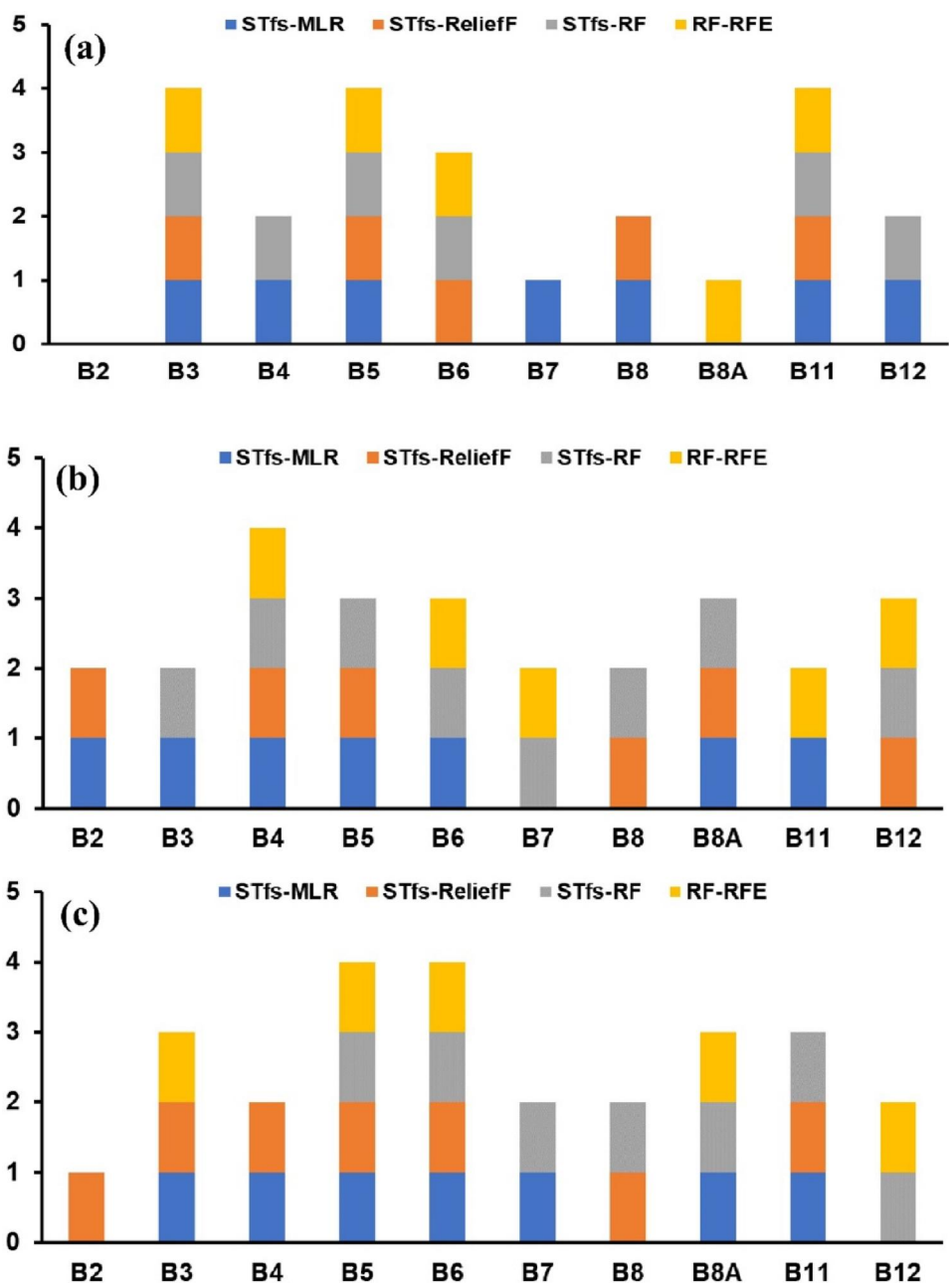


Figure 3. Similarities and differences in selected features between the proposed spectral triad feature selection (STfs) and Random Forest-Recursive Feature Elimination (RF-RFE) algorithms. Selected spectral bands for estimating (a) LAI, (b) LCab, and (c) CCC. The x-axis indicates Sentinel-2 spectral bands used in the analysis.

STfs algorithms to a well-established wrapper-based feature selection algorithm, i.e. RF-RFE, also shows minute differences in accuracies with R^2_{cv} differences of only 1% across all considered crop biophysical and biochemical parameters. In addition, depending on the crop parameters, the $RMSE_{cv}$ differences were marginally more favourable to STfs or

RF-RFE. All models achieved $RRMSE_{cv}$ less than 3%. Table 2 shows that feature subset selection results in better accuracies.

As shown in Figure 4, the distribution of predicted vs. observed BVs show no apparent differences between the entire feature space (i.e. No feature selection), RF-RFE, and STfs, except for LAI, where some values around $2 \text{ m}^2 \text{ m}^{-2}$ were overestimated as values between 3 and $4 \text{ m}^2 \text{ m}^{-2}$ by all subset and the entire feature space (Figure 4), resulting in $BIAS_{cv}$ of between 0.002 to $0.005 \text{ m}^2 \text{ m}^{-2}$ for STfs and RF-RFE. These overestimations are also visible in some areas (see the red box) on the spatial distribution maps (Figure 5(g)). Moreover, LAI values between $4 \text{ m}^2 \text{ m}^{-2}$ and $5.5 \text{ m}^2 \text{ m}^{-2}$ were underestimated by all datasets but severe in No feature selection where a $BIAS_{cv}$ of $0.15 \text{ m}^2 \text{ m}^{-2}$ was obtained. The optimal subsets by STfs and RF-RFE achieved better characterization of low LC_{ab} values (shown in cyan to blue in Figure 5(b,e)) than in the entire feature space.

4. Discussion

The advent of quasi-hyperspectral sensors such as Sentinel-2, characterized by many broad- and narrow-band wavelengths, present challenges of collinearity to machine learning algorithms, resulting in suboptimal performance. Therefore, this calls for optimizing the features for specific problems and applications through feature selection to ensure that the optimal subset (i.e. fewer bands that can achieve equivalent or better accuracy to the entire dataset) is used for prediction. In this study, the Spectral Triad feature selection

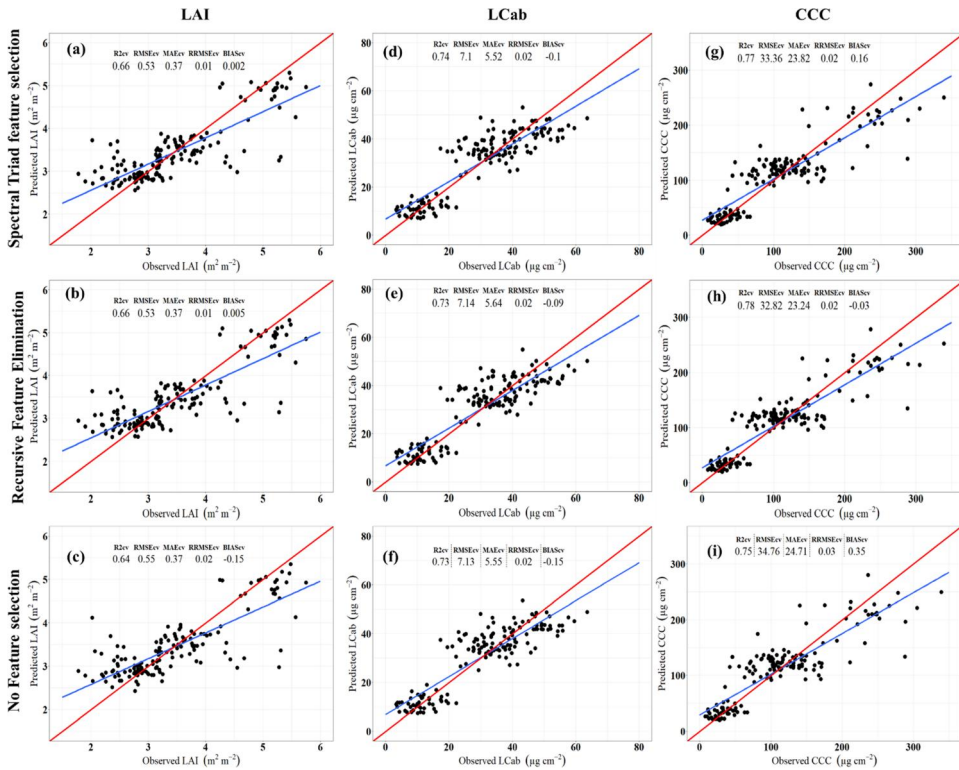


Figure 4. Comparison of the scatterplots of crop biophysical and biochemical parameters derived from the optimized Sentinel-2 MSI subsets selected by various feature selection techniques, i.e. spectral triad feature selection (a, d, g), Random Forest – Recursive Feature Elimination (b, c, h), and no feature selection (c, f, i).

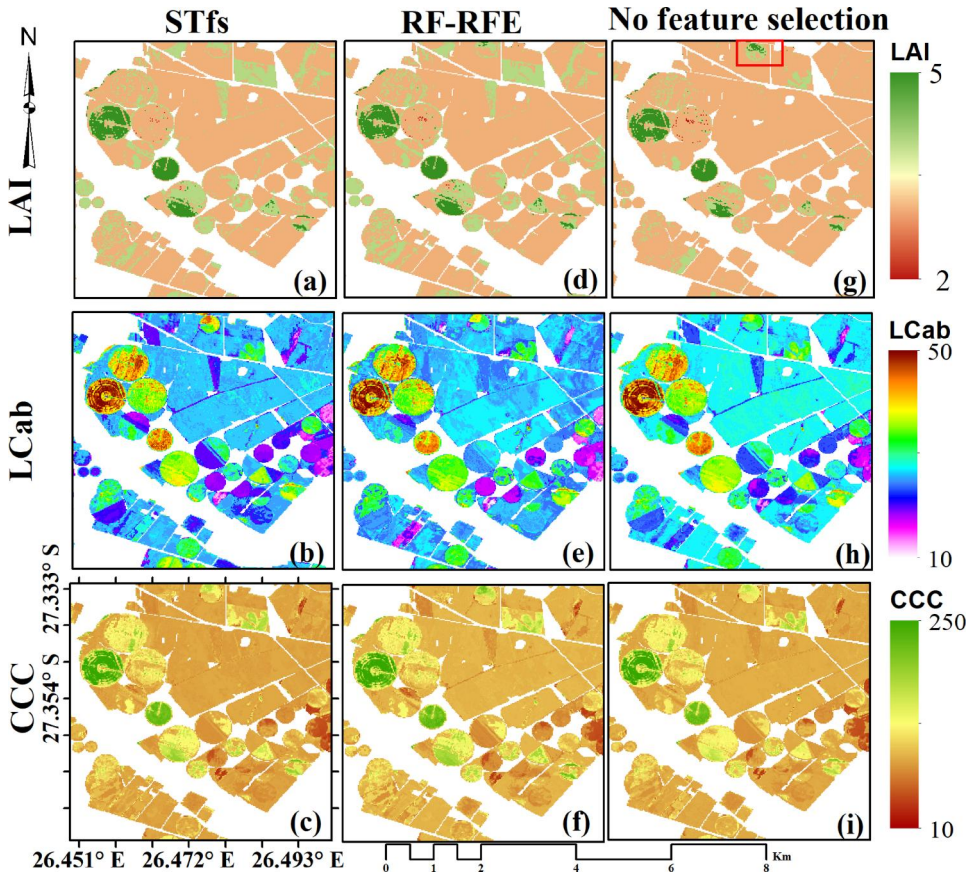


Figure 5. Maps of crop biophysical and biochemical parameters retrieved using the optimized Sentinel-2 MSI subsets selected by the novel spectral triad feature selection (a, b, and c), Random Forest-Recursive feature Elimination (d, e, and f), and entire feature space (i.e. no feature selection, g, h, and i).

(STfs) algorithm – based on music theory – is proposed, with two variants, i.e. filter-based STfs and wrapper-based STfs. Specifically, in filter-based mode, STfs was used with ReliefF algorithm, while in the wrapper-based mode, it was used with multiple linear regression and Random Forest (RF) algorithms (see Table 2 and Figure 3). The study found that feature selection algorithms selected fewer spectral bands, i.e. five to eight, varying by BV. This finding is consistent with studies by Verrelst et al. (2016) and Verrelst et al. (2012) that found that 4 to 9 spectral bands (out of hundreds from hyper-spectral sensors) were optimal for retrieving crop parameters. Moreover, LAI and CCC were characterized by similar bands selected by STfs-Relief (except for B2 selected for CCC but not LAI) and RF-RFE (except for B12 selected for CCC but not LAI) (see Figure 3), which indicates that similar plant traits influence the two crop parameters. The similarity of selected bands across different crop parameters may be due to the known co-variation of plant traits within various spectral regions covered by MSI (Ollinger 2011). This finding indicates that the novel STfs can select highly informative variables and is comparable to a well-established feature subset selection algorithm, i.e. RF-RFE. Therefore, STfs is promising for optimizing the feature space of quasi-hyperspectral data to reduce the complexity of the models and improve processing speeds. The fact that the spectral bands chosen by STfs and RF-RFE were dispersed across the whole MSI spectral

coverage in [Figure 3](#) is among the most important findings of this study. It highlights the significance of the MSI for various land applications, including supporting precision agriculture applications. Besides, studies reveal that the various plant traits impact the electromagnetic spectrum across the board, not just in certain places. For instance, although leaf pigments like chlorophyll are known to absorb heavily in the blue (400–500 nm) and red-edge (650–700 nm) regions, LAI affects the entire spectrum ([Ollinger 2011](#); [Verrelst et al. 2016](#)).

Certainly, feature subset selection is beneficial, resulting in better prediction accuracies (See [Table 2](#)). This observation is consistent with [Mutanga et al. \(2015\)](#), who found that three normalized ratios achieved better RMSE of calibration, i.e. 0.175 vs. 0.187 achieved by 28 normalised difference indices (NDIs), and five ratio indices (RIs) achieved 0.176 vs 0.188 achieved by the entire dataset of 56 RIs. In estimating crude protein and metabolizable energy using airborne hyperspectral imaging, [Pullanagari et al. \(2018\)](#) found that RF-RFE outperformed the full spectrum, achieving R^2_{cv} of 80% vs. 66% and 78% vs. 61%, respectively. Moreover, [Kganyago et al. \(2021\)](#) found that the sparse Partial Least Squares (i.e. embedded feature selection algorithm) yielded better performance (i.e. RMSE: $7.90 \mu\text{g cm}^{-2}$) over Gradient Boosting Machines—which achieved RMSE of $8.25 \mu\text{g cm}^{-2}$ with no feature selection—in retrieving LC_{ab} by using only seven variables. The optimized subsets in the current study performed better compared to the Sentinel-2 subsets used by [Delloye et al. \(2018\)](#) in evaluating the performance of the ANN algorithm for BV retrieval, i.e. 10 m bands ($n=3$, [B3, B4, B5]), SPOT5 bands ($n=4$, [B3, B4, B8 and B11]), Red-edge ($n=7$, [same as SPOT5 and B5, B6 and B7]), and All bands (all Sentinel-2 bands, excluding B2). In LAI retrieval, their results had RMSE values ranging from $0.55 \text{ m}^2 \text{ m}^{-2}$ to $1 \text{ m}^2 \text{ m}^{-2}$, while LC_{ab} RMSE values ranged from 11.03 to $13.94 \mu\text{g cm}^{-2}$ and for CCC, they achieved RMSE values from 0.35 to 0.51 g m^{-2} . Comparatively, the results obtained from STfs subsets are higher by $0.02 \text{ m}^2 \text{ m}^{-2}$ and $3.93 \mu\text{g cm}^{-2}$, for LAI and LC_{ab} , respectively. Another study utilizing all Sentinel-2 bands, Partial Least Squares Regression and RF algorithms found RMSE of $0.68 \text{ m}^2 \text{ m}^{-2}$ for LAI and $8.88 \mu\text{g cm}^{-2}$ for LC_{ab} , resulting in differences of $0.15 \text{ m}^2 \text{ m}^{-2}$ and $1.78 \mu\text{g cm}^{-2}$, respectively, with STfs optimized subsets.

In the current study, RF-RFE was consistently faster due to a relatively smaller subset, i.e. five spectral bands. STfs resulted in better training times when compared to the entire feature space (i.e. No feature selection). The relatively small feature subsets obtained by RF-RFE and STfs are critical for reducing model complexity and the chances of overfitting from redundant features ([R. R. Pullanagari et al. 2016](#)). The optimized feature spaces by STfs and RF-RFE in [Figure 4a](#) and [b](#), respectively, could reduce these overestimations to below $4 \text{ m}^2 \text{ m}^{-2}$, thus demonstrating the robustness of retrievals derived from optimal feature subsets. As it learns the significance of the features from the data, the STfs algorithm has the potential to be applied to any dataset, including hyperspectral data and datasets containing diverse variables such as a combination of spectral bands, vegetation indices, and textural measures.

While STfs algorithms selected various numbers of spectral bands per crop biophysical and biochemical parameters, RF-RFE consistently selected five spectral bands. This finding can be attributed to the search criterion used by the proposed STfs algorithm vs. the RFE algorithm. STfs uses spectral triads, i.e. combinations of three spectral bands, at each iteration and uses the evaluator (i.e. filter-band or machine learning algorithm) to estimate the relationship with the response variable, ranks, and selects a single most contributing spectral band in a triad. In contrast, RFE sequentially removes the least contributing bands in the regression models until there is no possible elimination without causing a loss in accuracy. The remaining spectral bands are considered the optimal subset. In STfs,

spectral bands have many chances of being selected because the relatively lowly contributing spectral bands in each triad may be a high contributor in another triad, while eliminated spectral bands in RFE are not considered again in evaluating the subset. Therefore, STfs is likely to retain more spectral bands than RFE, as observed in the current study. In highly dimensional data such as hyperspectral data, it is anticipated that the number of evaluated spectral triads and selected variables in a triad may still retain high dimensionality and cause a loss in computational efficiency. In such cases, it is recommended to specify the desired number of variables or subset size, in which case, the algorithm will create and evaluate new triads until the specified subset size has been reached. Future studies should test the algorithm with hyperspectral data and many datasets to evaluate the performance of STfs comprehensively. Other evaluation criteria could be incorporated, as well as other chord structures.

The STfs algorithm offers several advantages: (1) it provides an objective criterion for choosing the relevant variables, compared to arbitrarily chosen thresholds in filter-based approaches; (2) any evaluation criteria can be used to rank informative variables in each triad, thus allowing the user the flexibility and freedom of choice. For example, the users are not only limited to filter approaches but can employ more robust and explainable machine learning algorithms; (3) only relevant (i.e. most informative) variables, i.e. in relation to one another and the response variable, are selected to form the final subset for prediction, thus simultaneously dealing with problems of collinearity and dimensionality.

5. Conclusion

This study proposed the novel Spectral Triad feature selection (STfs) for optimizing the spectral feature spaces of quasi-hyperspectral sensors such as Sentinel-2, which are characterized by many typical broad-band and strategically located narrow-band wavebands. The performance of the proposed feature selection technique was benchmarked against the well-established Recursive Feature Elimination (RFE) coupled with a machine learning algorithm, i.e. Random Forest (RF), and the entire MSI spectral feature space, in retrieving essential crop biophysical and biochemical parameters. Overall, the results demonstrated that the MSI feature space could be optimized through feature selection, which results in slightly better or equivalent retrieval accuracies to the entire MSI feature space. Moreover, the spectral bands selected by STfs were consistent with those selected by a well-established algorithm, i.e. RF-RFE. This study showed that the proposed STfs algorithm has great potential to optimize the spectral feature space of quasi-hyperspectral sensors for rapid crop biophysical and biochemical parameter retrieval.

Acknowledgement

The authors would like to acknowledge the European Space Agency (ESA) and Copernicus Programme for providing Sentinel-2 data free of charge. The data was accessed through Sentinel Hub Cloud API for Satellite Imagery provided by ESA Network of Resources (NoR) sponsorship. We also appreciate the EU-AfriCultuReS Project (GA: 774652) and the South African National Space Agency (SANSA) for providing field data for this study. Mahlatse Kganyago received University Research Committee (URC) Research Grant (Grant No. 2023URC00563) from University of Johannesburg Faculty of Science.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Mahlatsé Kganyago  <http://orcid.org/0000-0001-9553-0378>

References

- Breiman L. 2001. Random forests. *Mach Learn.* 45, 5–32. doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- Breiman L, Cutler A, Liaw A, Wiener M. 2018. Package ‘RandomForest’ - Breiman and Cutler’s Random Forests for Classification and Regression. CRAN Repository.
- Daughtry CST, Walthall CL, Kim MS, Brown De Colstoun E, McMurtry JE. 2000. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Rem Sens Environ.* 74(2):229–239. doi:[10.1016/S0034-4257\(00\)00113-9](https://doi.org/10.1016/S0034-4257(00)00113-9).
- Delegido J, Verrelst J, Meza CM, Rivera JP, Alonso L, Moreno J. 2013. A red-edge spectral index for remote sensing estimation of green LAI over agroecosystems. *Europ J Agron.* 46:42–52. doi:[10.1016/j.eja.2012.12.001](https://doi.org/10.1016/j.eja.2012.12.001).
- Delloye C, Weiss M, Defourny P. 2018. Retrieval of the canopy chlorophyll content from sentinel-2 spectral bands to estimate nitrogen uptake in intensive winter wheat cropping systems. *Rem Sens Environ.* 216(July):245–261. Elsevier doi:[10.1016/j.rse.2018.06.037](https://doi.org/10.1016/j.rse.2018.06.037).
- Demarchi L, Kania A, Cieżkowski W, Piórkowski H, Oświecimska-Piasko Z, Chormański J. 2020. Recursive feature elimination and random forest classification of natura 2000 grasslands in lowland river valleys of poland based on airborne hyperspectral and LiDAR data fusion. *Rem Sens.* 12(11):1842. doi:[10.3390/rs12111842](https://doi.org/10.3390/rs12111842).
- Dimitrov P, Kamenova I, Roumenina E, Filchev L, Ilieva I, Jelev G, Gikov A. 2019. Estimation of biophysical and biochemical variables of winter wheat through sentinel-2 vegetation indices. *Bulg J Agric Sci.* 25(5):819–832.
- Fawagreh K, Gaber MM, Elyan E. 2014. Random forests: from early developments to recent advancements. *Syst Sci Control Eng.* 2(1):602–609. doi:[10.1080/21642583.2014.956265](https://doi.org/10.1080/21642583.2014.956265).
- Fitzgerald G, Rodriguez D, O’Leary G. 2010. Measuring and predicting canopy nitrogen nutrition in wheat using a spectral index-the canopy Chlorophyll Content Index (CCCI). *Field Crops Res.* 116(3): 318–324. doi:[10.1016/j.fcr.2010.01.010](https://doi.org/10.1016/j.fcr.2010.01.010).
- Frampton WJ, Dash J, Watmough G, Milton EJ. 2013. Evaluating the capabilities of sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS J Photogramm Remote Sens.* 82:83–92. doi:[10.1016/j.isprsjprs.2013.04.007](https://doi.org/10.1016/j.isprsjprs.2013.04.007).
- Gara TW, Skidmore AK, Darvishzadeh R, Wang T. 2019. Leaf to canopy upscaling approach affects the estimation of canopy traits. *GISci Rem Sens.* 56(4):554–575. doi:[10.1080/15481603.2018.1540170](https://doi.org/10.1080/15481603.2018.1540170).
- Georganos S, Grippa T, Vanhuysse S, Lennert M, Shimoni M, Wolff E. 2018. Very high resolution object-based land use-land cover urban classification using extreme gradient boosting. *IEEE Geosci Remote Sensing Lett.* 15(4):607–611. doi:[10.1109/LGRS.2018.2803259](https://doi.org/10.1109/LGRS.2018.2803259).
- Gislason PO, Benediktsson JA, Sveinsson JR. 2006. Random Forests for Land Cover Classification. In *Pattern Recog Lett.* 27(4):294–300. doi:[10.1016/j.patrec.2005.08.011](https://doi.org/10.1016/j.patrec.2005.08.011).
- Gitelson AA, Gritz Y, Merzlyak MN. 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J Plant Physiol.* 160(3):271–282. doi:[10.1078/0176-1617-00887](https://doi.org/10.1078/0176-1617-00887).
- Gitelson AA, Peng Y, Arkebauer TJ, Schepers J. 2014. Relationships between gross primary production, green LAI, and canopy chlorophyll content in maize: implications for remote sensing of primary production. *Rem Sens Environ.* 144:65–72. doi:[10.1016/j.rse.2014.01.004](https://doi.org/10.1016/j.rse.2014.01.004).
- Gregorutti B, Michel B, Saint-Pierre P. 2017. Correlation and variable importance in random forests. *Stat Comput.* 27(3):659–678. doi:[10.1007/s11222-016-9646-1](https://doi.org/10.1007/s11222-016-9646-1).
- Jacquemoud S, Baret F, Andrieu B, Danson FM, Jaggard K. 1995. Extraction of vegetation biophysical parameters by inversion of the PROSPECT+SAIL models on sugar beet canopy reflectance data. Application to TM and AVIRIS sensors. *Rem Sens Environ.* 52(3):163–172. doi:[10.1016/0034-4257\(95\)00018-V](https://doi.org/10.1016/0034-4257(95)00018-V).
- Jia F, Liu G, Liu D, Zhang Y, Fan W, Xing X. 2013. Wuguang Fan, and Xuexia Xing. Comparison of different methods for estimating nitrogen concentration in flue-cured tobacco leaves based on hyperspectral reflectance. *Field Crops Res.* 150:108–114. doi:[10.1016/j.fcr.2013.06.009](https://doi.org/10.1016/j.fcr.2013.06.009).
- Kganyago M. 2021. Using sentinel-2 observations to assess the consequences of the COVID-19 lockdown on winter cropping in bothaville and Harrismith, South Africa. *Remote Sens Lett.* 12(9):827–837. doi:[10.1080/2150704X.2021.1942582](https://doi.org/10.1080/2150704X.2021.1942582).

- Kganyago M, Adjorlolo C, Mhangara P. 2022. Exploring transferable techniques to retrieve crop biophysical and biochemical variables using sentinel-2 data. *Remote Sens.* 14(16):3968. doi:[10.3390/rs14163968](https://doi.org/10.3390/rs14163968).
- Kganyago M, Adjorlolo C, Sibanda M, Mhangara P, Laneve G, Alexandridis T. 2022. Testing sentinel-2 spectral configurations for estimating relevant crop biophysical and biochemical parameters for precision agriculture using tree-based and kernel-based algorithms. *Geocarto International*. 38(1):1–25. doi:[10.1080/10106049.2022.2146764](https://doi.org/10.1080/10106049.2022.2146764).
- Kganyago M, Mhangara P, Adjorlolo C. 2021. Estimating crop biophysical parameters using machine learning algorithms and sentinel-2 imagery. *Remote Sens.* 13(21):4314. doi:[10.3390/rs13214314](https://doi.org/10.3390/rs13214314).
- Kganyago M, Mhangara P, Alexandridis T, Laneve G, Ovakoglou G, Mashiyi N. 2020. Validation of sentinel-2 Leaf Area Index (LAI) product derived from SNAP toolbox and its comparison with global lai products in an african semi-arid agricultural landscape. *Rem Sens Lett.* 11(10):883–892. doi:[10.1080/2150704X.2020.1767823](https://doi.org/10.1080/2150704X.2020.1767823).
- Kganyago M, Odindi J, Adjorlolo C, Mhangara P. 2017. Selecting a subset of spectral bands for mapping invasive alien plants: A case of discriminating parthenium hysterophorus using field spectroscopy data. *Inter J Rem Sens.* 38(20):5608–5625. doi:[10.1080/01431161.2017.1343510](https://doi.org/10.1080/01431161.2017.1343510).
- Kononenko I, Robnik-Šikonja M, Pompe U. 1996. ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems. *Artificial Intelligence: methodology, Systems, Applications*. 1–15. <http://lkm.fri.uni-lj.si/rmarko/papers/kononenko96-aimsa.pdf>
- Li H, Chen Z-x, Jiang Z-w, Wu W-bin, Ren J-q, Liu B, Tuya H. 2017. Comparative analysis of GF-1, HJ-1, and Landsat-8 data for estimating the leaf area index of winter wheat. *J Integr Agric.* 16(2):266–285. doi:[10.1016/S2095-3119\(15\)61293-X](https://doi.org/10.1016/S2095-3119(15)61293-X).
- Mohanty S, Codell R. 2002. Sensitivity analysis methods for identifying influential parameters in a problem with a large number of random variables. *Management Information Systems*.
- Mutanga O, Adam E, Adjorlolo C, Abdel-Rahman EM. 2015. Evaluating the robustness of models developed from field spectral data in predicting african grass foliar nitrogen concentration using worldview-2 image as an independent test dataset. *Inter J Appl Earth Observ Geoinform.* 34(1):178–187. doi:[10.1016/j.jag.2014.08.008](https://doi.org/10.1016/j.jag.2014.08.008).
- Mutanga O, Adam E, Cho MA. 2012. High density biomass estimation for wetland vegetation using worldview-2 imagery and random forest regression algorithm. *Inter J Appl Earth Observ Geoinform.* 18(1):399–406. doi:[10.1016/j.jag.2012.03.012](https://doi.org/10.1016/j.jag.2012.03.012).
- Ollinger SV. 2011. Sources of variability in canopy reflectance and the convergent properties of plants. *New Phytol.* 189(2):375–394. doi:[10.1111/j.1469-8137.2010.03536.x](https://doi.org/10.1111/j.1469-8137.2010.03536.x).
- Pullanagari RR, Kereszturi G, Yule IJ. 2016. Mapping of macro and micro nutrients of mixed pastures using airborne AisaFENIX hyperspectral imagery. *ISPRS J Photogramm Remote Sens.* 117:1–10. doi:[10.1016/j.isprsjprs.2016.03.010](https://doi.org/10.1016/j.isprsjprs.2016.03.010).
- Pullanagari RR, Kereszturi G, Yule I. 2018. Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression. *Remote Sens.* 10(7):1117. MDPIdoi:[10.3390/rs10071117](https://doi.org/10.3390/rs10071117).
- Ramoelo A, Cho MA, Mathieu R, Madonsela S, van de Kerchove R, Kaszta Z, Wolff E. 2015. Monitoring grass nutrients and biomass as indicators of rangeland quality and quantity using random forest modelling and worldview-2 data. *Inter J Appl Earth Observ Geoinform.* 43:43–54. doi:[10.1016/j.jag.2014.12.010](https://doi.org/10.1016/j.jag.2014.12.010).
- Richter K, Atzberger C, Hank TB, Mauser W. 2012. Derivation of biophysical variables from earth observation data: validation and statistical measures. *J Appl Remote Sens.* 6(1):063557–1. doi:[10.1117/1.JRS.6.063557](https://doi.org/10.1117/1.JRS.6.063557).
- Robnik-Šikonja M, Kononenko I. 2003. Theoretical and empirical analysis of relief and rrelief. *Mach Learn.* 53(1/2):23–69. <http://lkm.fri.uni-lj.si/xaigor/slo/clanki/MLJ2003-FinalPaper.pdf>. doi:[10.1023/A:1025667309714](https://doi.org/10.1023/A:1025667309714).
- Romanski P. 2021. Package ‘FSelector.’ Package ‘FSelector’ 0.31.
- Shah SH, Angel Y, Houborg R, Ali S, McCabe MF. 2019. A random forest machine learning approach for the retrieval of leaf chlorophyll content in wheat. *Rem Sens.* 11(8):920. doi:[10.3390/rs11080920](https://doi.org/10.3390/rs11080920).
- Sibanda M, Mutanga O, Rouget M, Kumar L. 2017. Estimating biomass of native grass grown under complex management treatments using worldview-3 spectral derivatives. *Rem Sens.* 9(1):55. doi:[10.3390/rs9010055](https://doi.org/10.3390/rs9010055).
- Verrelst J, Alonso L, Camps-Valls G, Delegido J, Moreno J. 2012. Retrieval of vegetation biophysical parameters using gaussian process techniques. *IEEE Trans Geosci Rem Sens.* 50(5):1832–1843. doi:[10.1109/TGRS.2011.2168962](https://doi.org/10.1109/TGRS.2011.2168962).
- Verrelst J, Muñoz J, Alonso L, Delegido J, Rivera JP, Camps-Valls G, Moreno J. 2012. Machine learning regression algorithms for biophysical parameter retrieval: opportunities for sentinel-2 and -3. *Rem Sens Environ.* 118:127–139. doi:[10.1016/j.rse.2011.11.002](https://doi.org/10.1016/j.rse.2011.11.002).

- Verrelst J, Rivera JP, Alonso L, Guanter L, Moreno J. 2012. Evaluating machine learning regression algorithms for operational retrieval of biophysical parameters: opportunities for sentinel. In European Space Agency. (Special Publication) ESA SP707SP.
- Verrelst J, Rivera JP, Gitelson A, Delegido J, Moreno J, Camps-Valls G. 2016. Spectral band selection for vegetation properties retrieval using gaussian processes regression. *Inter J Appl Earth Observ Geoinform*. 52:554–567. doi:[10.1016/j.jag.2016.07.016](https://doi.org/10.1016/j.jag.2016.07.016).
- Verrelst J, Rivera JP, Veroustraete F, Muñoz-Marí J, Clevers JG, Camps-Valls G, Moreno J. 2015. Experimental sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods - a comparison. *ISPRS J Photogramm Remote Sens*. 108:260–272. doi:[10.1016/j.isprsjprs.2015.04.013](https://doi.org/10.1016/j.isprsjprs.2015.04.013).
- Vincini M, Calegari F, Casa R. 2016. Sensitivity of leaf chlorophyll empirical estimators obtained at sentinel-2 spectral resolution for different canopy structures. *Precision Agric*. 17(3):313–331. doi:[10.1007/s11119-015-9424-7](https://doi.org/10.1007/s11119-015-9424-7).
- Xu C, Ding Y, Zheng X, Wang Y, Zhang R, Zhang H, Dai Z, Xie Q. 2022. A comprehensive comparison of machine learning and feature selection methods for maize biomass estimation using sentinel-1 SAR, sentinel-2 vegetation indices, and biophysical variables. *Remote Sens*. 14(16):4083. doi:[10.3390/rs14164083](https://doi.org/10.3390/rs14164083).
- Ye F. 2018. Evolving the SVM model based on a hybrid method using swarm optimization techniques in combination with a genetic algorithm for medical diagnosis. *Multimed Tools Appl*. 77(3):3889–3918. doi:[10.1007/s11042-016-4233-1](https://doi.org/10.1007/s11042-016-4233-1).
- Yu L, Liu H. 2004. Efficient feature selection via analysis of relevance and redundancy. *J Mach Learn Res*. 5:1205–1224. doi:[10.5555/1005332.1044700](https://doi.org/10.5555/1005332.1044700).
- Zawadzki Z, Kosinski M. 2020. FSelectorRcpp: 'Rcpp' Implementation of 'FSelector' Entropy-Based Feature Selection Algorithms with a Sparse Matrix Support. R Package Version 0.3. 3.