1 Automated hyperspectral vegetation index derivation using a hyperparameter optimization framework for 2 high-throughput plant phenotyping

3

4 Joshua C.O. Koh^{1*}, Bikram P. Banerjee^{1*}, German Spangenberg^{2,3} and Surya Kant^{1,2,3}

5

6 1 Agriculture Victoria, Grains Innovation Park, 110 Natimuk Rd, Horsham, Victoria 3400, Australia

7 2 Agriculture Victoria, AgriBio, Centre for AgriBioscience, 5 Ring Road, Bundoora, Victoria 3083,

8 Australia

9 3 School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3083, Australia

10

11 *These authors contributed equally to this work.

12

13 Author for correspondence:

14 Surya Kant

15 Tel: +61(3) 4344 3179

16 Email: surya.kant@agriculture.vic.gov.au

17

18 Summary

- 19 • Hyperspectral vegetation indices (VIs) are widely deployed in agriculture remote sensing and
20 plant phenotyping to estimate plant biophysical and biochemical traits. However, existing VIs
21 consist mainly of simple 2-band indices which limits the net performance and often do not
22 generalize well for other traits than they were originally designed for.
- 23 • We present an automated hyperspectral vegetation index (AutoVI) system for the rapid
24 generation of novel 2- to 6-band trait-specific indices in a streamlined process covering model
25 selection, optimization and evaluation driven by the tree parzen estimator algorithm. Its
26 performance was tested in generating novel indices to estimate chlorophyll and sugar contents
27 in wheat.
- 28 • Results show that AutoVI can rapidly generate complex novel VIs (≥ 4 -band index) which
29 correlated strongly ($R^2 > 0.8$) with measured chlorophyll and sugar contents in wheat. AutoVI-
30 derived indices were used as features in simple and stepwise multiple linear regression for
31 chlorophyll and sugar content estimation, and outperformed results achieved with existing 47
32 VIs and those provided by partial least squares regression.

- The AutoVI system can deliver novel trait-specific VIs readily adoptable in high-throughput plant phenotyping platforms and should appeal to plant scientists and breeders. A graphical user interface of AutoVI is herein provided.

Key words: hyperspectral vegetation indices, automated vegetation index development, high-throughput plant phenotyping, hyperparameter optimization, chlorophyll estimation, sugar estimation.

Introduction

High-throughput plant phenotyping (HTP) is integral in meeting the demand for large-scale evaluation of genotypes in breeding programs and crop management systems (Tardieu *et al.*, 2017; Mir *et al.*, 2019). In recent years, controlled environment and field-based HTP platforms have been developed to monitor plants at the canopy or plot level for a large number of crop genotypes (Tardieu *et al.*, 2017; Mir *et al.*, 2019; Lu *et al.*, 2020). Central to the success of these HTP platforms is the use of various imaging sensors to acquire morphological, physiological, and biochemical parameters in a non-invasive manner. Hyperspectral imaging has been a promising HTP technology in measuring biochemical and morpho-physiological traits in a fast and non-destructive way by detecting signatures in the reflectance spectrum of vegetation in narrow (e.g. 1 - 2 nm in spectral resolution) and contiguous/broad (e.g. ≥ 20 nm in spectral resolution) spectral bands (Lu *et al.*, 2020). The recent availability of lightweight hyperspectral sensors has stimulated a rapid adoption of these sensors for use in unmanned aerial vehicle systems for HTP and precision agriculture (Adão *et al.*, 2017; Lu *et al.*, 2020). Hyperspectral data have been applied for the estimation of biophysical (e.g. leaf area index, green canopy cover) and biochemical (e.g. chlorophyll, nitrogen, sugar) traits for the leaf and canopy, as well as traits relating to the whole plant, particularly those linked with growth (e.g. plant biomass, plant height) and productivity (e.g. grain yield) (Adão *et al.*, 2017; Lu *et al.*, 2020). Apart from this, close-range hyperspectral imaging (ground-based or glasshouse) characterized by high spatial resolution and signal-to-noise ratio is also increasingly common in HTP facilities. These systems allow fine-scale investigation of vegetative features at the leaf or canopy level with applications in plant water content and biochemical compounds estimation, and detection of abiotic and biotic stresses in plants (Mishra *et al.*, 2017).

However, the large amount of data collected by hyperspectral sensors poses challenges in analytical implementations. Redundancy problems linked to the multicollinearity of bands and the curse of dimensionality impose high computational costs on analytical pipelines (Bajwa & Kulkarni, 2011;

67 Burger & Gowen, 2011). Multicollinearity happens when independent variables (e.g. wavebands) are
68 highly correlated, leading to inaccurate estimation of coefficients (effect of independent variables on
69 response/target trait) in a regression analysis and less reliable statistical inferences (Bajwa &
70 Kulkarni, 2011; Burger & Gowen, 2011). On the other hand, curse of dimensionality refers to the
71 problem of optimizing a given function or model due to the exponential increase in possible
72 solutions or sets of parameters associated with the increase in the number of variables (i.e. the
73 dimensionality), which often necessitates exhaustive enumeration of the solution/search space to
74 achieve satisfactory optimization (Piatetsky-Shapiro *et al.*, 2000). Dedicated efforts are often
75 required to develop efficient hyperspectral data processing workflows for a specific plant
76 phenotyping task (Aasen *et al.*, 2014; Aasen *et al.*, 2018). To this end, vegetation indices (VIs) offer
77 an alternate approach to hyperspectral data analysis which is simple and fast. The VIs are formulated
78 as ratios or algebraic combinations of vegetative reflectance at different wavebands, selected from
79 the visible (VIS, 400 – 700 nm), near infrared (NIR, 700 – 1000 nm) and the shortwave infrared (SWIR,
80 1000 – 2500 nm) regions (Silleos *et al.*, 2006; Xue & Su, 2017). The application of VIs for plant
81 phenotyping is straightforward, requiring the user to compute index values directly from the
82 relevant wavebands' reflectance and use them either as proxy measures to the target trait or as
83 features (or variables) in regression modelling to predict the target trait values. Since the
84 introduction of the first VI by Pearson and Miller (1972), more than 500 hyperspectral VIs have been
85 developed, demonstrating a strong and continued interest in the development and adoption of
86 novel VIs for specific remote sensing applications (Henrich *et al.*, 2017). However, existing VIs consist
87 predominantly of 2-band indices and to some extent, 3-band indices, which limits the amount of
88 information represented and thereby the net performance afforded by these VIs (Henrich *et al.*,
89 2017). In addition, existing VIs are designed for specific plant traits and often do not generalize well
90 for other traits. As such, particularly in agriculture, there is a continued demand for novel VIs that
91 can target specific traits associated with crop growth, biochemical parameters, yield, and quality.

92

93 The development of VIs is technically challenging and time-consuming, often requiring a
94 comprehensive understanding of the dynamic changes of the plant optical properties in relation to
95 the intrinsic biochemical or biophysical trait(s) of interest. To this end, a wide variety of experiments
96 have been established to acquire a comprehensive spectral library (Rao *et al.*, 2007; Chauhan &
97 Mohan, 2013). Ideally, knowledge on wavebands associated with plant traits may be enriched or
98 expanded with successive development of new VIs when different regions of the reflectance
99 spectrum corresponding to vegetative features are identified. Biochemical and biophysical traits can
100 be described comprehensively with more wavebands where each waveband adds supplementary

information. However, this is seldom the case as the VIs rarely constitute a complex or cohesive assemblage of wavebands but rather a limited selection of a few wavebands (usually ≤ 4 bands). In addition, multicollinearity of bands and the curse of high dimensionality inherent in hyperspectral data complicates the identification of wavebands linked to the underlying trait of interest. Several attempts have been made to accelerate the development of novel VIs, these include the use of correlation matrices between VIs and the target traits of interest to retrieve new waveband or index combinations (Thenkabail *et al.*, 2004; Aasen *et al.*, 2014; Xu *et al.*, 2019). Careful selection of hyperspectral features (wavebands) was effective in overcoming the curse of dimensionality and the selected bands could be combined to develop VIs (Aasen *et al.*, 2014; Aasen *et al.*, 2018). Recently, a brute force indices mining approach was applied in our lab to identify a new normalized difference chlorophyll index (NDCI_w) for the estimation of chlorophyll content in wheat (Banerjee *et al.*, 2020). However, these approaches are suited for the evaluation of a limited number of wavebands and/or index model combinations for the development of new VIs; and may be computationally less efficient when dealing with a greater number of wavebands and index models. An efficient VIs evaluation and waveband selection strategy is crucial to developing trait-specific hyperspectral VIs for HTP and agriculture remote sensing.

This study aims to report the development and evaluation of an automated hyperspectral VI (AutoVI) derivation system for the rapid generation of trait-specific novel 2-band to 6-band VIs based on a hyperparameter optimization framework. The term “hyperparameter optimization” (HPO) is often associated with the machine learning discipline in which specific algorithms are deployed to select optimum values in a defined search space for model parameters (values learned from data) and hyperparameters (values associated with the model function or architecture) to maximize the model performance (Bergstra *et al.*, 2011; Yu & Zhu, 2020). Building upon an HPO framework, AutoVI is designed to deliver an end-to-end VI development pipeline covering index model evaluation and optimum waveband selection with minimal user input. In this study, novel VIs for chlorophyll and sugar estimation in wheat were generated using AutoVI and compared against existing VIs as features in simple and multiple regression modelling, with the results also compared against those computed by partial least squares regression. The potential application of AutoVI for HTP and agriculture remote sensing is discussed.

Materials and Methods

Experiment in a high-throughput phenotyping facility

Data used in this study was collected in a high-throughput controlled-environment phenotyping facility in the Plant Phenomics Victoria, Horsham (PPVH), Agriculture Victoria as previously described (Banerjee *et al.*, 2020). In brief, the PPVH facility is equipped with a conveyor belt system, automated weighing and watering stations, and an automated phenotyping Scanalyzer 3D system (LemnatecGmbH, Aachen, Germany), which includes a hyperspectral imaging sensor. For the experiment conducted at the PPVH, wheat plants were grown at 2, 5, 10, and 20 mM nitrogen (N) levels. One plant per pot was grown in a nutrient-free growth medium, consisting of perlite covered with a layer of vermiculite. Individual pots were weighed and equalized to a fixed pot weight and watered uniformly. The pots were loaded onto the system 10 days after the emergence of seedlings. Nutrient solution (4 mM MgSO₄, 4 mM KCl, 5 mM CaCl₂, 3 mM KH₂PO₄/K₂HPO₄-pH 6.0, 0.1 mM Fe-EDTA, 10 µM MnCl₂, 10 µM ZnSO₄, 2 µM CuSO₄, 50 µM H₂BO₃, and 0.2 µM Na₂MoO₄) with the indicated N concentration was supplied as 100 ml per pot every week. The growing conditions were 16 h (24 °C) day / 8 h (15 °C) night. The experiment was conducted as biological repeats with 20 replicate plants per N treatment. A subset of five plants per N treatment were destructively harvested at 14, 21, 28 and 35 days after sowing (DAS) and a total of 80 samples were collected for biochemical assays.

Hyperspectral image acquisition and processing

Plants were scanned in an imaging station with a push broom hyperspectral sensor (Micro-Hyperspec, VNIR-E Series, Headwall Photonics, Fitchburg, MA, USA) over a spectral range of 475 – 1710 nm in three viewing angles (0°, 120°, and 240°). Raw data acquired in 12-bit digital numbers (DNs) were transformed to radiance following spectral and radiometric calibration using a white reference target (Spectralon panel, Labsphere Inc, North Sutton, US) and dark reference (spectrum collected with halogen lamps turned off) in a data-acquisition software (Hyperspec III, Headwall Photonics, Inc, Massachusetts, US). Further processing was applied to remove inter-channel variation and correct illumination variations (Banerjee *et al.*, 2020). Non-plant pixels (cage, pot, soil, and background) in the hyperspectral image were first classified using the spectral information divergence method (Chein, 1999) and a binary mask was applied to segment out the remaining pixels (i.e. the plant pixels). Detected plant pixels were averaged to generate a representative reflectance spectrum with 256 spectral bands and resampled to a spectral width of 1 nm using a linear resampling approach (Lewitt, 1990). A total of 47 published VIs within the spectral range of 475 – 1710 nm was then computed (Table S1).

Biochemical assays

Whole plant samples were finely ground using a pestle and mortar with liquid nitrogen, then subsampled separately for chlorophyll and sugar analysis and stored at -80°C until biochemical analysis. For chlorophyll analysis, chlorophyll was extracted from 100 mg of sample with 100% methanol and centrifuged for 10 min at 10,016 g; this process was repeated twice. Extracts were analyzed by recording the absorbance at 750, 665, 652, and 470 nm using a UV-VIS spectrophotometer (Shimadzu UV-1800, Shimadzu Inc., Kyoto, Japan). Total chlorophyll was calculated using the formula described in Lichtenthaler (1987) and ranged between 396.2 – 821.9 $\mu\text{g/g}$ in this study. For sugar analysis, soluble sugars were extracted from 100 mg samples with 80% ethanol and centrifuged for 10 min at 10,016 g; this process was repeated twice. Total soluble sugars were assayed according to the colorimetric method described in DuBois *et al.* (1956) and ranged between 2600 – 28300 $\mu\text{g/g}$ in this study.

Automated hyperspectral vegetation index (AutoVI) derivation

An automated system for hyperspectral vegetation index derivation named AutoVI, was developed to deliver effective VIs in a streamlined process covering model selection, model parameter generation, model parameter tuning, and model evaluation driven by a hyperparameter optimization framework, i.e. the optimizer, for plant phenotyping (Fig. 1). The optimizer is the core of the AutoVI method which seeks to generate a VI model for the desired trait based on a model evaluation metric, i.e. the objective function score (R^2 , see below) using an optimization algorithm given time and computing resource constraints. However, unlike simple optimization challenges which typically search for optimal solutions for a single model or function (static search space), multiple index models (dynamic search spaces) are optimized and evaluated in AutoVI (Fig. 1). This is made possible using dynamic parameter programming or “define-by-run” coding (Akiba *et al.*, 2019), which generates the search space or set of model parameters during code execution depending on the index model or equation under evaluation.

The mathematical expression (or model) defining a VI is primarily needed to combine two or more wavebands to decipher certain biochemical or biophysical traits of interest. The development of a new VI requires the selection of both a suitable model and a set of wavebands. A library of 33 index models was created (Table S2) after an exhaustive review of more than 500 previously developed VIs (Henrich *et al.*, 2017). The listed model equations are of varying complexities, differing in the number of distinct wavebands ($N_{\text{wb}} = 2$ to 6, i.e. B1, B2,... B6) and the number of coefficients ($N_{\text{cf}} = 1$

to 5, i.e. α , β ,... p). The AutoVI system begins with model selection (step 1, Fig. 1) in which the optimizer selects a mathematical model at random from the library of 33 index models and generates the model parameters corresponding to the N_{wb} and N_{cf} (step 2, Fig. 1). This is followed by model parameter tuning (step 3, Fig. 1) where optimum wavebands and coefficient values (between 0 – 1) are selected to maximize the objective function score (step 4, Fig. 1), which in this study is the coefficient of determination, R^2 , derived from a simple linear regression fitted to calculated index values and ground truth values for the measured trait of interest (in this case chlorophyll or sugar content). One run of optimization (steps 1 to 4) is referred to as a single iteration; the AutoVI system is programmed to repeat the optimization process until a pre-determined number of iterations is reached. At each iteration, an index model gets selected, and the computed objective function score (R^2) is compared to the previous iteration. Additionally, unique sets of model parameters are computed for the respective index models. The repeated iterations seek to maximize the objective function score, R^2 , by selecting the best candidate model and an optimum set of model parameters.

In this study, the tree parzen estimator (TPE) (Bergstra *et al.*, 2011; Yu & Zhu, 2020) was implemented as the optimization algorithm in AutoVI as it is widely employed for hyperparameter optimization in machine learning problems and showed better accuracy and efficiency compared to other algorithms when dealing with dynamic search spaces (Bergstra *et al.*, 2015; Akiba *et al.*, 2019; Yu & Zhu, 2020). The TPE algorithm is a variant of Bayesian optimization approaches which tries to construct a probabilistic model, also known as a “surrogate” model for mapping hyperparameters based on the probability of an objective function score given the set of hyperparameters. The AutoVI system, including the TPE algorithm customized to handle hyperspectral data was implemented in Python using the open source hyperparameter optimization library, Optuna version 2.0 (Akiba *et al.*, 2019) with default settings. A graphical user interface (GUI) for AutoVI and source codes for TPE implementation are hosted on the public repository, GitHub (see Data availability). The AutoVI system was tested on an AMD Threadripper 3970X (32-cores) system with 256 GB RAM at the SmartSense iHub, Agriculture Victoria.

AutoVI for chlorophyll content estimation

We evaluated the ability of AutoVI to derive high quality novel hyperspectral VIs for plant phenotyping using wheat’s total chlorophyll content as a biochemical trait. Chlorophyll content, either measured or estimated, can be a direct indicator of a plant’s primary production and has been used to determine N status and stress response of crop plants (Richardson *et al.*, 2002; Murchie & Lawson, 2013). The existing dataset (n=80) was randomly split 65:35 into training and test datasets,

with both datasets having the same sample distribution (i.e. stratified sampling) according to time points. The AutoVI system was trained on the training dataset to derive novel indices for chlorophyll content estimation, with the performance of these indices evaluated using the test dataset.

One possible issue with any optimization system is a selection bias towards index models with lower complexity, e.g. models with N_{wb} between 2 and 3 compared to those with higher complexity, e.g. models with $N_{wb} \geq 4$. This is because the size of the solution search space increases exponentially with the increase in the number of input features i.e. N_{wb} (Winston, 1992; Yao & Liu, 1997). Consequently, more computational time or resource is required to optimize the complex models ($N_{wb} \geq 4$) compared to the simpler models ($N_{wb} \leq 3$) but their performance tend to scale better over time and do not plateau as fast as simpler models. When all model computations are grouped together, simpler models tend to outperform complex models in early stages of optimization, causing these to be favored by the optimizer and leading towards a ‘locally maximal solution’, which is the tendency of the computation to get stuck at a sub-optimal solution (Hinneburg & Keim, 1999). To address this issue in AutoVI, computations were performed on model groups consisting of models with the same N_{wb} (Fig. 2). Five parallel instances of AutoVI corresponding to the five model groups (M2, M3,... M6, equivalent to $N_w = 2, 3, \dots 6$) were executed with 20,000 iterations each with coefficients fixed at 1 (Fig. 2). These were repeated five times with the best performing index model from each group logged at each repetition (Fig. 2). This allows the reliable identification of the best index model for each model group and the overall best index model. The effect of longer optimizations and inclusion of coefficient on model performance was determined using the overall best index model, with five repeated AutoVI computations at 20,000 and 40,000 iterations with and without coefficient tuning.

To determine the quality of AutoVI-derived indices for chlorophyll estimation, they were used as features in simple linear regression (SLR) modelling to predict chlorophyll content. SLR with each of the derived indices was first trained on the training dataset and then used to predict chlorophyll values for the test dataset. Model performance was evaluated using the R^2 score calculated between predicted and actual chlorophyll values. In addition, performance for a stepwise multiple linear regression (SMLR) model (see below) with VIs selected from 25 AutoVI-derived indices (Fig. 2) is also included for comparison. Results achieved using AutoVI-indices were compared to those produced by SLR and SMLR with 47 published VIs, in addition to results provided by partial least-squares regression (PLSR) modelling using the full spectrum of reflectance values, i.e., reflectance values from all 1,235 wavebands (see below). For comparison across different regression models, additional

performance metrics such as root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are also provided.

AutoVI for sugar content estimation

Sugar plays an important role in the osmotic adjustment of plants in response to drought stress. Studies have shown that genotypes that show higher accumulation of sugar content in leaves or stems are more drought tolerant (Adams *et al.*, 2013; Piaskowski *et al.*, 2016). Using the grouped model evaluation approach, AutoVI computations were conducted across five repetitions with 20,000 iterations each with the coefficients fixed at 1. The existing dataset was split 65:35 into training and test datasets as described previously, with the training of AutoVI conducted on the training dataset and validation of derived indices performed on the test dataset. Results for SLR and SMLR using AutoVI-derived indices were compared to those produced using 47 published VIs and PLSR modelling, as described in the previous section.

Stepwise Multiple Linear Regression

Stepwise multiple linear regression (SMLR) is a feature selection method which iteratively adds (forward selection) or removes (backward selection) features to a multiple linear regression model to improve model performance, as indicated by an evaluation metric or score. In this study, we implemented a stepwise forward selection strategy based on a five-fold cross validated R^2 score of a multiple linear regression model using the Python package, scikit-learn version 0.24. The maximum number of features to select was set to between 1 – 20 and selection was performed on 25 AutoVI-derived indices and 47 published VIs for both chlorophyll and sugar estimation on the training dataset. A multiple linear regression model was then fitted to the training dataset using the optimum selected features and used to predict target values (chlorophyll or sugar content) for the test dataset.

Partial Least Squares Regression

Partial least-squares regression (PLSR) modelling is one of the most effective methods for plant trait prediction using hyperspectral data (Ely *et al.*, 2019; Wu *et al.*, 2019; Burnett *et al.*, 2021). PLSR was designed to address both the collinearity between predictors, i.e., the different wavebands of a reflectance spectrum, and the large number of predictor variables when compared to trait observations. This study implemented PLSR modelling using the Python package, scikit-learn version 0.24 for chlorophyll and sugar content estimation. The optimal number of PLSR components was first determined based on five-fold cross validated R^2 scores of PLSR models fitted on the training

dataset with the number of components set to between 1 – 20 (Fig. S1). A PLSR model was then fitted to the training dataset using the optimum number of components (n=6 for chlorophyll and n=7 for sugar, Fig. S1) and used to predict target values (chlorophyll or sugar contents) for the test dataset.

Results

AutoVI for chlorophyll content estimation

AutoVI performance was measured using R^2 scores generated by simple linear regression models on the test dataset with the respective AutoVI-derived indices as features. AutoVI performance across five repetitions was relatively stable, with the grouped evaluation strategy allowing for comparison across different model groups and identification of the best performing index model within each model group (Fig. 3). Between model groups, the M4 group ($N_{wb}=4$) had the best mean R^2 of 0.7818, followed by M5 ($N_{wb}=5$) with R^2 of 0.7747, M6 ($N_{wb}=6$) with R^2 of 0.7689, M2 ($N_{wb}=2$) with R^2 of 0.7637 and finally M3 ($N_{wb}=3$) with R^2 of 0.7555. Overall, Index25 ($N_{wb}=4$, $N_{cf}=1$) produced the best VI ($R^2=0.8007$) and generated the best results across all five repetitions within the M4 group (Fig. 3). The performance of the best VIs according to model group is summarized in Table 1. Wavebands selected by AutoVI for the best chlorophyll indices derive predominantly from the red (600 – 720 nm) and red-edge (720 – 780 nm) regions, with a few wavebands from the blue (470 – 490 nm), near infrared (NIR, 1000 – 1300 nm) and shortwave infrared (SWIR, 1600 - 1700 nm) regions (Table 1). The R^2 score produced by the best VI, termed hereafter as AutoVI chlorophyll index (AutoVI-Chl), was achieved using wavebands of 610 nm, 716 nm, 1384 nm and 1607 nm, as depicted in equation 1:

Equation 1

$$AutoVI - Chl = \frac{(R_{1607} - R_{1384}) - (R_{1607} - R_{716}) \frac{R_{1607}}{R_{1384}}}{* \frac{(R_{610} - R_{1384})}{(R_{610} + R_{1384} + 1)}}$$

where, R_{wb} represent the reflectance measured at a discrete waveband (wb).

The performance of AutoVI-Chl with or without a coefficient variable, alpha (α), as depicted in its original equation (Index25, Table S2) was determined across five computational repetitions of 20,000 and 40,000 iterations (Fig. 4). At 20,000 iterations, the inclusion of the coefficient had minimal impact on AutoVI-Chl performance, as boxplots for R^2 scores obtained with the coefficient (min=0.7657, median=0.7771, max=0.7959) and without the coefficient (min=0.7673, median=0.7875, max=0.7922) were comparable (Fig. 4). However, AutoVI computational time when

the coefficient was included (~2.1 hours for 5 repetitions) was up to 1.6x higher than without the coefficient (~1.3 hours for 5 repetitions), suggesting that inclusion of coefficient(s) in AutoVI optimizations will likely incur computational costs. At 40,000 iterations, AutoVI-Chl performance deteriorated significantly with or without the coefficient (Fig. 4), suggesting that overfitting, where a model performs significantly better on the training (i.e. overfitted) but not the test dataset (i.e. unable to generalize to new data), is likely to occur with longer AutoVI runs. Based on these results, the recommended starting point for AutoVI training is up to 20,000 iterations and without coefficient tuning.

The quality of AutoVI-derived indices for chlorophyll content estimation was evaluated further against 47 published VIs, as features in simple linear regression (SLR) modelling. First, the best SLR model resulting from AutoVI-indices and the best SLR model with existing VIs was compared (Table 2). The model with AutoVI-Chl ($R^2=0.8007$, RMSE=38.52, MAE=30.51, MAPE=4.69%) significantly outperformed the model with the normalized difference chlorophyll index, NDCI ($R^2=0.6018$, RMSE=54.45, MAE=46.05, MAPE=7.09%) (Table 2, S3). Next, stepwise multiple linear regression (SMLR) models using the optimum subset of features selected from AutoVI-indices and existing VIs were compared (Table 2). For the existing VIs, SMLR with 7 VIs selected (Table S4) led to a significant improvement in model performance ($R^2=0.7136$, RMSE=46.17, MAE=38.12, MAPE=5.88%) compared to SLR with NDCI but was still inferior to SLR with AutoVI-Chl; SMLR with four AutoVI-indices selected (Table S5) did not perform better ($R^2=0.7989$, RMSE=38.69, MAE=30.98, MAPE=4.88%) compared to SLR with AutoVI-Chl. Finally, PLSR modelling performance for chlorophyll estimation was included as an additional benchmark for comparison. The PLSR model ($R^2=0.7379$, RMSE=44.17, MAE=33.81, MAPE=5.36%) did not perform as well as SLR with AutoVI-Chl or SMLR with the selected AutoVI-indices (Table 2). Overall, the best modelling performance was provided by SLR with AutoVI-Chl. These results support AutoVI as an efficient system for novel trait-specific hyperspectral VI derivation.

AutoVI for sugar content estimation

AutoVI performance across five repetitions was relatively stable, with the M6 group having the best mean R^2 of 0.8201, followed by M3 group with R^2 of 0.8127, M4 group with R^2 of 0.7933, M5 group with R^2 of 0.7877 and finally M2 group with R^2 of 0.7591 (Fig. 5). The overall best VI was produced by Index33 ($N_{wb}=6$, $N_{cf}=1$), which also generated the best results for three repetitions within the M6 group (Fig. 5). The performance of the best VIs according to model group is summarized in Table 3. Wavebands selected by AutoVI for the best sugar indices derive predominantly from the shortwave

infrared (1400 – 1700 nm) and near infrared (770 – 1370 nm) regions, with a few bands from the VIS (499 – 644 nm) region (Table 3). The R^2 score produced by the best VI, termed hereafter as AutoVI sugar index (AutoVI-Sgr), was achieved using wavebands of 499 nm, 773 nm, 1179 nm, 1291 nm, 1425 nm and 1661 nm, as depicted in equation 2:

Equation 2

$$AutoVI - Sgr = \frac{(R_{773} - R_{1425}) - (R_{773} - R_{1179}) \frac{R_{773}}{R_{1425}}}{2 * \frac{(R_{1661} - R_{1425})}{(R_{1661} + R_{1425} + R_{499} + R_{1291} + 1)}}$$

where, R_{wb} represent the reflectance measured at a discrete waveband (wb).

The quality of AutoVI-derived indices for sugar content estimation was evaluated further against 47 published VIs, as features in simple linear regression (SLR) modelling. The best SLR performance with AutoVI-derived indices was achieved using AutoVI-Sgr ($R^2=0.8339$, RMSE=2612.41, MAE=2148.15, MAPE=20.17%), which significantly outperformed the best SLR model achieved with the published VI, Gitelson and Merzlyak Index 2, GMI2 ($R^2=0.4695$, RMSE=4668.35, MAE=3939.59, MAPE=38.96%). In general, SLR modelling performance with existing VIs was very poor (Table S6). SMLR with five existing VIs selected (Table S7) produced dramatically better results ($R^2=0.7387$, RMSE=3276.50, MAE=2401.14, MAPE=22.30%) compared to SLR with GMI2 but was still inferior to the model with AutoVI-Sgr (Table 4). On the other hand, SMLR with four AutoVI-indices selected (Table S8) performed better ($R^2=0.8587$, RMSE=2409.19, MAE=2071.47, MAPE=19.16%) compared to SLR with AutoVI-Sgr (Table 4). PLSR modelling performance for sugar estimation was also included as a benchmark for comparison. The PLSR model had similar performance ($R^2=0.8322$, RMSE=2625.99, MAE=2212.89, MAPE=21.19%) as SLR with AutoVI-Sgr but was outperformed by SMLR with the AutoVI-indices (Table 4). These results further support AutoVI as an efficient system for novel VI derivation, with AutoVI-indices as high-quality features for trait prediction.

Discussion

In this study, we describe the design and implementation of an automated system for hyperspectral vegetation index derivation, AutoVI, for plant phenotyping using as examples chlorophyll and sugar content estimation in wheat. In both cases, indices generated by AutoVI significantly outperformed existing VIs in simple and multiple linear regression modelling, with performance exceeding that of a more complex model such as PLSR. The success of AutoVI can be attributed to the use of a highly performant hyperparameter optimization algorithm, tree parzen estimator (TPE) (Bergstra *et al.*, 2011; Akiba *et al.*, 2019).

Results in this study showed that AutoVI was able to efficiently generate high-quality 2- to 6-band indices specific to chlorophyll and sugar content. A review of existing literature highlights that most VIs currently deployed consist predominantly of 2-band indices, and to a lesser extent 3-band indices (Henrich *et al.*, 2017). Indices with 4 bands or more are rare, presumably due to the difficulty in optimizing such indices (e.g. band selection) attributed to the curse of high dimensionality and issues associated with the multicollinearity of bands. Although a myriad of hyperspectral band selection strategy exists (Sun & Du, 2019), most of these were explicitly developed for image classification or regression models. Previous studies which focused on generating novel hyperspectral VIs used single or multiple correlation matrices between VI pairs and the trait of interest to uncover new band or index combinations (Thenkabail *et al.*, 2004; Aasen *et al.*, 2014; Xu *et al.*, 2019). However, these approaches are computationally expensive as every possible combination of available bands (filtered or unfiltered) needs to be computed. Therefore, previous efforts have focused only on limited combinations between a few bands (2- and 3-band indices), thereby also limiting the net performance of the derived VIs. Consequently, the limited availability of specific VIs as biomarkers for plant traits has forced researchers to generalize upon the applicability of existing VIs, for example, the NDVI (2-band index), across almost all aspects of research and analysis, leading to suboptimal results. The AutoVI system is well-positioned to address this issue through the rapid generation of trait-specific novel 2-band to 6-band VIs without requiring band filtering or dimension reduction techniques to limit the number of input hyperspectral bands prior to processing.

Biochemical constituents in plants absorb electromagnetic energy in specific wavelength regions. Vegetation indices profiled around these characteristic spectral absorption regions can detect or estimate the biochemical trait of interest. The AutoVI system is constructed to automate the identification of these critical spectral regions using the underlying TPE algorithm. For chlorophyll, the selected wavebands centred around the red (600 – 700nm) and red-edge (700 – 740 nm) regions, with a few bands from the blue (470 – 490 nm), NIR (1000 – 1300 nm) and SWIR (1600 - 1700 nm) regions. Chlorophylls (chlorophyll *a* and *b*) are the most important plant pigments which function as photoreceptors and catalysts for photosynthesis, the photochemical synthesis of carbohydrates (Blackburn, 2006). As such, chlorophyll content in leaves and canopies is a key indicator of physiological measures such as photosynthetic capacity, developmental stage, productivity and stress (Richardson *et al.*, 2002; Murchie & Lawson, 2013). Studies have shown that reflectance of wavelengths in the red region (~530 – 630 nm, and a narrower band around 700 nm) is most sensitive to chlorophyll pigment concentrations across the normal range found in most leaves and

canopies (Lichtenthaler *et al.*, 1996; Gitelson *et al.*, 2005). In addition, research has also shown that bands within the red-edge region (RE, 680 – 740 nm), which delineates the border between chlorophyll absorption in red wavelengths and leaf scattering in the NIR wavelengths, are strongly correlated with chlorophyll content (Curran *et al.*, 1991; Gitelson *et al.*, 1996). Existing chlorophyll VIs consist mainly of 2-band indices derived from ratios of narrow bands within regions of spectrum sensitive to chlorophyll pigments (VIS-RE, 400 – 740 nm) and those areas not sensitive to the pigments and/or related to some other control on reflectance (typically NIR, 750 – 900 nm) (Blackburn, 2006; Wu *et al.*, 2008). Wavebands selected by AutoVI agree with published studies and additional wavebands selected from the SWIR region likely enhanced the sensitivity of some of the AutoVI-indices to chlorophyll by acting as control on reflectance.

The sugar metabolic pathway is intrinsically linked with the regulation of plant growth and development, and response to stress (Julius *et al.*, 2017; Kaur *et al.*, 2021). Studies have shown that abiotic stress such as drought or heat triggers sugar accumulation, particularly soluble sugars such as sucrose in plants (Lemoine *et al.*, 2013; Zhou *et al.*, 2017). In this study, AutoVI generated novel indices with a strong correlation to total sugar content, with specific wavebands selected mainly from the SWIR (1400 – 1700 nm) and NIR (770 – 1370 nm) regions. It is known that leaf reflectance properties in the SWIR region (1300 – 2500 nm) is governed by water content and biochemical compounds such as cellulose, sugars and starch (Elvidge, 1990; Kokaly *et al.*, 2009). Indeed, more recent studies have used near infrared spectroscopy (750 – 2500 nm) coupled with machine learning and/or statistical models to estimate soluble carbohydrates, including total sugar in various plant tissue and organ such as leaf and stem (Adams *et al.*, 2013; Piaskowski *et al.*, 2016). Research in rice has also identified wavebands in the NIR region (800 – 1100 nm) as being important for sugar content estimation in leaves (Das *et al.*, 2018). These studies provide support for bands selected by AutoVI as being specific to total sugar content. Models for sugar estimation in this study were less robust overall (e.g. MAPE = 19.16 – 22.30%) compared to models for chlorophyll estimation (e.g. MAPE = 4.69 – 5.23%) as reflected in the higher error scores. This may be due to difficulties in obtaining high quality spectral signatures from the SWIR region as water vapor is known to obscure spectral signatures for biochemical compounds in this region in plants (Elvidge, 1990; Kokaly *et al.*, 2009). Nevertheless, it is noteworthy that AutoVI was able to select bands in the SWIR region whilst avoiding the water absorption peak at 1450 nm (Elvidge, 1990; Kokaly *et al.*, 2009).

Novel AutoVI-derived indices for chlorophyll and sugar content estimation were first compared against 47 published VIs as features in simple linear regression (SLR). This provided a good baseline

to compare VIs as model performance would depend solely on the quality of the VI. For chlorophyll estimation, the best SLR achieved with the AutoVI-derived index, AutoVI-Chl ($R^2=0.8007$) significantly outperformed the best SLR with the published VI, NDCl ($R^2=0.6018$). Similarly, for sugar estimation, the best performance provided by SLR with AutoVI-Sgr ($R^2=0.8339$) significantly outperformed the SLR with the published VI, GM12 ($R^2=0.4695$). Next, the performance of a stepwise multiple linear regression (SMLR) with features selected from existing VIs was compared to the SLR models with AutoVI-indices. Results showed that although SMLR did significantly improve performance for both chlorophyll ($R^2=0.7136$) and sugar ($R^2=0.7387$) estimation with published VIs, these were still inferior compared to SLR results achieved using the AutoVI-indices. SMLR was also performed using the AutoVI-indices, with performance enhancement observed for sugar ($R^2=0.8587$) but not chlorophyll ($R^2=0.7989$) estimation. Finally, model performance for chlorophyll and sugar estimation using AutoVI-indices were compared to results produced by PLSR, a well-established method for plant trait modelling using hyperspectral data (Ely *et al.*, 2019; Wu *et al.*, 2019; Burnett *et al.*, 2021). PLSR represents a different modelling approach as it projects the entire spectrum of reflectance values (or the predictor variables) into a smaller number of variables (or components) whilst simultaneously maximizes the correlation between the response and the variables (Geladi & Kowalski, 1986; Wold *et al.*, 2001). With PLSR, what is being compared to the models with AutoVI-indices is the quality of feature representation or learning as depicted by the projected components in PLSR in contrast to the functions encoded in the AutoVI-indices. For both chlorophyll and sugar estimation, SLR and SMLR with AutoVI-indices outperformed PLSR ($R^2=0.7379$ for chlorophyll; $R^2=0.8322$ for sugar). Impressively, SLR with the best AutoVI-derived index (AutoVI-Chl or AutoVI-Sgr) produced better results and outperformed more complex approaches such as SMLR (except for sugar estimation) and PLSR. These results provide strong support for AutoVI-derived indices as high-quality features and affirm AutoVI as a high-performing system for novel trait-specific hyperspectral VI derivation.

However, AutoVI-indices, including AutoVI-Chl and AutoVI-Sgr, are not perfected VIs and will benefit from further studies incorporating more data collected from multiple genotypes across different environments and growth stages. Particularly for field-based phenotyping, AutoVI-indices should ideally be derived using canopy spectra, as these are known to differ from leaf spectra (Croft *et al.*, 2014). Depending on the underlying trait, it may be worthwhile to collect samples from different plant tissues or organ. For example, studies have shown that genotypes which accumulate more sugar in leaf and stem tissues are more heat and/or drought tolerant (Piaskowski *et al.*, 2016; Zhou *et al.*, 2017). Fortunately, AutoVI does not impose any size or dimensional constraint on the input data and is expected to work with data derived from different hyperspectral sensors and/or spectra

sources. Depending on the data provided, AutoVI can be customized to deliver trait-specific VIs according to species, genotype, growth stage and environment, making it a powerful and versatile tool for both novice and expert users alike. In addition, compared to machine learning and statistical modelling approaches where model optimization and deployment remains technically challenging, AutoVI-derived indices can be easily computed and readily deployed for HTP without requiring complex hardware or software resources.

In conclusion, results in our study demonstrate that AutoVI is an efficient and powerful tool for deriving high-quality hyperspectral VIs for plant phenotyping. AutoVI-derived VIs outperformed existing VIs and delivered strong performance in SLR and SMLR modelling for chlorophyll and sugar content estimation, producing results superior to PLSR. The AutoVI system is expected to accelerate the development of novel VIs for plant/crop traits which will find wide application in HTP and agriculture remote sensing vital to improving breeding program and crop management efficiencies.

Table 1. Performance of the best AutoVI-derived indices according to model group for chlorophyll estimation

Group	Index Model	Selected Wavelengths (nm)	R ²	RMSE	MAE	MAPE
M2	2	637, 711	0.7683	41.52	33.78	5.14%
M3	15	607, 716, 1712	0.7629	42.01	34.27	5.23%
M4	25	610, 716, 1384, 1607	0.8007	38.52	30.51	4.69%
M5	26	497, 650, 707, 778, 1017	0.7813	40.35	32.74	5.13%
M6	32	477, 635, 713, 1033, 1035, 1668	0.7776	40.69	33.01	5.03%

Performance metrics calculated for simple linear regression using the indicated AutoVI-derived index for chlorophyll estimation on the test dataset. The best scores are indicated in bold.

Table 2. Comparison between different regression models for chlorophyll estimation

Model	Feature(s)	R ²	RMSE	MAE	MAPE
SLR	AutoVI-Chl	0.8007	38.52	30.51	4.69%
SLR	NDCI	0.6018	54.45	46.05	7.09%
SMLR	7 Existing VIs	0.7136	46.17	38.12	5.88%
SMLR	4 AutoVI-Indices	0.7989	38.69	30.98	4.88%
PLSR	Reflectance Values	0.7379	44.17	33.81	5.36%

Performance metrics calculated for simple linear regression (SLR) and stepwise multiple regression (SMLR) using AutoVI-derived indices or existing 47 vegetation indices, in addition to partial least squares regression (PLSR) using the full spectrum of reflectance values for chlorophyll estimation on the test dataset. The best scores are indicated in bold.

Table 3. Performance of the best AutoVI-derived indices according to model group for sugar estimation

Group	Index Model	Selected Wavelengths (nm)	R ²	RMSE	MAE	MAPE
M2	3	1420, 1588	0.7693	3078.89	2493.21	22.94%
M3	15	1271, 1417, 1650	0.8209	2712.72	2237.33	20.31%
M4	24	1371, 1422, 1650, 1712	0.8055	2827.07	2466.19	23.57%
M5	29	644, 990, 1406, 1418, 1649	0.7860	2964.88	2387.41	21.66%
M6	33	499, 773, 1179, 1291, 1425, 1661	0.8339	2612.41	2148.15	20.17%

Performance metrics calculated for simple linear regression using the indicated AutoVI-derived index for chlorophyll estimation on the test dataset. The best scores are indicated in bold.

Table 4. Comparison between different regression models for sugar estimation

Model	Feature(s)	R ²	RMSE	MAE	MAPE
SLR	AutoVI-Sgr	0.8339	2612.41	2148.15	20.17%
SLR	GMI2	0.4695	4668.95	3939.59	38.96%
SMLR	5 Existing VIs	0.7387	3276.50	2401.14	22.30%
SMLR	4 AutoVI-Indices	0.8587	2409.19	2071.47	19.16%
PLSR	Reflectance Values	0.8322	2625.99	2212.89	21.19%

Performance metrics calculated for simple linear regression (SLR) and stepwise multiple regression (SMLR) using AutoVI-derived indices or existing 47 vegetation indices, in addition to partial least squares regression (PLSR) using the full spectrum of reflectance values for sugar estimation on the test dataset. The best scores are indicated in bold.

Author contributions

JCOK and BPB designed the AutoVI system and analyzed the data. JCOK implemented the AutoVI system and GUI version in Python. BPB researched and collated the list of index models used in AutoVI. JCOK, BPB, GS and SK conceived and designed the study. JCOK, BPB, GS and SK wrote and edited the paper. All authors read and approved the final manuscript.

Data availability

All relevant source codes and datasets, including a GUI implementation of AutoVI in Python for Windows 10 64-bit operating system required to reproduce results reported in this study are available at www.github.com/AVR-SmartSense/AutoVI.

Figure legends

Figure 1. Overview of the AutoVI framework. The tree parzen estimator (TPE) algorithm is implemented as the optimizer for AutoVI. A single iteration consists of model selection, model

parameter generation, model parameter tuning and model evaluation. AutoVI is programmed to repeat the optimization process until a pre-determined number of iterations is reached. The repeated iterations seek to maximize the objective function score, R^2 , by selecting the best candidate model and an optimum set of hyperparameters.

Figure 2. Flowchart for grouped model evaluations in AutoVI. Index models were grouped according to the number of wavebands (N_{wb}) and parallel computations with 20,000 iterations each were performed on the model groups (M2, M3, M4, M5 and M6). The best index model for each model group and the overall best index model were identified from five repeated computations.

Figure 3. Best-performing index models from grouped model evaluations for total chlorophyll content estimation. The best index models from grouped (M2, M3, M4, M5 and M6) model evaluations were identified across five repeated AutoVI computations. Index model names are indicated within the bar figure. The overall best index model was Index25 from the M4 group with R^2 score of 0.8007 (indicated by green dashed line).

Figure 4. Effect of coefficient tuning and longer iterations on AutoVI chlorophyll index (AutoVI-Chl) performance. The inclusion (Yes) or exclusion (No) of a single coefficient, α , on AutoVI-Chl performance was determined across five repeated computations at 20,000 (20k) and 40,000 (40k) iterations each. The distribution of R^2 scores are shown in the boxplot.

Figure 5. Best-performing index models from grouped model evaluations for total sugar content estimation. The best index models from grouped (M2, M3, M4, M5 and M6) model evaluations were identified across five repeated AutoVI computations. Index model names are indicated within the bar figure. The overall best index model was Index33 from the M6 group with R^2 score of 0.8339 (indicated by green dashed line).

Supporting Information

Figure S1. Optimization of number of components in PLSR for chlorophyll and sugar estimation.

Table S1. List of 47 published vegetation indices.

Table S2. Index models for AutoVI selection.

Table S3. Performance of 47 published vegetation indices for chlorophyll content estimation.

Table S4. Stepwise multiple linear regression with 47 published vegetation indices for chlorophyll estimation

Table S5. Stepwise multiple linear regression with AutoVI-derived indices for chlorophyll estimation

Table S6. Performance of 47 published vegetation indices for sugar content estimation.

Table S7. Stepwise multiple linear regression with 47 published vegetation indices for sugar estimation

Table S8. Stepwise multiple linear regression with AutoVI-derived indices for sugar estimation

References

Aasen H, Gnyp ML, Miao Y, Bareth G. 2014. Automated Hyperspectral Vegetation Index Retrieval from Multiple Correlation Matrices with HyperCor. *Photogrammetric Engineering & Remote Sensing* **80**(8): 785-795.

Aasen H, Honkavaara E, Lucieer A, Zarco-Tejada PJ. 2018. Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: a review of sensor technology, measurement procedures, and data correction workflows. *Remote Sensing* **10**(7): 1091.

Adams HD, Germino MJ, Breshears DD, Barron-Gafford GA, Guardiola-Claramonte M, Zou CB, Huxman TE. 2013. Nonstructural leaf carbohydrate dynamics of *Pinus edulis* during drought-induced tree mortality reveal role for carbon metabolism in mortality mechanism. *New Phytologist* **197**(4): 1142-1151.

Adão T, Hruška J, Pádua L, Bessa J, Peres E, Morais R, Sousa JJ. 2017. Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry. *Remote Sensing* **9**(11): 1110.

Akiba T, Sano S, Yanase T, Ohta T, Koyama M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Bajwa SG, Kulkarni SS. 2011. *Hyperspectral data mining*: Boca Raton, London, New York: CRC Press/Taylor and Francis Group.

Banerjee BP, Joshi S, Thoday-Kennedy E, Pasam RK, Tibbits J, Hayden M, Spangenberg G, Kant S. 2020. High-throughput phenotyping using digital and hyperspectral imaging-derived biomarkers for genotypic nitrogen response. *Journal of Experimental Botany*.

Bergstra J, Bardenet R, Bengio Y, Kégl B 2011. Algorithms for hyper-parameter optimization. *Proceedings of the 24th International Conference on Neural Information Processing Systems*. Granada, Spain: Curran Associates Inc. 2546–2554.

618 **Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD. 2015.** Hyperopt: a Python library for model
619 selection and hyperparameter optimization. *Computational Science & Discovery* **8**(1):
620 014008.

621 **Blackburn GA. 2006.** Hyperspectral remote sensing of plant pigments. *Journal of Experimental*
622 *Botany* **58**(4): 855-867.

623 **Burger J, Gowen A. 2011.** Data handling in hyperspectral image analysis. *Chemometrics and*
624 *Intelligent Laboratory Systems* **108**(1): 13-22.

625 **Burnett AC, Anderson J, Davidson KJ, Ely KS, Lamour J, Li Q, Morrison BD, Yang D, Rogers A, Serbin**
626 **SP. 2021.** A best-practice guide to predicting plant traits from leaf-level hyperspectral data
627 using partial least squares regression. *Journal of Experimental Botany*.

628 **Chauhan HJ, Mohan BK. 2013.** Development of agricultural crops spectral library and classification
629 of crops using Hyperion hyperspectral data. *Journal of Remote Sensing Technology* **1**(1): 9.

630 **Chein IC 1999.** Spectral information divergence for hyperspectral image analysis. *IEEE 1999*
631 *International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293).*
632 509-511 vol.501.

633 **Croft H, Chen JM, Zhang Y. 2014.** The applicability of empirical vegetation indices for determining
634 leaf chlorophyll content over different leaf and canopy structures. *Ecological Complexity* **17**:
635 119-130.

636 **Curran PJ, Dungan JL, Macler BA, Plummer SE. 1991.** The effect of a red leaf pigment on the
637 relationship between red edge and chlorophyll concentration. *Remote Sensing of*
638 *Environment* **35**(1): 69-76.

639 **Das B, Sahoo RN, Pargal S, Krishna G, Verma R, Chinnusamy V, Sehgal VK, Gupta VK, Dash SK,**
640 **Swain P. 2018.** Quantitative monitoring of sucrose, reducing sugar and total sugar dynamics
641 for phenotyping of water-deficit stress tolerance in rice through spectroscopy and
642 chemometrics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **192**:
643 41-51.

644 **DuBois M, Gilles KA, Hamilton JK, Rebers PA, Smith F. 1956.** Colorimetric Method for Determination
645 of Sugars and Related Substances. *Analytical Chemistry* **28**(3): 350-356.

646 **Elvidge CD. 1990.** Visible and near infrared reflectance characteristics of dry plant materials.
647 *International Journal of Remote Sensing* **11**(10): 1775-1795.

648 **Ely KS, Burnett AC, Lieberman-Cribbin W, Serbin SP, Rogers A. 2019.** Spectroscopy can predict key
649 leaf traits associated with source–sink balance and carbon–nitrogen status. *Journal of*
650 *Experimental Botany* **70**(6): 1789-1799.

651 **Geladi P, Kowalski BR. 1986.** Partial least-squares regression: a tutorial. *Analytica Chimica Acta* **185**:
652 1-17.

653 **Gitelson AA, Merzlyak MN, Lichtenthaler HK. 1996.** Detection of Red Edge Position and Chlorophyll
654 Content by Reflectance Measurements Near 700 nm. *Journal of Plant Physiology* **148**(3):
655 501-508.

656 **Gitelson AA, Viña A, Ciganda V, Rundquist DC, Arkebauer TJ. 2005.** Remote estimation of canopy
657 chlorophyll content in crops. *Geophysical Research Letters* **32**(8).

658 **Henrich V, Krauss G, Götze C, Sandow C 2017.** Index Database: A database for remote sensing
659 indices.

660 **Hinneburg A, Keim DA 1999.** Optimal grid-clustering: Towards breaking the curse of dimensionality
661 in high-dimensional clustering.

662 **Julius BT, Leach KA, Tran TM, Mertz RA, Braun DM. 2017.** Sugar Transporters in Plants: New Insights
663 and Discoveries. *Plant and Cell Physiology* **58**(9): 1442-1460.

664 **Kaur H, Manna M, Thakur T, Gautam V, Salvi P. 2021.** Imperative role of sugar signaling and
665 transport during drought stress responses in plants. *Physiologia Plantarum* **171**(4): 833-848.

666 **Kokaly RF, Asner GP, Ollinger SV, Martin ME, Wessman CA. 2009.** Characterizing canopy
667 biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote*
668 *Sensing of Environment* **113**: S78-S91.

669 **Lemoine R, La Camera S, Atanassova R, Dédaldéchamp F, Allario T, Pourtau N, Bonnemain J-L, Laloi**
670 **M, Coutos-Thévenot P, Maurousset L, et al. 2013.** Source-to-sink transport of sugar and
671 regulation by environmental factors. *Frontiers in Plant Science* **4**(272).

672 **Lichtenthaler HK 1987.** [34] Chlorophylls and carotenoids: Pigments of photosynthetic
673 biomembranes. *Methods in Enzymology*: Academic Press, 350-382.

674 **Lichtenthaler HK, Gitelson A, Lang M. 1996.** Non-Destructive Determination of Chlorophyll Content
675 of Leaves of a Green and an Aurea Mutant of Tobacco by Reflectance Measurements.
676 *Journal of Plant Physiology* **148**(3): 483-493.

677 **Lu B, Dao PD, Liu J, He Y, Shang J. 2020.** Recent Advances of Hyperspectral Imaging Technology and
678 Applications in Agriculture. *Remote Sensing* **12**(16): 2659.

679 **Mir RR, Reynolds M, Pinto F, Khan MA, Bhat MA. 2019.** High-throughput phenotyping for crop
680 improvement in the genomics era. *Plant Sci* **282**: 60-72.

681 **Mishra P, Asaari MSM, Herrero-Langreo A, Lohumi S, Diezma B, Scheunders P. 2017.** Close range
682 hyperspectral imaging of plants: A review. *Biosystems Engineering* **164**: 49-67.

683 **Murchie EH, Lawson T. 2013.** Chlorophyll fluorescence analysis: a guide to good practice and
684 understanding some new applications. *Journal of Experimental Botany* **64**(13): 3983-3998.

685 **Pearson RL, Miller LD 1972.** Remote mapping of standing crop biomass for estimation of the
686 productivity of the shortgrass prairie, Pawnee National Grasslands, Colorado.

687 **Piaskowski JL, Brown D, Campbell KG. 2016.** Near-Infrared Calibration of Soluble Stem
688 Carbohydrates for Predicting Drought Tolerance in Spring Wheat. *Agronomy Journal* **108**(1):
689 285-293.

690 **Piatetsky-Shapiro G, Bosch XE, Jung T 2000.** High-Dimensional Data Analysis : The Curses and
691 Blessings of Dimensionality.

692 **Rao NR, Garg PK, Ghosh SK. 2007.** Development of an agricultural crops spectral library and
693 classification of crops at cultivar level using hyperspectral data. *Precision Agriculture* **8**(4):
694 173-185.

695 **Richardson AD, Duigan SP, Berlyn GP. 2002.** An evaluation of noninvasive methods to estimate
696 foliar chlorophyll content. *New Phytologist* **153**(1): 185-194.

697 **Silleos NG, Alexandridis TK, Gitas IZ, Perakis K. 2006.** Vegetation Indices: Advances Made in Biomass
698 Estimation and Vegetation Monitoring in the Last 30 Years. *Geocarto International* **21**(4): 21-
699 28.

700 **Sun W, Du Q. 2019.** Hyperspectral Band Selection: A Review. *IEEE Geoscience and Remote Sensing*
701 *Magazine* **7**(2): 118-139.

702 **Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M. 2017.** Plant Phenomics, From Sensors to
703 Knowledge. *Curr Biol* **27**(15): R770-r783.

704 **Thenkabail PS, Enclona EA, Ashton MS, Van Der Meer B. 2004.** Accuracy assessments of
705 hyperspectral waveband performance for vegetation analysis applications. *Remote Sensing*
706 *of Environment* **91**(3): 354-376.

707 **Winston PH. 1992.** *Artificial intelligence (3rd ed.)*: Addison-Wesley Longman Publishing Co., Inc.

708 **Wold S, Sjöström M, Eriksson L. 2001.** PLS-regression: a basic tool of chemometrics. *Chemometrics*
709 *and Intelligent Laboratory Systems* **58**(2): 109-130.

710 **Wu C, Niu Z, Tang Q, Huang W. 2008.** Estimating chlorophyll content from hyperspectral vegetation
711 indices: Modeling and validation. *Agricultural and Forest Meteorology* **148**(8): 1230-1241.

712 **Wu J, Rogers A, Albert LP, Ely K, Prohaska N, Wolfe BT, Oliveira Jr RC, Saleska SR, Serbin SP. 2019.**
713 Leaf reflectance spectroscopy captures variation in carboxylation capacity across species,
714 canopy environment and leaf age in lowland moist tropical forests. *New Phytologist* **224**(2):
715 663-674.

716 **Xu M, Liu R, Chen JM, Liu Y, Shang R, Ju W, Wu C, Huang W. 2019.** Retrieving leaf chlorophyll
717 content using a matrix-based vegetation index combination approach. *Remote Sensing of*
718 *Environment* **224**: 60-73.

719 **Xue J, Su B. 2017.** Significant Remote Sensing Vegetation Indices: A Review of Developments and
720 Applications. *Journal of Sensors* **2017**: 1353691.

721 **Yao X, Liu Y 1997.** Fast evolution strategies. Berlin, Heidelberg: Springer Berlin Heidelberg. 149-161.

722 **Yu T, Zhu H. 2020.** Hyper-Parameter Optimization: A Review of Algorithms and Applications. *ArXiv*
723 **abs/2003.05689.**

724 **Zhou R, Kjær KH, Rosenqvist E, Yu X, Wu Z, Ottosen C-O. 2017.** Physiological Response to Heat
725 Stress During Seedling and Anthesis Stage in Tomato Genotypes Differing in Heat Tolerance.
726 *Journal of Agronomy and Crop Science* **203**(1): 68-80.

727