frontiers in Plant Science

ORIGINAL RESEARCH
published: 12 January 2021
doi: 10.3389/fpls.2020.624273

Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean

*Mohsen Yoosefzadeh-Najafabadi[1], Hugh J. Earl[1], Dan Tulpan[2], John Sulik[1] and Milad Eskandari[1]**

[1] Department of Plant Agriculture, University of Guelph, Guelph, ON, Canada, [2] Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada

Recent substantial advances in high-throughput field phenotyping have provided plant breeders with affordable and efficient tools for evaluating a large number of genotypes for important agronomic traits at early growth stages. Nevertheless, the implementation of large datasets generated by high-throughput phenotyping tools such as hyperspectral reflectance in cultivar development programs is still challenging due to the essential need for intensive knowledge in computational and statistical analyses. In this study, the robustness of three common machine learning (ML) algorithms, multilayer perceptron (MLP), support vector machine (SVM), and random forest (RF), were evaluated for predicting soybean (*Glycine max*) seed yield using hyperspectral reflectance. For this aim, the hyperspectral reflectance data for the whole spectra ranged from 395 to 1005 nm, which were collected at the R4 and R5 growth stages on 250 soybean genotypes grown in four environments. The recursive feature elimination (RFE) approach was performed to reduce the dimensionality of the hyperspectral reflectance data and select variables with the largest importance values. The results indicated that R5 is more informative stage for measuring hyperspectral reflectance to predict seed yields. The 395 nm reflectance band was also identified as the high ranked band in predicting the soybean seed yield. By considering either full or selected variables as the input variables, the ML algorithms were evaluated individually and combined-version using the ensemble–stacking (E–S) method to predict the soybean yield. The RF algorithm had the highest performance with a value of 84% yield classification accuracy among all the individual tested algorithms. Therefore, by selecting RF as the metaClassifier for E–S method, the prediction accuracy increased to 0.93, using all variables, and 0.87, using selected variables showing the success of using E–S as one of the ensemble techniques. This study demonstrated that soybean breeders could implement E–S algorithm using either the full or selected spectra reflectance to select the high-yielding soybean genotypes, among a large number of genotypes, at early growth stages.

Keywords: artificial intelligence, data-driven model, ensemble methods, high-throughput phenotyping, random forest, recursive feature elimination

OPEN ACCESS

Edited by:
Yiannis Ampatzidis, University of Florida, United States

Reviewed by:
Omar Vergara-Diaz, University of Barcelona, Spain
Michael Gomez Selvaraj, Consultative Group on International Agricultural Research (CGIAR), United States

****Correspondence:***
Milad Eskandari
meskanda@uoguelph.ca

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *31 October 2020*
Accepted: *10 December 2020*
Published: *12 January 2021*

Citation:
Yoosefzadeh-Najafabadi M, Earl HJ, Tulpan D, Sulik J and Eskandari M (2021) Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean. Front. Plant Sci. 11:624273. doi: 10.3389/fpls.2020.624273

INTRODUCTION

The world population is projected to exceed nine billion individuals by 2050, which will require significant improvements in the yield of major crops that contribute to global food security (Tilman et al., 2009; Foley et al., 2011; Alexandratos and Bruinsma, 2012; Dubey et al., 2019). Increasing the yield is the primary goal of most plant breeding programs for major crops, such as soybean (*Glycine max*), which is the world's most widely grown leguminous crop and an important source of protein and oil for food and feed (Hartman et al., 2011). In the area of plant breeding, however, measuring primary traits, such as yield, which is under influenced by a combination of quantitative and qualitative traits, in large breeding populations consisting of several thousand genotypes is time and labor-consuming (Araus and Cairns, 2014; Cai et al., 2016; Xiong et al., 2018). Breeding for yield is known as a highly complex and non-linear process due to the genetic and environmental factors (Collins et al., 2008). Therefore, breeding approaches that are established based on secondary traits (e.g., yield component traits and reflectance bands), which are strongly correlated with the primary trait, enable plant breeders to efficiently recognize promising lines at early growth stages (Ma et al., 2001; Jin et al., 2010; Montesinos-López et al., 2017).

The combination of high-throughput genotyping and phenotyping technologies have enabled plant breeders to make their early growth stage selections more accurate while it reduced the evaluation time and cost in their breeding programs (Rutkoski et al., 2016). Although there has been significant progress in high-throughput genotyping in recent years with a direct impact on current plant breeding challenges (Araus and Cairns, 2014; Tardieu et al., 2017; Araus et al., 2018), acquisition of high-throughput field phenotyping is still a bottleneck in most breeding programs (Furbank and Tester, 2011; Araus et al., 2018).

Remote sensing of spectral reflectance is considered as an efficient high-throughput phenotyping tool (Araus and Cairns, 2014; Tardieu et al., 2017), which aims to measure the spectral reflectance efficiently at several plant growth and development stages in large breeding populations (Rutkoski et al., 2016). It is well documented that the spectral properties are genotype-specific and influenced by the anatomy, morphology, and physiology of plants (Kycko et al., 2018; Schweiger et al., 2018) and, therefore, can be used for screening plant genotypes with different agronomic potentials.

Analyzing large datasets consisting of spectral reflectance data requires intensive computational and statistical analyses, which is still challenging in many plant breeding programs (Lopez-Cruz et al., 2020). Nowadays, machine learning algorithms have drawn attention from researchers to develop model-based breeding methods that can improve the efficiency of breeding processes (Hesami et al., 2020a). Recently, one of the most common artificial neural networks (ANNs), the multilayer perceptron (MLP) developed by Pal and Mitra (1992), has been broadly used for modeling and predicting complex traits, such as yield, in different breeding programs (Geetha, 2020). MLP can be considered as a non-linear computational method employed for various tasks such as classification and regression of complex systems (Chen and Wang, 2020; Hesami et al., 2020b). This algorithm is able to detect the connection and relationship between the input and output (target) variables and quantify the inherent knowledge existing in the datasets (Ghorbani et al., 2016; Hesami et al., 2020b). This algorithm includes several highly interconnected processing neurons that can be used in parallel to detect a solution for a specific problem (Ghorbani et al., 2016; Geetha, 2020). Support vector machines (SVMs), developed by Vapnik (2000), are known as one of the powerful and easy to use machine learning algorithms that can recognize patterns and behavior of non-linear relationships (Auria and Moro, 2008; Su et al., 2017). Some of the advantages of SVMs over MLP are linked to the complexity of the networks. SVMs usually use a large number of learning problem formulations leading to solving a quadratic optimization problem (Feng et al., 2020; Hesami et al., 2020b). In theory, SVM has to be better performance because of using structural risk minimization inductive principles rather than the empirical risk minimization inductive principle (Belayneh et al., 2014). In addition to MLP and SVM, random forest (RF) (Breiman, 2001) is another method for data modeling with a computational efficient training phase and very high generalization accuracy. RF has been broadly used in areas such as object recognition (Lepetit et al., 2005), skin detection (Khan et al., 2010), plant phenomics (Falk et al., 2020), and genomics (Mokry et al., 2013).

Machine learning algorithms are subject to overfitting, mainly because of limited training data and dependent on single predictive models (Ali et al., 2014; Feng et al., 2020). Ensemble techniques, in which a group of algorithms are exploited to combine all the possible predictions for the ultimate prediction used to address this shortage (Dietterich, 2000). By using ensemble models, the predictive performances were improved for yield prediction in Alfalfa (Feng et al., 2020), Nicosia wastewater treatment plant (Nourani et al., 2018), and plant lncRNAs (Simopoulos et al., 2018). Boosting, bagging, and stacking are three of the most commonly used ensemble models (Dietterich, 2000; Feng et al., 2020). The bagging method was first introduced by Breiman (1996) as a variance reduction approach for different algorithms such as decision trees or other algorithms that employed variable selection and fitting in a linear model (Galar et al., 2011). Boosting algorithms have been introduced by Schapire (1999) to serve as the alternative for the bagging method (Drucker and Cortes, 1996). Unlike bagging methods, which are parallel ensemble techniques, boosting methods are known as sequential ensemble techniques of base models by exploiting the dependencies of each algorithm (Dietterich, 2000; Feng et al., 2020). Many studies reported the successfulness of using bagging-RF and stochastic gradient boosting in predicting the yield of different crops (Pal, 2007; Gandhi et al., 2016; Aghighi et al., 2018; Zhang Z. et al., 2019). Bagging and boosting ensemble techniques commonly combine homogeneous algorithms for interpretation, while stacking methods tend to use heterogeneous algorithms and adjust the difference between them to increase precision (Dietterich, 2000; Zhou, 2012; Feng et al., 2020).

The successful use of machine learning algorithms for predicting the performance of different agronomic traits,

including yield, are reported in Alfalfa (Feng et al., 2020), *Senecio* species (Carvalho et al., 2013), grassland (Feilhauer et al., 2017; Rocha et al., 2018), and chrysanthemum (Hesami et al., 2019). However, the application of machine learning algorithms for predicting soybean yield from hyperspectral reflectance data is still unexplored and required comprehensive studies. Ensemble-based methods have successfully been applied to improve the prediction accuracies in artificial intelligence and computer vision (Ali et al., 2014; Feng et al., 2018, 2020; Ju et al., 2018) and, therefore, they may improve the accuracy of the yield prediction in this study. Thus, the main objectives of this study are: (1) to investigate the potential use of hyperspectral reflectance for predicting soybean yield, (2) to identify appropriate time-point of soybean growth stages for collecting hyperspectral reflectance to maximize yield prediction accuracy, and (3) to have a comparative study of individual and ensemble machine learning algorithms to improve the accuracy of predicting yield. The results of this study might help soybean breeders to increase the efficiency of selecting superior lines by estimating the final yield at early growth stages using spectral reflectance combined with machine learning approaches.

MATERIALS AND METHODS

Experimental Locations and Plant Materials

The research was conducted at the University of Guelph, Ridgetown Campus, in 2018 and 2019. A panel of 250 soybean genotypes was grown under field condition at two locations: Ridgetown (42°27′14.8″N 81°52′48.0″W, 200 m above sea level) and Palmyra (42°25′50.1″N 81°45′06.9″W, 195 m above sea level), in Ontario, Canada, during two consecutive growing seasons in 2018 and 2019 (**Figure 1**).

The soybean genotypes used in this study were the core germplasms of the soybean breeding program at the University of Guelph, Ridgetown, that have been collected in the past 35 years and used for genetic studies and cultivar developments. The experiments were conducted using randomized complete block designs (RCBD) with two replications in four environments (two locations × two years). Overall, there were 500 soybean plots per environment and 1000 soybean plots per year. In order to reduce the possible spatial variability in the field, each experiment was analyzed by nearest-neighbor analysis (NNA) as one of the error control strategies by using double covariate analysis (Stroup and Mulitze, 1991; Bowley, 1999; Katsileros et al., 2015). Each plot consisted of five rows, each 4.2 m long with a row spacing of 43 cm. The seeding rate was 57 seeds/m^2. At the end of the season, the three inside rows were machine harvested for estimating total seed yields (Ton ha^{-1}).

Phenotypic Evaluations

Yield Collection

Soybean seed yield (Ton ha^{-1}) was measured using three out of five harvested rows for each plot and adjusted to a 13% moisture level. The best linear unbiased prediction (BLUP) as a mixed model was used to calculate the average seed yield production for each soybean genotype across different environments (Goldberger, 1962).

Hyperspectral Reflectance Data Collection

In this study, the focus was on the spectral reflectance bands that are typically classified as the visible (VIS) and near-infrared (NIR) spectral components (Albetis et al., 2017). Canopy hyperspectral reflectance measurements were collected during the soybean growth and development stages at R4, where pods are 1.91 cm long at one of the four uppermost nodes, and R5, where seeds are 0.31 cm long in pods at one of the four uppermost nodes (Pedersen et al., 2004).

Each soybean genotype's hyperspectral reflectance properties were collected using a UniSpec-DC Spectral Analysis System (PP Systems International, Inc., 110 Haverhill Road, Suite 301 Amesbury, MA, United States). The machine covers 250 reflectance bands between 350 nm and 1,100 nm with a bandwidth of 3 nm. The field-of-view of the sensor was approximately 25° and covered a sample area of 0.25 m^2. Dark reference was used for calibrating the dual channels, and Spectralon panels were used to characterize incoming solar radiation. For each plot, three measurements were recorded, and their average, calculated by the BLUP model, was used as the reflectance band datapoint. All of the measurements were performed as close to solar noon as possible—the data for each stage were collected in 1 day, from 11:00 AM to 2:00 PM, to minimize the signal noise associated with the environment.

Data Pre-processing and Statistical Analyses

The existence of noise during hyperspectral reflectance measurement is inevitable, typically caused by sensors and electronic fluctuations (Ozaki et al., 2006). Therefore, it would be critical to have a pre-processing step for the collected data in order to increase the accuracy of the study. The hyperspectral data and yield of 250 soybean genotypes were pre-processed using the R software (version 3.6.1) to remove potential noises that randomly occur across the whole spectra resulting in misinterpretations. After checking the quality of reflectance bands and detecting outliers by using principal component analysis (PCA) for each genotype, 245 genotypes were selected for further analyses (Serneels and Verdonck, 2008). As a result of sensor-specific artifacts, reflectance bands at the two edges of the hyperspectral reflectance spectrum, 350–395 nm and 1,005–1,100 nm, were removed from the original data. The collected contiguous hyperspectral reflectance data was also reduced from 395 to 1005 nm with a 3 nm interval to a 10 nm interval leading to a total of 62 variables. Data scaling and centering were applied in order to improve reflectance properties in the pre-processing and the pre-treatment steps (Rossel, 2008). For each reflectance band variable, the Savitzky–Golay filter was applied for improving the signal-to-noise ratio (Savitzky and Golay, 1964).

As shown in **Figure 2**, the measured soybean yield was divided into four classes with equal numbers (∼) of data points in ascending order: Low (0–24.99% of total yield), medium-low (25–49.99% of total yield), medium-high (50–74.99% of total yield), and high (75–100% of total yield) yield. While

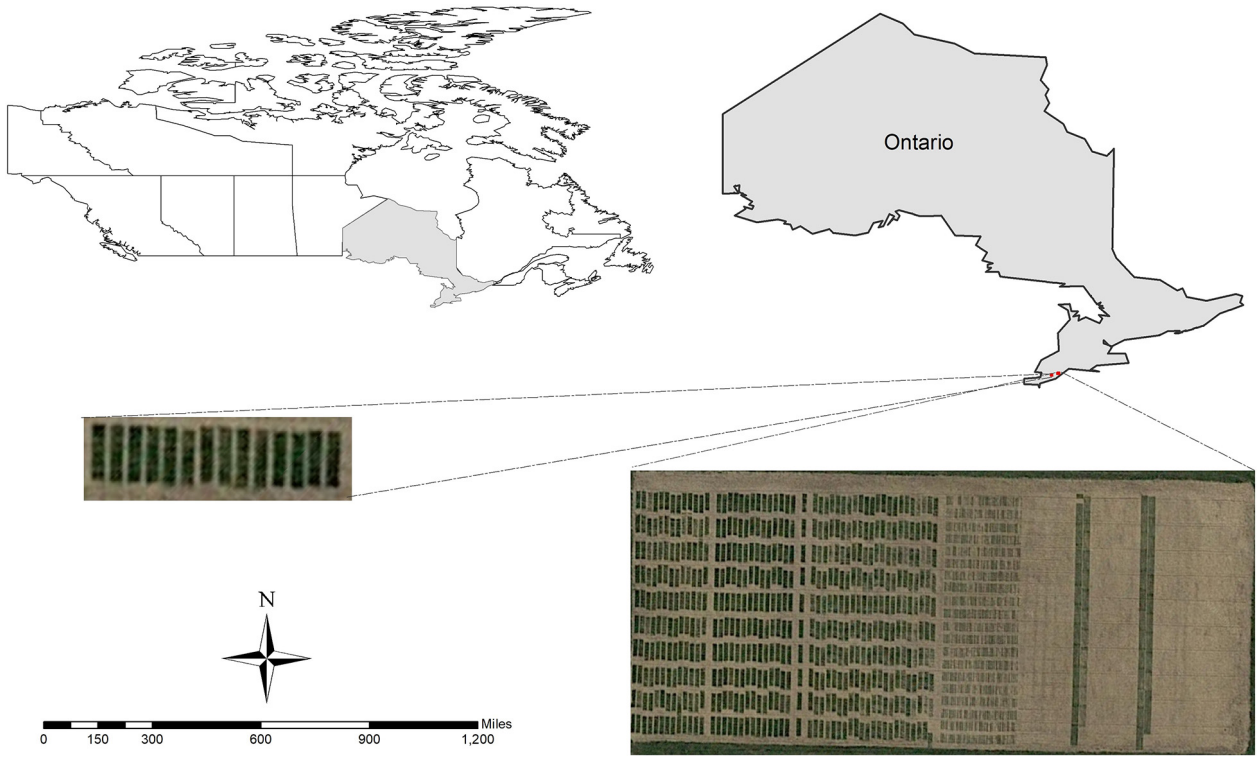

FIGURE 1 | The location of the experiments in 2018 and 2019.

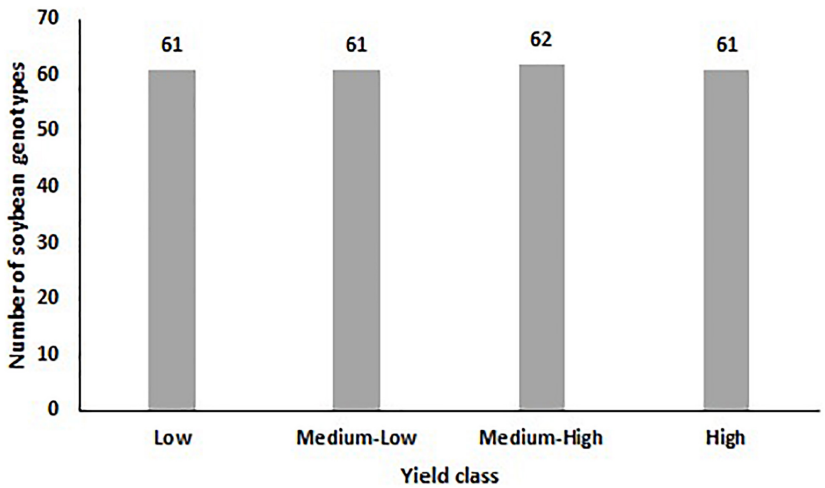

FIGURE 2 | The distribution of soybean genotypes in each yield class.

62 reflectance bands were considered as input variables, the classified yield was chosen as the output variable.

Variable Selection

Feature or variable selection is usually applied before developing the machine learning algorithms for reducing the data dimensionality, specifically in the small training datasets. One of the common approaches for variable selection is the recursive feature elimination (RFE) approach, which is easy to configure and effectively select the most relevant variables that predict the output (Chen and Jeong, 2007). Therefore, the RFE was run to indicate the initial variable importance scores and eliminate the reflectance band variables with the lowest importance score. Afterward, the process was recursively repeated until the ranking was indicated for all the reflectance bands. The package

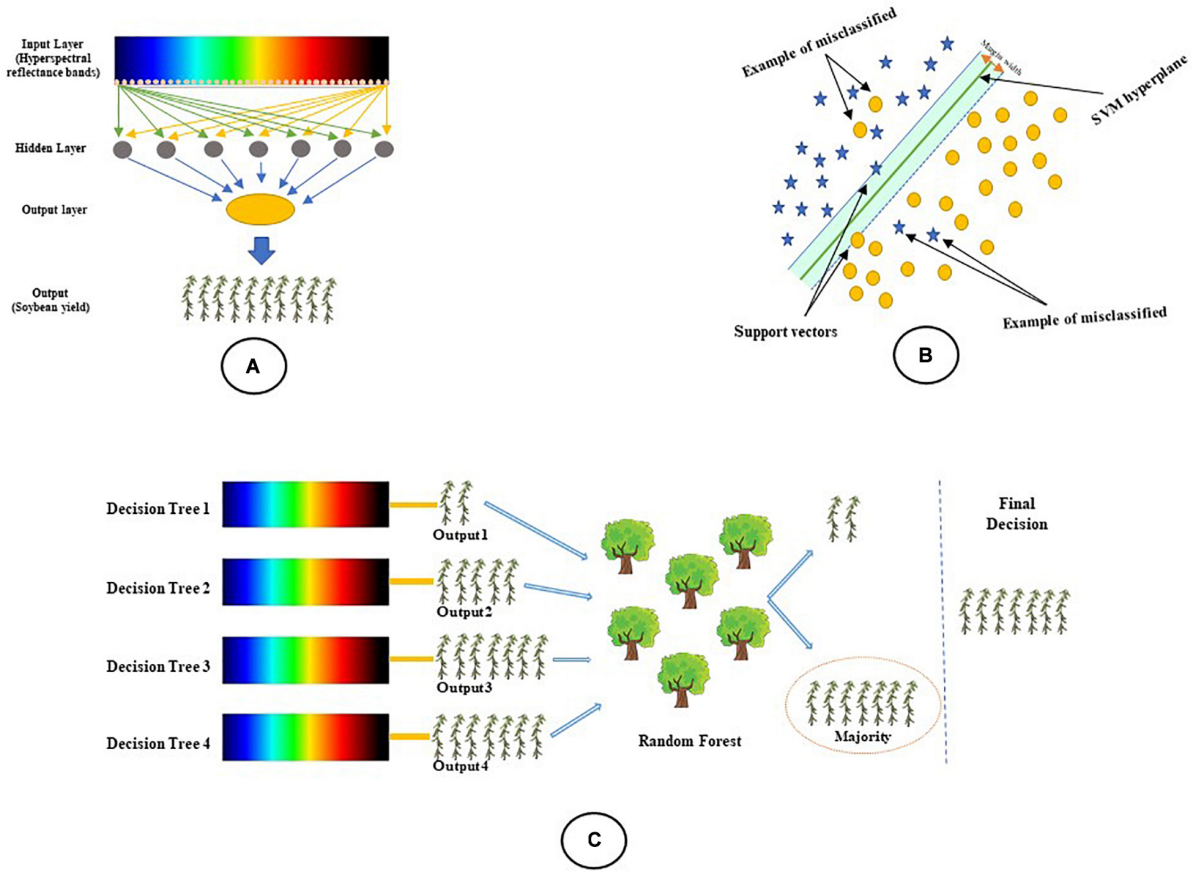

FIGURE 3 | A schematic representation of the machine learning algorithms used in this study to classify the soybean yield using reflectance bands: **(A)** Multilayer perceptron, **(B)** Support vector machine, and **(C)** Random forest.

caret (Kuhn, 2008) in R software version 3.6.1 was used for running RFE.

Data-Driven Modeling

Three of the most commonly used algorithms in the literature, multilayer perceptron (MLP), the support vector machine (SVM), and random forest (RF) (Filippi and Jensen, 2006; Chen et al., 2007; De Castro et al., 2012; Makantasis et al., 2015; Zhang N. et al., 2019; Šestak et al., 2019), were chosen and used for predicting the soybean yield. **Figure 3A** shows the MLP algorithm including an input layer, one or more hidden layers, and an output layer of completely interconnected neurons. Each neuron unit produces an output based on a sigmoid function derived from a linear combination of outputs from a previous layer (Wang et al., 2009). SVM (**Figure 3B**) is a set of related supervised learning methods that can recognize patterns used for classification analyses (Suykens and Vandewalle, 1999; Shao et al., 2012). The objective of SVM is to use hyperplanes for determining the optimal separation of yield classes. The random forest (RF) approach generates a series of trees representing a subset *n* of independent observations (**Figure 3C**). A detailed description of these machine learning algorithms can be found in Taillardat et al. (2016) and Meinshausen (2006). All of the relevant parameters in each machine learning algorithm were optimized based on the input variables.

We employed an ensemble method based on a stacking strategy (E–S) to improve the prediction performance. The results from individual algorithms were collected and combined together via the stacking procedure described in Dietterich (2000), where an algorithm with the highest accuracy performance was selected as the metaClassifier for this ensemble model. The Weka software version 3.9.4 (Hall et al., 2009) was used for running all machine learning algorithms and the ensemble method.

Quantification of Machine Learning Performance

In this study, the fivefold cross-validation strategy (Siegmann and Jarmer, 2015) with 10 repetitions was used to measure the classification quality of all the tested ML algorithms (**Figure 4**). In order to evaluate each algorithm, the values of precision (Eq. 1), recall (Eq. 2) as a measure of sensitivity, F-measure (Eq. 3), and Matthews correlation coefficient (MCC, Eq. 4) for validation

dataset were measured using the following formulas:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

where *TP* stands for the number of true positives, *TN* is the number of true negatives, *FP* stands for the number of false positives, and *FN* is the number of false negatives.

Visualizing and Analyzing

The Microsoft Excel software (2016), *ggplot2* (Wickham, 2011), and *ggvis* (Dennis, 2016) packages in the R software version 3.6.1 were used to conduct statistical analyses and visualize the results.

RESULTS

Yield Statistics and Spectral Profiles

In the current study, the average yield of 245 soybean genotypes, evaluated in four environments, ranged from 2.58 to 5.71 ton ha^{-1} with a mean and standard deviation of 4.22 and 0.57 ton ha^{-1}, respectively. The minimum, mean, and maximum values of each reflectance bands evaluated for all the genotypes across the four environments at the R4 and R5 growth stages are reported in **Figure 5**. At both growth stages, while the reflectance values showed small ranges of variation among the genotypes between 395 and 695 nm, the bands greater than 705 nm showed large variations within the population.

Variable Selection

The importance values of all 62 reflectance band variables for predicting yield were estimated using the RFE strategy for the R4 (**Table 1**) and R5 (**Table 2**) growth stages. For the R4 growth stage, the 1005 nm and the 605 nm bands had the highest and lowest importance values (%) for classifying the soybean yield, respectively. Based on RFE analysis, 56 of the reflectance bands were selected for training the algorithm, as shown in **Figure 6A**. At the R5 growth stage, the highest and lowest importance values (%) were found at 395 nm and the 725 nm bands, respectively. Out of 62 reflectance bands, 21 reflectance bands were selected to train the algorithms based on RFE strategy, which were considered selected variables (-VS) for further analyses. Among the 21 selected reflectance bands, three bands were in the violet, six in the blue, two in the green, eight in the red, and two were in the near-infrared (NIR) regions of the spectrum (**Figure 6B**). By using RFE for the R4 growth stage dataset, the top five high importance reflectance bands were located in the violet and NIR regions of the spectra. However, for R5, the violet and red regions had the top five high importance reflectance bands (**Tables 1**, **2**). The violet region, specifically the 395 nm band, had the highest importance values in both growth stages. The plotting of the soybean yield versus reflectance values at 395 nm (R5 growth stage) illustrated that the values for 395 nm in the high yielding class ranged from 0.009 to 0.016 which lower than values for the low yielding class, ranged from 0.020 to 0.029 (**Figure 7**). The difference between the reflectance values of high and low yielding classes was statistically significant at the significance level of 0.05 (data were not shown). Among all the tested bands, the 395 nm band measured at R5 was considered as the best solo reflectance band for discriminating soybeans for their yield potential.

Growth Stage Comparison

In order to investigate which of the growth stages is better for collecting reflectance data and predicting the soybean yield, the reflectance bands collected at each soybean growth stage were analyzed using the three machine learning algorithms. The average classification accuracy for validation dataset ranged between 12 and 43% using the R4 data and between 12 and 99% using the R5 data, which indicated that the R5 soybean growth stage is, in general, a better stage for collecting reflectance data if the goal is to predict the yield (**Figure 8**). Therefore, R5 was selected for further machine learning algorithm analyses. The results of individual and ensemble machine learning algorithms using R4 data are available in **Figure 8** and **Supplementary Figure 1**.

Comparative Analysis of the Developed Models

All three algorithms (RF, MLP, and SVM), as well as the E–S model, were trained using both full (62 bands) and selected (21 bands) variables at R5, and the summaries of the confusion matrices were presented in **Supplementary Table 1**. Regarding the comparative analyses of individual algorithms using all variables, RF, MLP, and SVM had the highest to lowest MCC values equal to 0.84, 0.76, and 0.66, respectively (**Figure 9A**). For the selected variables, the MCC values for RF and MLP declined to 0.80 and 0.73, respectively, while the value for SVM slightly increased to 0.73. The E–S method outperformed all the individual algorithms obtaining an MCC value of 0.93, using all variables, and 0.87, using selected variables (**Figure 9A**). In general, among all the individual tested algorithms, the RF algorithm had the highest performance with the values of 84 and 80% yield classification accuracy using all variables and selected variables, respectively.

When using full variables, the precision values for RF, MLP, and SVM were 0.91, 0.83, and 0.82, respectively. However, by using selected variables, the precision values for RF and MLP declined to 0.87 and 0.78, respectively, while the SVM performance was improved (0.87) when compared against all variables (**Figure 9B**). The E–S model had a precision of 0.96 using all variables and 0.90 using selected variables. Using all variables, the highest recall value was obtained for RF with a value of 0.84, followed by MLP and SVM with the values of 0.83 and 0.68, respectively.

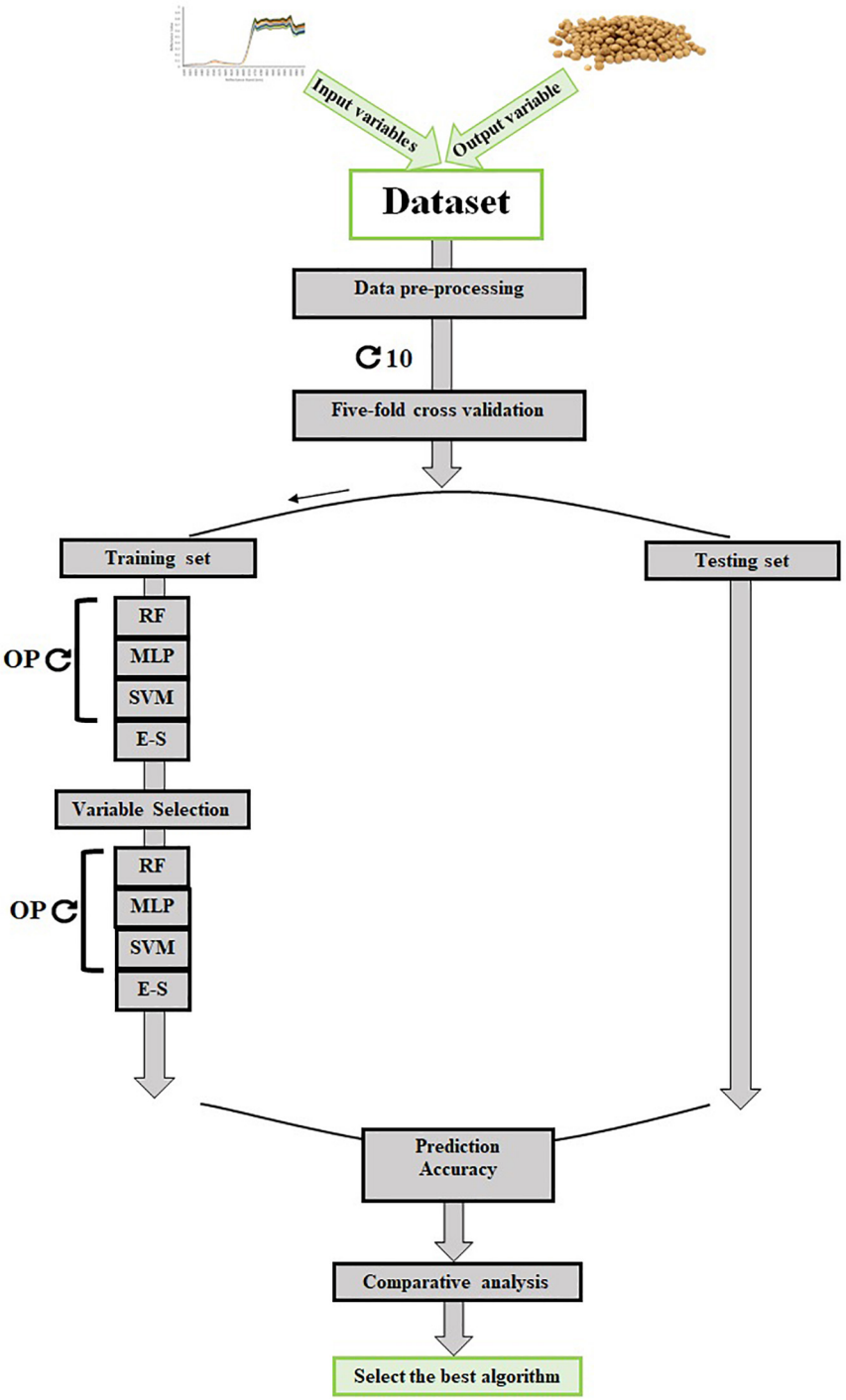

FIGURE 4 | The scheme of data collection and machine learning algorithm development and validation. OP, optimizing parameters; MLP, multilayer perceptron; SVM, support vector machine; RF, random forest; E–S, ensemble–stacking strategy.

However, the recall value of SVM increased to 0.72 using selected variables. The recall values of RF and MLP declined when selected variables were used (**Figure 9C**). E–S had the highest recall values, with 0.94 and 0.90 for full and selected variables, respectively.

To have a better interpretation of precision and recall values, the F-measure was evaluated for each and every algorithm. Using all variables, the F-measures of RF, MLP, and SVM were estimated to be 0.87, 0.81, and 0.71, respectively (**Figure 9D**). F-measure values were decreased for RF (0.84)

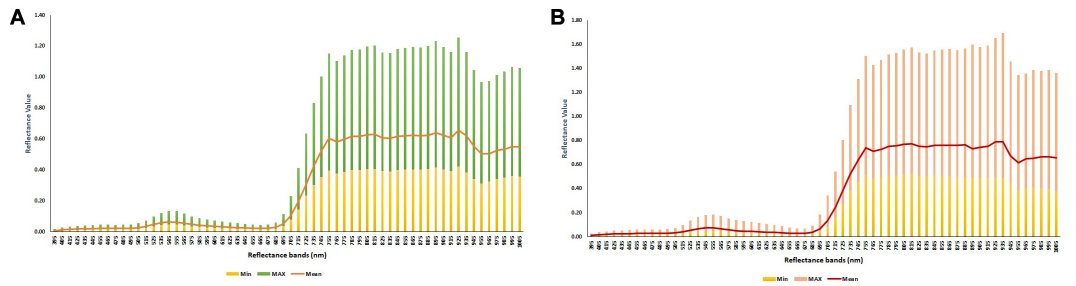

FIGURE 5 | The minimum, mean, and maximum values of each reflectance band were measured for 245 soybean genotypes evaluated at **(A)** R4 and **(B)** R5 growth stages at four different field environments.

TABLE 1 | Reflectance band ranking using the recursive feature elimination (RFE) strategy at R4 soybean growth stage.

Reflectance band (nm)	Ranking	Reflectance band (nm)	Ranking
1005	1	775	32
395	2	655	33
945	3	675	34
755	4	665	35
985	5	825	36
995	6	485	37
705	7	695	38
745	8	475	39
965	9	615	40
955	10	465	41
715	11	515	42
975	12	625	43
905	13	685	44
725	14	565	45
885	15	575	46
875	16	535	47
895	17	645	48
765	18	545	49
845	19	635	50
915	20	555	51
865	21	415	52
735	22	505	53
925	23	595	54
855	24	455	55
805	25	585	56
795	26	445	57
935	27	425	58
835	28	435	59
815	29	525	60
405	30	495	61
785	31	605	62

TABLE 2 | Reflectance band ranking using the recursive feature elimination (RFE) strategy at R5 soybean growth stage.

Reflectance band (nm)	Ranking	Reflectance band (nm)	Ranking
395	1	965	32
665	2	845	33
675	3	865	34
655	4	905	35
405	5	915	36
685	6	515	37
645	7	695	38
435	8	885	39
445	9	875	40
635	10	895	41
475	11	585	42
485	12	925	43
495	13	935	44
415	14	575	45
625	15	955	46
425	16	715	47
455	17	755	48
465	18	535	49
765	19	975	50
615	20	555	51
775	21	985	52
795	22	995	53
815	23	525	54
805	24	545	55
785	25	565	56
505	26	945	57
605	27	705	58
825	28	1005	59
855	29	745	60
595	30	735	61
835	31	725	62

and MLP (0.80) using selected variables. However, for SVM algorithm, the F-measure value was increased to 0.77 using selected variables. The E–S algorithm overperformed all the individual machine learning algorithms by having an F-measure value of 0.94, using all variables, and 0.90, using selected variables.

DISCUSSION

One of the objectives of this study was to find the best growth stage for collecting reflectance data suitable for predicting soybean yields. In this study, the hyperspectral reflectance data were collected at the reproductive stages of R4 and R5, in which

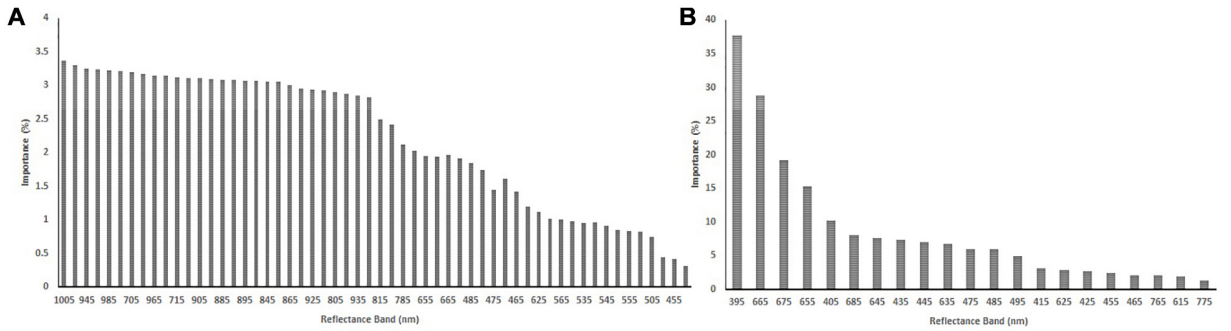

FIGURE 6 | The importance of selected variables based on the recursive feature elimination (RFE) strategy for soybean reflectance bands measured at R4 **(A)** and R5 **(B)** soybean growth stages.

pods and seeds are developed. R4 and R5 are known as critical growth and development stages in soybean since stresses can impose significant impacts on the yield at these stages and, therefore, soybean genotypes with different levels of tolerance to stresses can be discriminated from one another at these stages (Sweeney et al., 2003). For example, the results of a study by Eck (1987) showed that imposing soybeans to water deficit stress during the R1 to R3 growth stages reduce seed yields up to 9-13%. However, imposing the same soybeans to water deficit stress during R4 to R5 reduced seed yields up to 46%. Water deficit stress can less influence the total seed yield when occurring anytime beyond the R5 growth and development stage. Therefore, measuring hyperspectral reflectance at R4 and R5 would be more informative for predicting the soybean yield classes since the final yield production has been to some extent established at these two stages for all the genotypes. Our results indicated that R5 is a better stage to measure reflectance bands for predicting the yield. Ma et al. (1996) reported significant correlations between leaf photosynthetic rates and leaf greenness at R4 and R5, while this correlation was not significant at R6. In Soybean, the leaf photosynthetic rate can be changed significantly during the growth stages that, in turn, can empower different genotypes to recover the yield losses caused by temporary environmental stresses (Ferris et al., 1998; Siebers et al., 2015). Studies showed that environmental stresses at R5 can damage the soybean yield greater than that at R4 (Fehr et al., 1981) since the plants have less time to recover for yield before physiological maturity. It can be hypothesized that predicting yield of genotypes with different genetic potential by using reflectance bands that are measured at R5 is more reliable since the final yield productions have already been established, to some extent, for all the genotypes. The current study showed that the reflectance bands collected at R5 are more reliable and informative for predicting yield than the data collected at R4.

Several studies reported the strong correlation between reflectance bands and yield in different crop plants such as alfalfa (Kayad et al., 2016; Feng et al., 2020), wheat (Prey and Schmidhalter, 2019), maize (Lane et al., 2020), rice (Wan et al., 2020), and sugarcane (Verma et al., 2020). The visible reflectance bands can be splitted into three main regions, red (650–700 nm), green (495–570 nm), and violet–blue (390–495 nm) (Hennessy et al., 2020). Most studies were emphasized the importance of red spectral bands or the combined use of red and red edge bands as one solid index in predicting the total yield (Jolly et al., 2005; Filippa et al., 2018; Lykhovyd, 2020; Phan et al., 2020; Tiwari and Shukla, 2020). In this study, we identified highly ranked bands in the violet and red regions for classifying the soybean seed yield, centered at 395 nm, 665 nm, and 675 nm (**Table 2**). The violet and red spectral regions can be associated with the absorption of plant pigments such as carotenoid, anthocyanins, and chlorophyll (Merzlyak et al., 2003; Richter et al., 2016; Hennessy et al., 2020). Carotenoid plays a pivotal role in discrimination of senescent leaves (Richter et al., 2016; Hennessy et al., 2020). The importance of 395 nm band in soybean yield prediction at R5 growth stage can refer to the fact that soybean at R5 growth stage initiates the senescence and decay of chlorophyll resulting in better discrimination of the genotypes with different photosystems functioning and photoprotection capabilities. However, there is no report on the solid effect of the 395 nm reflectance band in the physiological process of soybean or any other plants.

In order to have accurate yield prediction and avoid model overfitting, machine learning algorithms may benefit from using a variable selection process to reduce the dimensionality of the data to an appropriate level (Hennessy et al., 2020). Existing variable selection methods can be categorized based on their applications, complexities, and accuracy (Zheng et al., 2020). One of the most commonly used variable selection methods is the RFE approach that provides an acceptable performance with moderate computational exertions (Guyon et al., 2002; Granitto et al., 2006). The successful use of RFE to reduce the number of input variables has been reported in many studies (Granitto et al., 2006; Chen and Jeong, 2007; Feng et al., 2020). The efficiency of using selected variables for predicting classified soybean yield over full reflectance band variables was evaluated using the RFE method. Using RFE method might decrease the value of the parameters such as precision, recall, MCC, and F-measure to avoid overfitting (Loughrey and Cunningham, 2004). This is a small price to pay, especially if the decrease in performance is not significant. Among all the tested individual machine learning algorithms, RF had the highest performance using either full or

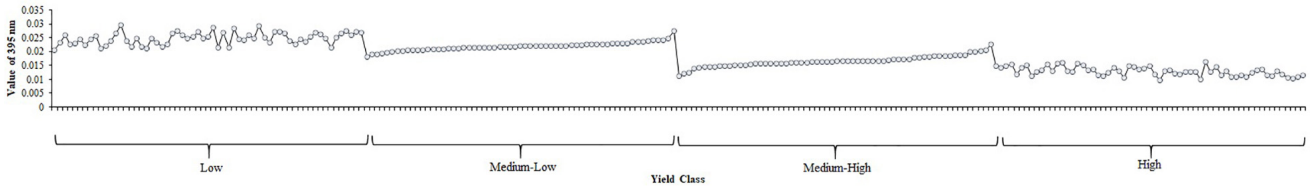

FIGURE 7 | The soybean yield classes versus the 395 nm reflectance band at R5 growth stages.

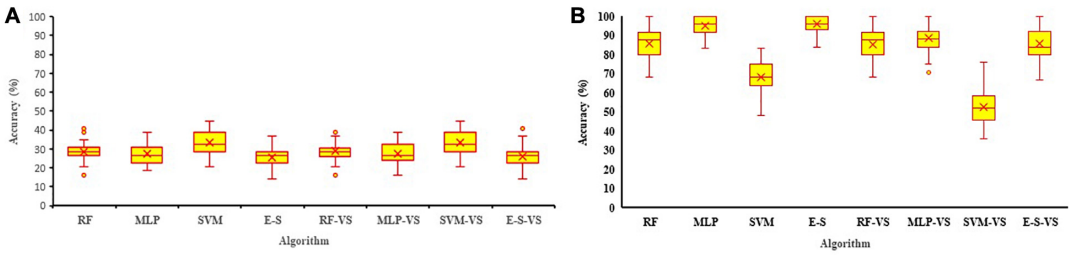

FIGURE 8 | The accuracy of RF, MLP, SVM, and E–S algorithms for predicting soybean yield using full and RFE selected variables (-VS) measured at R4 **(A)** and R5 **(B)** soybean growth stages in four environments. The mean performance was shown as × in each figure. MLP, multilayer perceptron; SVM, support vector machine; RF, random forest; E–S, ensemble–stacking strategy; RFE, recursive feature elimination.

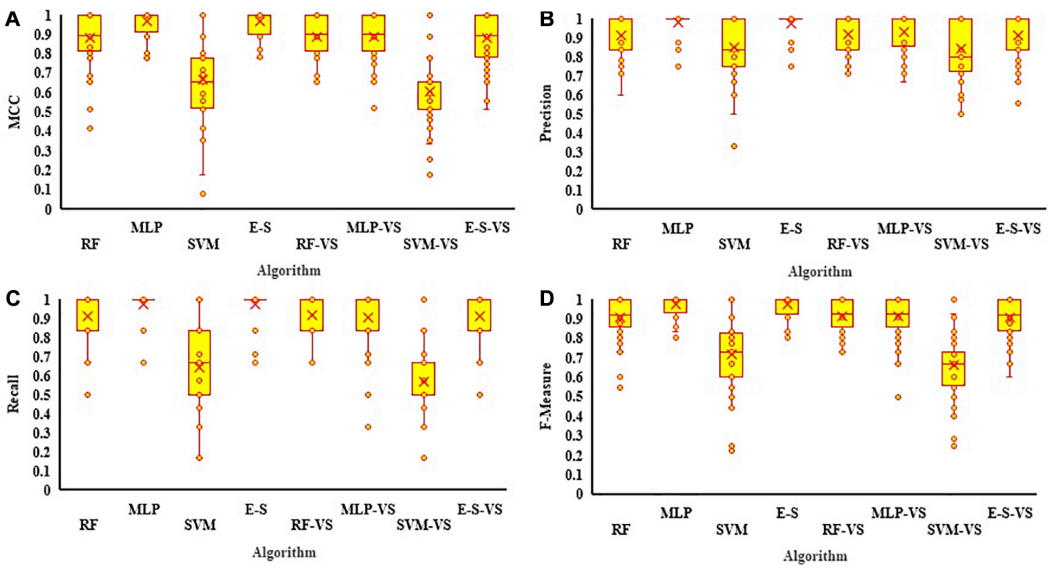

FIGURE 9 | The estimate values of **(A)** Matthews correlation coefficient (MCC), **(B)** Precision, **(C)** Recall, and **(D)** F-Measure for RF, MLP, SVM, and the E–S algorithms used for predicting soybean yield from all and selected variables collected (-VS) at the R5 growth stage. The mean performance is indicated with × in each figure. MLP, multilayer perceptron; SVM, support vector machine; RF, random forest; E–S, ensemble–stacking strategy; RFE, recursive feature elimination.

selected variables. This high performance may come from the nature of the RF algorithm, in which trees are trained using multiple random subsamples of the original dataset. This feature gives RF this ability to generate better and more stable predictions for new instances not necessarily included in the training dataset (Liaw and Wiener, 2002).

Multilayer perceptron was another machine learning algorithm that was exploited in this study. MLP was previously applied in different areas such as weed science (Tamouridou et al., 2017) or drought tolerance (Etminan et al., 2019), but not in soybean for yield prediction. This study found MLP to be the second-best machine learning algorithm for predicting the soybean yield. Previous studies reported a high likelihood of overfitting for neural network algorithms (Lawrence and Giles, 2000; Murakoshi, 2005). For MLP, common parameters such as the number of hidden layers, the

number of neurons in each layer, or training time can be used to control overfitting; however, the degree of overfitting would vary throughout the input variables (Lawrence and Giles, 2000).

Support vector machine is also one of the most common machine learning algorithms that have been broadly used in different areas such as plant tissue culture (Hesami and Jones, 2020), image classification (Lin et al., 2011), genes classification (Duan et al., 2005), and drug disambiguation (Björne et al., 2013). SVM is usually used when scientists have to deal with large numbers of features and high sparsity (Nguyen and De la Torre, 2010). Although the prediction accuracy of the SVM algorithm was lower than the values for RF and MLP in this study, its performance was slightly increased when the selected variables were used. An increase in SVM performance using selected variables was also reported in previous studies (Su and Yang, 2008; Tan et al., 2010; Alirezanejad et al., 2020). It might be due to this fact that selecting relevant variables can improve the performance of SVM through ameliorating its feature interpretability, computational efficiency, and generalization performance (Nguyen and De la Torre, 2010; Roy et al., 2015).

In order to see we can improve the prediction accuracy in this study through the combined use of the machine learning algorithms, RF, MLP, and SVM were used in constructing E–S, and RF was chosen as the metaClassifier for this ensemble algorithm. By using the E–S approach, we improved the prediction accuracy using either full or selected variables. A successful use of the E–S method has recently been reported for predicting the yield in alfalfa (Feng et al., 2020). When using the E–S approach, it is necessary to include self-sufficient, independent, and diverse ML algorithms in the analyses (Araya et al., 2017; Feng et al., 2020), which have a minimum dependency from one another and sufficient powers to predict the dependent variable, soybean yield classes in this study (Araya et al., 2017; Feng et al., 2020). The above criteria are important to be considered when individual ML algorithms are selected to combine in a given E–S analysis. In this study, RF, MLP, and SVM are selected as individual algorithms to be used in the E–S analyses because of their independent prediction methods as well as having different training approaches. By using the E–S approach, the prediction accuracy increased to 0.93, using all variables, and to 0.87, using selected variables, showing the success of using E–S as one of the ensemble techniques.

CONCLUSION

Pre-harvest soybean yield classifications and estimations are important for grain policy-making and food security across worldwide. Spectral reflectance is considered as an efficient phenotyping tool that can help breeders to make their selections at lower cost at a fast pace. The objectives of this study were to demonstrate the best soybean growth stage for measuring hyperspectral reflectance and evaluating the three most commonly used ML algorithms along with introducing the E–S method in predicting the soybean yield using reflectance variables. Soybean R5 growth stage was identified as the better stage than R4 for measuring hyperspectral reflectance. In addition to using 62 reflectance bands as the full variables, the RFE method was used to reduce the dimensionality of the data, and therefore, 21 most important bands were selected as the selected reflectance variables. Using both full and selected reflectance variables, RF overperformed all individual algorithms. Therefore, RF was selected as the metaClassifier for E–S. E–S had the highest prediction accuracy as one of the ensembles combined approaches compared to an individual ML algorithm. Therefore, E–S was recommended as a reliable and appropriate ML algorithm for predicting the soybean yield using reflectance variables. This study provides an applicable pipeline for using hyperspectral reflectance data and suitable ML algorithms for the development of high yielding soybeans, which can be used in large soybean breeding programs for selecting high-yielding soybeans at pre-harvesting stages. The developed methodology in this study can open a reliable and new window in using spectral reflectance for selecting high yielding genotypes in different crops.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

MY-N performed the experiments, modeled, summed up, and wrote the manuscript. HE, DT, and JS revised the manuscript. ME designed and lead the experiments and revised the manuscript. All authors have read and approved the final manuscript.

FUNDING

This project was funded in part by the Grain Farmers of Ontario (GFO) and SeCan. The funding bodies did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge the technical assistance of Dr. Sepideh Torabi, Mr. Bryan Stirling, Mr. John Kobler, Mr. Robert Brandt, and all the soybean breeding crew at the University of Guelph, Ridgetown Campus.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.624273/full#supplementary-material

REFERENCES

Aghighi, H., Azadbakht, M., Ashourloo, D., Shahrabi, H. S., and Radiom, S. (2018). Machine learning regression techniques for the silage maize yield prediction using time-series images of landsat 8 OLI. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* 11, 4563–4577. doi: 10.1109/JSTARS.2018.2823361

Albetis, J., Duthoit, S., Guttler, F., Jacquin, A., Goulard, M., Poilvé, H., et al. (2017). Detection of flavescence dorée grapevine disease using unmanned aerial vehicle (UAV) multispectral imagery. *Remote Sens.* 9:308. doi: 10.3390/rs9040308

Alexandratos, N., and Bruinsma, J. (2012). *World Agriculture Towards 2030/2050: the 2012 Revision*. Rome: Food and Agriculture Organization of the United Nations, Agricultural Development Economics Division (ESA).

Ali, I., Cawkwell, F., Green, S., and Dwyer, N. (2014). "Application of statistical and machine learning models for grassland yield estimation based on a hypertemporal satellite remote sensing time series," in *Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium*, (IEEE), 5060–5063. doi: 10.1109/IGARSS.2014.6947634

Alirezanejad, M., Enayatifar, R., Motameni, H., and Nematzadeh, H. (2020). Heuristic filter feature selection methods for medical datasets. *Genomics* 112, 1173–1181. doi: 10.1016/j.ygeno.2019.07.002

Araus, J. L., and Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends Plant Sci.* 19, 52–61. doi: 10.1016/j.tplants.2013.09.008

Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., and Cairns, J. E. (2018). Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci.* 23, 451–466. doi: 10.1016/j.tplants.2018.02.001

Araya, D. B., Grolinger, K., ElYamany, H. F., Capretz, M. A., and Bitsuamlak, G. (2017). An ensemble learning framework for anomaly detection in building energy consumption. *Energy Build.* 144, 191–206. doi: 10.1016/j.enbuild.2017.02.058

Auria, L., and Moro, R. A. (2008). *Support Vector Machines (SVM) as a Technique for Solvency Analysis*. Berlin: Deutsches Institut für Wirtschaftsforschung (DIW). DIW Discussion Papers, No. 811 doi: 10.2139/ssrn.1424949

Belayneh, A., Adamowski, J., Khalil, B., and Ozga-Zielinski, B. (2014). Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. *J. Hydrol.* 508, 418–429. doi: 10.1016/j.jhydrol.2013.10.052

Björne, J., Kaewphan, S., and Salakoski, T. (2013). "UTurku: drug named entity recognition and drug-drug interaction extraction using SVM classification and domain knowledge," in *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (* SEM),: Proceedings of the Seventh International Workshop on Semantic Evaluation*, (SemEval), 651–659.

Bowley, S. (1999). *A hitchhiker's guide to statistics in plant biology*. Guelph, Ont: Any Old Subject Books.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cai, G., Yang, Q., Chen, H., Yang, Q., Zhang, C., Fan, C., et al. (2016). Genetic dissection of plant architecture and yield-related traits in *Brassica napus*. *Sci. Rep.* 6:21625. doi: 10.1038/srep21625

Carvalho, S., Macel, M., Schlerf, M., Moghaddam, F. E., Mulder, P. P., Skidmore, A. K., et al. (2013). Changes in plant defense chemistry (pyrrolizidine alkaloids) revealed through high-resolution spectroscopy. *ISPRS J. Photogrammetry Remote Sens.* 80, 51–60. doi: 10.1016/j.isprsjprs.2013.03.004

Chen, J.-C., and Wang, Y.-M. (2020). Comparing activation functions in modeling shoreline variation using multilayer perceptron neural network. *Water* 12:1281. doi: 10.3390/w12051281

Chen, L., Huang, J., Wang, F., and Tang, Y. (2007). Comparison between back propagation neural network and regression models for the estimation of pigment content in rice leaves and panicles using hyperspectral data. *Int. J. Remote Sens.* 28, 3457–3478. doi: 10.1080/01431160601024242

Chen, X.-W., and Jeong, J. C. (2007). "Enhanced recursive feature elimination," in *Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, (IEEE), 429–435. doi: 10.1109/ICMLA.2007.35

Collins, N. C., Tardieu, F., and Tuberosa, R. (2008). Quantitative trait loci and crop performance under abiotic stress: where do we stand? *Plant Physiol.* 147, 469–486. doi: 10.1104/pp.108.118117

De Castro, A.-I., Jurado-Expósito, M., Gómez-Casero, M.-T., and López-Granados, F. (2012). Applying neural networks to hyperspectral and multispectral field data for discrimination of cruciferous weeds in winter crops. *Sci. World J.* 2012:630390. doi: 10.1100/2012/630390

Dennis, T. (2016). "Using R and ggvis to create interactive graphics for exploratory data analysis," in *Data Visualization: a Guide to Visual Storytelling for Libraries*, ed. L. Magnuson. (Lanham, MD: Rowman & Littlefield).

Dietterich, T. G. (2000). "Ensemble methods in machine learning," in *Proceedings of the International Workshop on Multiple Classifier Systems*, (Springer), 1–15. doi: 10.1007/3-540-45014-9_1

Drucker, H., and Cortes, C. (1996). Boosting decision trees. *Adv. Neural Inform. Process. Systems* 8, 479–485.

Duan, K.-B., Rajapakse, J. C., Wang, H., and Azuaje, F. (2005). Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Trans. Nanobiosci.* 4, 228–234. doi: 10.1109/TNB.2005.853657

Dubey, A., Kumar, A., Abd_Allah, E. F., Hashem, A., and Khan, M. L. (2019). Growing more with less: breeding and developing drought resilient soybean to improve food security. *Ecol. Indicators* 105, 425–437. doi: 10.1016/j.ecolind.2018.03.003

Eck, H. V. (1987). Plant water stress at various growth stages and growth and yield of soybeans. *Field Crops Res.* 17, 1–16. doi: 10.1016/0378-4290(87)90077-3

Etminan, A., Pour-Aboughadareh, A., Mohammadi, R., Shooshtari, L., Yousefiazarkhanian, M., and Moradkhani, H. (2019). Determining the best drought tolerance indices using artificial neural network (ANN): insight into application of intelligent agriculture in agronomy and plant breeding. *Cereal Res. Commun.* 47, 170–181. doi: 10.1556/0806.46.2018.057

Falk, K. G., Jubery, T. Z., Mirnezami, S. V., Parmley, K. A., Sarkar, S., Singh, A., et al. (2020). Computer vision and machine learning enabled soybean root phenotyping pipeline. *Plant Methods* 16:5. doi: 10.1186/s13007-019-0550-555

Fehr, W., Lawrence, B., and Thompson, T. (1981). Critical stages of development for defoliation of soybean 1. *Crop Sci.* 21, 259–262. doi: 10.2135/cropsci1981.0011183X002100020014x

Feilhauer, H., Somers, B., and van der Linden, S. (2017). Optical trait indicators for remote sensing of plant species composition: predictive power and seasonal variability. *Ecol. Indicators* 73, 825–833. doi: 10.1016/j.ecolind.2016.11.003

Feng, L., Zhang, Z., Ma, Y., Du, Q., Williams, P., Drewry, J., et al. (2020). Alfalfa yield prediction using UAV-Based hyperspectral imagery and ensemble learning. *Remote Sens.* 12:2028. doi: 10.3390/rs12122028

Feng, P., Ma, J., Sun, C., Xu, X., and Ma, Y. (2018). A novel dynamic android malware detection system with ensemble learning. *IEEE Access* 6, 30996–31011. doi: 10.1109/ACCESS.2018.2844349

Ferris, R., Wheeler, T., Hadley, P., and Ellis, R. (1998). Recovery of photosynthesis after environmental stress in soybean grown under elevated CO2. *Crop Sci.* 38, 948–955. doi: 10.2135/cropsci1998.0011183X003800040012x

Filippa, G., Cremonese, E., Migliavacca, M., Galvagno, M., Sonnentag, O., Humphreys, E., et al. (2018). NDVI derived from near-infrared-enabled digital cameras: applicability across different plant functional types. *Agricult. Forest Meteorol.* 249, 275–285. doi: 10.1016/j.agrformet.2017.11.003

Filippi, A. M., and Jensen, J. R. (2006). Fuzzy learning vector quantization for hyperspectral coastal vegetation classification. *Remote Sens. Environ.* 100, 512–530. doi: 10.1016/j.rse.2005.11.007

Foley, J. A., Ramankutty, N., Brauman, K. A., Cassidy, E. S., Gerber, J. S., Johnston, M., et al. (2011). Solutions for a cultivated planet. *Nature* 478, 337–342. doi: 10.1038/nature10452

Furbank, R. T., and Tester, M. (2011). Phenomics–technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* 16, 635–644. doi: 10.1016/j.tplants.2011.09.005

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Systems Man Cybernet. Part C (Applications and Reviews)* 42, 463–484. doi: 10.1109/TSMCC.2011.2161285

Gandhi, N., Armstrong, L. J., Petkar, O., and Tripathy, A. K. (2016). "Rice crop yield prediction in India using support vector machines," in *Proceedings of the 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, (IEEE), 1–5. doi: 10.1109/JCSSE.2016.7748856

Geetha, M. (2020). Forecasting the crop yield production in trichy district using fuzzy C-Means algorithm and multilayer perceptron (MLP). *Int. J. Knowledge Systems Sci. (IJKSS)* 11, 83–98. doi: 10.4018/IJKSS.2020070105

Ghorbani, M. A., Zadeh, H. A., Isazadeh, M., and Terzi, O. (2016). A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction. *Environ. Earth Sci.* 75:476. doi: 10.1007/s12665-015-5096-x

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *J. Am. Statist. Assoc.* 57, 369–375. doi: 10.1080/01621459.1962.10480665

Granitto, P. M., Furlanello, C., Biasioli, F., and Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometr. Intell. Lab. Systems* 83, 83–90. doi: 10.1016/j.chemolab.2006.01.007

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422. doi: 10.1023/A:1012487302797

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11, 10–18. doi: 10.1145/1656274.1656278

Hartman, G. L., West, E. D., and Herman, T. K. (2011). Crops that feed the World 2. soybean—worldwide production, use, and constraints caused by pathogens and pests. *Food Security* 3, 5–17. doi: 10.1007/s12571-010-0108-x

Hennessy, A., Clarke, K., and Lewis, M. (2020). Hyperspectral classification of plants: a review of waveband selection generalisability. *Remote Sens.* 12:113. doi: 10.3390/rs12010113

Hesami, M., Condori-Apfata, J. A., Valderrama Valencia, M., and Mohammadi, M. (2020a). Application of artificial neural network for modeling and studying in vitro genotype-independent shoot regeneration in wheat. *Appl. Sci.* 10:5370. doi: 10.3390/app10155370

Hesami, M., Naderi, R., Tohidfar, M., and Yoosefzadeh-Najafabadi, M. (2020b). Development of support vector machine-based model and comparative analysis with artificial neural network for modeling the plant tissue culture procedures: effect of plant growth regulators on somatic embryogenesis of chrysanthemum, as a case study. *Plant Methods* 16, 1–15. doi: 10.1186/s13007-020-00655-9

Hesami, M., and Jones, A. M. P. (2020). Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture. *Appl. Microbiol. Biotechnol.* 104, 9449–9485. doi: 10.1007/s00253-020-10888-10882

Hesami, M., Naderi, R., Tohidfar, M., and Yoosefzadeh-Najafabadi, M. (2019). Application of adaptive neuro-fuzzy inference system-non-dominated sorting genetic algorithm-II (ANFIS-NSGAII) for modeling and optimizing somatic embryogenesis of chrysanthemum. *Front. Plant Sci.* 10:869. doi: 10.3389/fpls.2019.00869

Jin, J., Liu, X., Wang, G., Mi, L., Shen, Z., Chen, X., et al. (2010). Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. *Field Crops Res.* 115, 116–123. doi: 10.1016/j.fcr.2009.10.016

Jolly, W. M., Nemani, R., and Running, S. W. (2005). A generalized, bioclimatic index to predict foliar phenology in response to climate. *Global Change Biol.* 11, 619–632. doi: 10.1111/j.1365-2486.2005.00930.x

Ju, C., Bibaut, A., and van der Laan, M. (2018). The relative performance of ensemble methods with deep convolutional neural networks for image classification. *J. Appl. Statist.* 45, 2800–2818. doi: 10.1080/02664763.2018.1441383

Katsileros, A., Drosou, K., and Koukouvinos, C. (2015). Evaluation of nearest neighbor methods in wheat genotype experiments. *Commun. Biometry Crop Sci.* 10, 115–123.

Kayad, A. G., Al-Gaadi, K. A., Tola, E., Madugundu, R., Zeyada, A. M., and Kalaitzidis, C. (2016). Assessing the spatial variability of alfalfa yield using satellite imagery and ground-based data. *PLoS One* 11:e0157166. doi: 10.1371/journal.pone.0157166

Khan, R., Hanbury, A., and Stoettinger, J. (2010). "Skin detection: a random forest approach," in *Proceedings of the 2010 IEEE International Conference on Image Processing*, (IEEE), 4613–4616. doi: 10.1109/ICIP.2010.5651638

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Statist. Software* 28, 1–26. doi: 10.18637/jss.v028.i05

Kycko, M., Zagajewski, B., Lavender, S., Romanowska, E., and Zwijacz-Kozica, M. (2018). The impact of tourist traffic on the condition and cell structures of alpine swards. *Remote Sens.* 10:220. doi: 10.3390/rs10020220

Lane, H. M., Murray, S. C., Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Rooney, D. K., et al. (2020). Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels. *Plant Phenome J.* 3:e20002. doi: 10.1002/ppj2.20002

Lawrence, S., and Giles, C. L. (2000). "Overfitting and neural networks: conjugate gradient and backpropagation," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, (IEEE), 114–119. doi: 10.1109/IJCNN.2000.857823

Lepetit, V., Lagger, P., and Fua, P. (2005). "Randomized trees for real-time keypoint recognition," in *Proceedings of the 2005 IEEE*, (IEEE), 775–781. doi: 10.1109/CVPR.2005.288

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R news* 2, 18–22.

Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., et al. (2011). "Large-scale image classification: fast feature extraction and SVM training," in *Proceedings of the CVPR*, (IEEE), 1689–1696. doi: 10.1109/CVPR.2011.5995477

Lopez-Cruz, M., Olson, E., Rovere, G., Crossa, J., Dreisigacker, S., Mondal, S., et al. (2020). Regularized selection indices for breeding value prediction using hyper-spectral image data. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-020-65011-2

Loughrey, J., and Cunningham, P. (2004). "Overfitting in wrapper-based feature subset selection: the harder you try the worse it gets," in *Proceedings of the International Conference on Innovative Techniques and Applications of Artificial Intelligence*, (Springer), 33–43. doi: 10.1007/1-84628-102-4_3

Lykhovyd, P. (2020). Sweet corn yield simulation using normalized difference vegetation index and leaf area index. *J. Ecol. Eng.* 21, 228–236 doi: 10.12911/22998993/118274

Ma, B., Dwyer, L. M., Costa, C., Cober, E. R., and Morrison, M. J. (2001). Early prediction of soybean yield from canopy reflectance measurements. *Agronomy J.* 93, 1227–1234. doi: 10.2134/agronj2001.1227

Ma, B., Morrison, M. J., and Dwyer, L. M. (1996). Canopy light reflectance and field greenness to assess nitrogen fertilization and yield of maize. *Agronomy J.* 88, 915–920. doi: 10.2134/agronj1996.00021962003600060011x

Makantasis, K., Karantzalos, K., Doulamis, A., and Doulamis, N. (2015). "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, (IEEE), 4959–4962. doi: 10.1109/IGARSS.2015.7326945

Meinshausen, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* 7, 983–999.

Merzlyak, M. N., Solovchenko, A. E., and Gitelson, A. A. (2003). Reflectance spectral features and non-destructive estimation of chlorophyll, carotenoid and anthocyanin content in apple fruit. *Postharvest Biol. Technol.* 27, 197–211. doi: 10.1016/S0925-5214(02)00066-2

Mokry, F. B., Higa, R. H., de Alvarenga Mudadu, M., de Lima, A. O., Meirelles, S. L. C., da Silva, M. V. G. B., et al. (2013). Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. *BMC Genet.* 14:47. doi: 10.1186/1471-2156-14-47

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., los Campos, G., Alvarado, G., Suchismita, M., et al. (2017). Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. *Plant Methods* 13:4. doi: 10.1186/s13007-016-0154-2

Murakoshi, K. (2005). Avoiding overfitting in multilayer perceptrons with feeling-of-knowing using self-organizing maps. *BioSystems* 80, 37–40. doi: 10.1016/j.biosystems.2004.09.031

Nguyen, M. H., and De la Torre, F. (2010). Optimal feature selection for support vector machines. *Pattern Recogn.* 43, 584–591. doi: 10.1016/j.patcog.2009.09.003

Nourani, V., Elkiran, G., and Abba, S. (2018). Wastewater treatment plant performance analysis using artificial intelligence–an ensemble approach. *Water Sci. Technol.* 78, 2064–2076. doi: 10.2166/wst.2018.477

Ozaki, Y., McClure, W. F., and Christy, A. A. (2006). *Near-infrared Spectroscopy in Food Science and Technology*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/0470047704

Pal, M. (2007). Ensemble learning with decision tree for remote sensing classification. *World Academy Sci. Eng. Technol.* 36, 258–260.

Pal, S. K., and Mitra, S. (1992). Multilayer perceptron, fuzzy sets, classifiaction. *IEEE Trans. Neural Networks* 3, 683–697. doi: 10.1109/72.159058

Pedersen, P., Kumudini, S., Board, J., and Conley, S. (2004). *Soybean Growth and Development*. University Extension Ames, IA: Iowa State University.

Phan, P., Chen, N., Xu, L., and Chen, Z. (2020). Using multi-temporal MODIS NDVI data to monitor tea status and forecast yield: a case study at tanuyen, laichau, vietnam. *Remote Sens.* 12:1814. doi: 10.3390/rs12111814

Prey, L., and Schmidhalter, U. (2019). Simulation of satellite reflectance data using high-frequency ground based hyperspectral canopy measurements for in-season estimation of grain yield and grain nitrogen status in winter wheat. *ISPRS J. Photogrammetry Remote Sens.* 149, 176–187. doi: 10.1016/j.isprsjprs.2019.01.023

Richter, R., Reu, B., Wirth, C., Doktor, D., and Vohland, M. (2016). The use of airborne hyperspectral data for tree species classification in a species-rich Central European forest area. *Int. J. Appl. Earth Observation Geoinform.* 52, 464–474. doi: 10.1016/j.jag.2016.07.018

Rocha, A., Groen, T., Skidmore, A., Darvishzadeh, R., and Willemen, L. (2018). Machine learning using hyperspectral data inaccurately predicts plant traits under spatial dependency. *Remote Sens.* 10:1263. doi: 10.3390/rs10081263

Rossel, R. A. V. (2008). ParLeS: software for chemometric analysis of spectroscopic data. *Chemometrics Intell. Lab. Systems* 90, 72–83. doi: 10.1016/j.chemolab.2007.06.006

Roy, K., Kar, S., and Das, R. N. (2015). "Statistical methods in QSAR/QSPR," in *A Primer on QSAR/QSPR Modeling*, eds K. Roy, S. Kar, and R. N. Das (Berlin: Springer), 37–59. doi: 10.1007/978-3-319-17281-1_2

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Párez, L. G., Crossa, J., et al. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3 (Bethesda)* 6, 2799–2808. doi: 10.1534/g3.116.032888

Savitzky, A., and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36, 1627–1639. doi: 10.1021/ac60214a047

Schapire, R. E. (1999). "A brief introduction to boosting," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, (Florham Park, NJ), 1401–1406.

Schweiger, A. K., Cavender-Bares, J., Townsend, P. A., Hobbie, S. E., Madritch, M. D., Wang, R., et al. (2018). Plant spectral diversity integrates functional and phylogenetic components of biodiversity and predicts ecosystem function. *Nat. Ecol. Evol.* 2, 976–982. doi: 10.1038/s41559-018-0551-1

Serneels, S., and Verdonck, T. (2008). Principal component analysis for data containing outliers and missing elements. *Comput. Statist. Data Anal.* 52, 1712–1727. doi: 10.1016/j.csda.2007.05.024

Šestak, I., Zgorelec, Ž, Perèin, A., Mesiæ, M., and Galiæ, M. (2019). "Prediction of soybean leaf nitrogen content using proximal field spectroscopy," in *Proceedings of the 54th Croatian & 14th International Symposium on Agriculture*, (Vodice).

Shao, Y., Zhao, C., Bao, Y., and He, Y. (2012). Quantification of nitrogen status in rice by least squares support vector machines and reflectance spectroscopy. *Food Bioprocess Technol.* 5, 100–107. doi: 10.1007/s11947-009-0267-y

Siebers, M. H., Yendrek, C. R., Drag, D., Locke, A. M., Rios Acosta, L., Leakey, A. D., et al. (2015). Heat waves imposed during early pod development in soybean (G lycine max) cause significant yield loss despite a rapid recovery from oxidative stress. *Global Change Biol.* 21, 3114–3125. doi: 10.1111/gcb.12935

Siegmann, B., and Jarmer, T. (2015). Comparison of different regression models and validation techniques for the assessment of wheat leaf area index from hyperspectral data. *Int. J. Remote Sens.* 36, 4519–4534. doi: 10.1080/01431161.2015.1084438

Simopoulos, C. M., Weretilnyk, E. A., and Golding, G. B. (2018). Prediction of plant lncRNA by ensemble machine learning classifiers. *BMC Genom.* 19:316. doi: 10.1186/s12864-018-4665-2

Stroup, W., and Mulitze, D. (1991). Nearest neighbor adjusted best linear unbiased prediction. *Am. Statistician* 45, 194–200. doi: 10.1080/00031305.1991.10475801

Su, C.-T., and Yang, C.-H. (2008). Feature selection for the SVM: an application to hypertension diagnosis. *Exp. Systems Appl.* 34, 754–763. doi: 10.1016/j.eswa.2006.10.010

Su, Q., Lu, W., Du, D., Chen, F., Niu, B., and Chou, K.-C. (2017). Prediction of the aquatic toxicity of aromatic compounds to tetrahymena pyriformis through support vector regression. *Oncotarget* 8, 49359–49369. doi: 10.18632/oncotarget.17210

Suykens, J. A., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Process. Lett.* 9, 293–300. doi: 10.1023/A:1018628609742

Sweeney, D. W., Long, J. H., and Kirkham, M. (2003). A single irrigation to improve early maturing soybean yield and quality. *Soil Sci. Soc. Am. J.* 67, 235–240. doi: 10.2136/sssaj2003.2350

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Rev.* 144, 2375–2393. doi: 10.1175/MWR-D-15-0260.1

Tamouridou, A. A., Alexandridis, T. K., Pantazi, X. E., Lagopodi, A. L., Kashefi, J., Kasampalis, D., et al. (2017). Application of multilayer perceptron with automatic relevance determination on weed mapping using UAV multispectral imagery. *Sensors* 17:2307. doi: 10.3390/s17102307

Tan, M., Wang, L., and Tsang, I. W. (2010). "Learning sparse svm for feature selection on very high dimensional datasets," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, (Haifa).

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., and Bennett, M. (2017). Plant phenomics, from sensors to knowledge. *Curr. Biol.* 27, R770–R783. doi: 10.1016/j.cub.2017.05.055

Tilman, D., Socolow, R., Foley, J. A., Hill, J., Larson, E., Lynd, L., et al. (2009). Beneficial biofuels—the food, energy, and environment trilemma. *Science* 325, 270–271. doi: 10.1126/science.1177970

Tiwari, P., and Shukla, P. (2020). "Artificial neural network-based crop yield prediction using NDVI, SPI, VCI feature vectors," in *Proceedings of the Information and Communication Technology for Sustainable Development Proceedings of ICT4SD*, (Springer), 585–594. doi: 10.1007/978-981-13-7166-0_58

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. New York, NY: SpringerVerlag. doi: 10.1007/978-1-4757-3264-1

Verma, A. K., Garg, P. K., Hari Prasad, K., and Dadhwal, V. K. (2020). Modelling of sugarcane yield using LISS-IV data based on ground LAI and yield observations. *Geocarto Int.* 35, 887–904. doi: 10.1080/10106049.2018.1544291

Wan, L., Cen, H., Zhu, J., Zhang, J., Zhu, Y., Sun, D., et al. (2020). Grain yield prediction of rice using multi-temporal UAV-based RGB and multispectral images and model transfer–a case study of small farmlands in the South of China. *Agricult. Forest Meteorol.* 291:108096. doi: 10.1016/j.agrformet.2020.108096

Wang, Y., Wang, F., Huang, J., Wang, X., and Liu, Z. (2009). Validation of artificial neural network techniques in the estimation of nitrogen concentration in rape using canopy hyperspectral reflectance data. *Int. J. Remote Sens.* 30, 4493–4505. doi: 10.1080/01431160802577998

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Rev. Comput. Statist.* 3, 180–185. doi: 10.1002/wics.147

Xiong, Q., Tang, G., Zhong, L., He, H., and Chen, X. (2018). Response to nitrogen deficiency and compensation on physiological characteristics, yield formation, and nitrogen utilization of rice. *Front. Plant Sci.* 9:1075. doi: 10.3389/fpls.2018.01075

Zhang, N., Pan, Y., Feng, H., Zhao, X., Yang, X., Ding, C., et al. (2019). Development of Fusarium head blight classification index using hyperspectral microscopy images of winter wheat spikelets. *Biosystems Eng.* 186, 83–99. doi: 10.1016/j.biosystemseng.2019.06.008

Zhang, Z., Jin, Y., Chen, B., and Brown, P. (2019). California almond yield prediction at the orchard level with a machine learning approach. *Front. Plant Sci.* 10:809. doi: 10.3389/fpls.2019.00809

Zheng, W., Zhu, X., Wen, G., Zhu, Y., Yu, H., and Gan, J. (2020). Unsupervised feature selection by self-paced learning regularization. *Pattern Recogn. Lett.* 132, 4–11. doi: 10.1016/j.patrec.2018.06.029

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: CRC press. doi: 10.1201/b12207

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.