PLOS ONE

RESEARCH ARTICLE

Application of machine learning and genetic optimization algorithms for modeling and optimizing soybean yield using its component traits

Mohsen Yoosefzadeh-Najafabadi[1], **Dan Tulpan**[2], **Milad Eskandari**[1]*

1 Department of Plant Agriculture, University of Guelph, Guelph, Ontario, Canada, **2** Department of Animal Biosciences, University of Guelph, Guelph, Ontario, Canada

* meskanda@uoguelph.ca

OPEN ACCESS

Citation: Yoosefzadeh-Najafabadi M, Tulpan D, Eskandari M (2021) Application of machine learning and genetic optimization algorithms for modeling and optimizing soybean yield using its component traits. PLoS ONE 16(4): e0250665. https://doi.org/10.1371/journal.pone.0250665

Editor: Qiang Zeng, South China University of Technology, CHINA

Received: January 24, 2021

Accepted: April 12, 2021

Published: April 30, 2021

Data Availability Statement: The data are uploaded in an open access file in Github, and can be accessed through the following link: https://github.com/Mohsen1080/Available-Datasets.

Funding: This project was funded in part by Grain Farmers of Ontario (GFO) and SeCan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Improving genetic yield potential in major food grade crops such as soybean *(Glycine max* L.) is the most sustainable way to address the growing global food demand and its security concerns. Yield is a complex trait and reliant on various related variables called yield components. In this study, the five most important yield component traits in soybean were measured using a panel of 250 genotypes grown in four environments. These traits were the number of nodes per plant (NP), number of non-reproductive nodes per plant (NRNP), number of reproductive nodes per plant (RNP), number of pods per plant (PP), and the ratio of number of pods to number of nodes per plant (P/N). These data were used for predicting the total soybean seed yield using the Multilayer Perceptron (MLP), Radial Basis Function (RBF), and Random Forest (RF), machine learning (ML) algorithms, individually and collectively through an ensemble method based on bagging strategy (E-B). The RBF algorithm with highest Coefficient of Determination (R^2) value of 0.81 and the lowest Mean Absolute Errors (MAE) and Root Mean Square Error (RMSE) values of 148.61 kg.ha^{-1}, and 185.31 kg.ha^{-1}, respectively, was the most accurate algorithm and, therefore, selected as the meta-Classifier for the E-B algorithm. Using the E-B algorithm, we were able to increase the prediction accuracy by improving the values of R^2, MAE, and RMSE by 0.1, 0.24 kg.ha^{-1}, and 0.96 kg.ha^{-1}, respectively. Furthermore, for the first time in this study, we allied the E-B with the genetic algorithm (GA) to model the optimum values of yield components in an ideotype genotype in which the yield is maximized. The results revealed a better understanding of the relationships between soybean yield and its components, which can be used for selecting parental lines and designing promising crosses for developing cultivars with improved genetic yield potential.

Introduction

Soybean (*Glycine max* L. Merrill) is the world's most widely grown leguminous crop and an important oil and protein source for food and feed [1]. Because of its nutritional and

Competing interests: The authors have declared that no competing interests exist.

pharmaceutical properties, soybean is also considered as an important source of healthy plant-based food products in the human diet. While the global demand for soybean is increasing significantly [2], the current average annual genetic gain for yield seem not to be able to cope with the growing demand [3]. One of the probable main reasons for this low genetic gain is the inefficient selection criteria that are currently used in breeding programs for selecting genotypes with desirable genetic yield potentials [4].

Despite the major advances in molecular technologies and their potential implications in breeding programs, generating reliable phenotypic data and analyzing big datasets have been remained the major bottlenecks [5]. Plant breeding programs, including soybeans, are continuously relying on the evaluation of yield and important agronomic traits for making selections and defining commercial products [6]. If a trait is under controlled by a few major genes, and so with limited environmental effects, designing molecular marker tools would be sufficient for selecting desirable genotypes in a given breeding population [7]. However, for complex traits such as yield, which are highly influenced by the environment and controlled by numerous genes with minor effects, the dissection of traits underlying the yield can be beneficial for selecting genotypes with improved genetic potentials [8]. In general, soybean breeders have made extensive use of the classical phenotypic selection approaches to evaluate and exploit total seed yield as the main selection criterion in their cultivar development programs [8, 9]. However, genetic gains for yield have generally been low and inconsistent across different environments [8]. The general low genetic gains for yield can be to a large extent because of the nature of this trait, in which several secondary traits drive the final production directly or indirectly [3]. Thus, one possible way to increase the yield genetic gains in new cultivars is to improve their yield component traits [5, 7, 8].

Yield formation in plant species is mostly governed by yield component traits [10–12]. Traits such as the number of nodes per plant (NP), number of non-reproductive nodes per plant (NRNP), number of reproductive nodes per plant (RNP), and number of pods per plant (PP) are considered as the major yield components in soybean [13]. Therefore, the increase in yield production can be regulated by selecting the soybeans with greater performance in their yield components [13, 14]. Several studies have been conducted to describe yield improvement in plants via improving yield components [15–18]. For example, positive correlations between soybean yield and the number of pods [19] were reported to be the most significant contributor to yield gain in Northeast China [20]. By considering yield components, soybean breeders can model and predict optimum conditions in which the highest yield production can achieve [21]. However, developing reliable prediction models built upon several yield component traits requires dealing with large datasets that are generated from the evaluation of large breeding populations across multi-environments.

Due to the essential need for advanced skills in computational and mathematical analyses, exploiting large datasets in many public breeding programs is still a bottleneck. Machine Learning (ML) algorithms, as one of the reliable and efficient computational approaches, were successfully implemented in different fields of study, such as traffic crash frequency modeling [22, 23], environmental science [24], engineering [25], and medicine [26]. Previously, Zeng, Huang (22) developed neural network, as one of the ML algorithms for exploring the non-linear relationship between risk factors and crash frequency. They reported the successfulness of using neural network algorithms to eliminate the overfitting and deal with lack-box characteristic [22]. Also, several studies reported the efficiency of ML algorithms in better detection of genomic regions associated with a trait of interest [26–28].

One of the recent agriculture trends is the use of ML algorithms for analyzing big data [29, 30]. Emerging ML algorithms in agriculture have created new opportunities to quantify and understand the intensive data process in agriculture [29, 31]. In a simple form, ML algorithms

can be defined as machines with the ability to learn without explicitly programmed [29, 32]. Theoretically, each algorithm is involved in a specific learning process from training data to perform a task of clustering, predicting, and classifying new datasets using the knowledge attained during the learning process [31]. Various ML algorithms have been developed and can be implemented for complex interactions between features [8, 29, 32, 33]. Multilayer Perceptron (MLP), for example, is known as the common neural network algorithm that is widely used in different areas such as plant sciences [30], remote sensing [34], environmental sciences [35], and engineering [36]. Like other neural network algorithms, MLP is built upon many neurons in which each neuron has its own specific weight [37]. In any case that one neuron is insufficient to explain the algorithm, MLP will be useful by providing multi-neurons [38, 39]. Radial Basis Function (RBF) is another ML algorithm commonly used in plant sciences [40–42]. RBF is reported to be successful for predictions wherever relevant features are used [40]. However, its performance for predicting soybean yield from its components is still unknown. Random Forest (RF) is another algorithm that its performance was evaluated in this study. RF has drawn many researchers' attention because of adequate performances in various fields, including plant science [30, 43], animal science [44, 45], human science [46], and remote sensing [47].

One of the major impediments of using individual ML algorithms is the high probability of overfitting in single predictive algorithms [48]. To overcome this obstacle, the ensemble techniques can be employed [49]. Ensemble techniques are known as the most influential development in the application of ML algorithms [50], in which combined algorithms are exploited to improve prediction accuracies by reducing overfitting rates [48–50]. Three commonly used ensemble algorithms are stacking, boosting, and bagging methods that are used according to the nature of the dataset and the individual ML algorithms that are used [50]. The success of using ensemble techniques was reported in different areas such as plant science [30], engineering [51], and computer sciences [52].

In soybean cultivar development programs, an optimum selection among important yield components can significantly improve the yield genetic gain. Therefore, the implementation of optimization methods in this field is of particular interest. Optimization algorithms for improving important traits are becoming more and more attractive in plant science [40, 53]. Genetic Algorithm (GA) is known as one of the most well-known single objective optimization methods designed and developed by Holland [54], as a searching algorithm based on natural selection. GA searching algorithm is based on Darwin's notion that more stable organisms across different environments survive better than the others [55].

Although the successful uses of ML ensemble methods have been reported in different agriculture-related fields [30, 48, 53, 56, 57], the potential use of these algorithms to predict soybean yield using yield components remains unknown. Therefore, this study aimed to investigate the potential use of soybean yield components for predicting the final seed yield using individual ML algorithms as well as ensemble learning methods. In addition, linking the best machine learning algorithm with GA for estimating the optimized values of the yield components for maximizing the soybean yield was investigated. The outcomes of this study can pave the way to understand the importance of soybean yield components in determining the total seed yield and implement the proposed pipeline for making the genotypic selection more accurate.

Material and methods

Plant material and experimental design

Two hundred and fifty soybean genotypes were grown under field conditions at two locations: Palmyra (42˚25'50.1"N 81˚45'06.9"W, 195 m above sea level) and Ridgetown (42˚27'14.8"N

81˚52'48.0"W, 200m above sea level), in Ontario, Canada in 2018–2019. The population used in this study was selected from the core germplasms of soybean breeding programs at the University of Guelph, Ridgetown campus that have been created over 35 years and used for cultivar development and genetic studies. The experiment was conducted using randomized complete block designs (RCBD) with two replications in four environments (two locations × two years), consisting of 2000 phenotypic plots in total. Each plot consisted of five rows, each 4.2 m long and 40 cm spacing between each row. The seeding rates used in this study were 50–57 seeds per m^2.

Phenotypic evaluations

Seed yield (ton ha^{-1}) of each plot was measured by harvesting three middle rows and adjusting to a 13% moisture level. Soybean yield components, including NP, NRNP, RNP, and PP, were hand-measured by randomly selected ten plants from each phenotypic plot for each genotype. Also, the PP to NP ratio (P/N) for each genotype was calculated using the following equation:

$$P/N = \frac{\mathrm{PP}}{\mathrm{NP}} \tag{1}$$

where *PP* indicates the number of pods per plant, and *NP* indicates the total number of nodes per plant.

Data pre-processing, correlation coefficient, and statistical analyses

All the phenotypic plot-based data were adjusted for spatial variations within the fields using nearest neighbor analysis (NNA) as one of the spatial error control methods to reduce and minimize the possible error in the field data [58, 59]. The yield components were collected for 250 soybean genotypes from 2000 plots across four environments. The average value of each yield component for each genotype was estimated through the best linear unbiased prediction (BLUP) as a mixed model [60]. For BLUP analysis, the environment factor was selected as a fixed effect, and the genotype factor was considered as a random effect. Afterward, all 250 data-points were used for constructing training and testing datasets. In this study, all the yield component traits such as NP, NRNP, RNP, PP, and P/N were considered as input variables for predicting the soybean yield as the output variable. In order to improve the prediction accuracy of input variables, data scaling and centering were applied for the pre-processing and pre-treatment steps [61]. Before performing ML algorithms, the principal component (PC) analysis was applied to identify outliers; however, no outlier was detected. The Pearson coefficient of correlations between seed yield and yield components were estimated using the R software version 3.6.1.

Data-driven modeling

MLP, RBF, and RF are the most commonly used ML algorithms with distinct functions and abilities that are used for predicting yield in plant crop species [37, 48, 62]. The MLP, as one of the most common feed-forward neural networks, consists of input, hidden, and output layers of interconnected neurons [63]. RBF is another type of neural network that used approximate multivariate functions [64]. RBF has the same functionality as that of MLP but in an effective way for using in more than one dimension [65]. RF is also another commonly used ML algorithm that generates combined trees representing *n* number of independent observations [66]. The final prediction in RF is determined based on the average predictions of all possible independent trees [48]. In order to improve the prediction performance of individual ML

algorithm, an ensemble method was proposed based on the bagging strategy (E-B). The steps for the E-B analyses are as follows: (1) applying and training MLP, RBF, and RF, independently; (2) selecting the best ML algorithm, based on the validation set, as the metaClassifier for the E-B; and (3) combining the prediction results in order to improve the prediction accuracy of the soybean yield [67]. All the used parameters in each algorithm were optimized based on the training dataset. The Weka software version 3.9.4 [68] was used to run all the tested ML algorithms and the E-B method for analyzing training and validation sets.

Optimization process via GA

For obtimizing the values of NP, PP, RNP, NRNP, and P/N in a theoretical maximized yield ideotype genotype, the best ML algorithm was linked to the genetic algorithm (GA). In order to have the best implement of GA, some parameters such as mutation rate, crossover fraction, and the number of chromosomes should be determined. For that, optimum values of the mentioned parameters are estimated by the trial and error methods. In brief, a chromosome is known as a set of variables that define as a possible solution to the problem. In this study, all the possible combinations of the yield component traits were considered as different chromosomes to form the maximum soybean yield production. All proposed chromosomes are constructed from the initial population, which is the first step in the GA optimization process [53]. Crossover in GA is the process of creating a new generation of chromosomes by mixing the existing chromosomes [69]. In this study, the possible crossover between two chromosomes, each containing a combination pattern of the yield components, was evaluated by using a two-point crossover, which is known as one of the common crossover fractions. Mutation is another parameter in GA, which is used to control the local minima in the population [69]. By using a mutation rate, the possibility of having similar chromosomes will be decreased, and therefore, the possible local minima are decreased [53]. In this study, the mutation and crossover rates as well as the number of chromosomes were set to 0.1, 0.7, and 100, respectively (Fig 1). Also, the Roulette wheel was used for selecting elite populations for crossover to obtain the appropriate fitness. In order to achieve the generation number, the generational practice was iteratively implemented. The lower and upper bounds of the dataset were considered as the constraints in the optimization process (Fig 1).

Quantification of model performance and error estimations

The original dataset consists of 250 observations was randomly divided into training and validation sets based on the five k-fold cross-validation method [70] with ten repetitions (Fig 2). To quantify the performance of the ML algorithms for predicting soybean seed yield from the yield components, the following statistical measurements between independent reference values (Y) and estimated values (Y′) were applied: The Root Mean Square Error (RMSE, Eq 2) and the Mean Absolute Errors (MAE, Eq 3) of the validation set.

$$RMSE = \sqrt{\frac{\sum (Y' - Y)^2}{n}} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{n} |Y'_i - YI|}{n} \tag{3}$$

where Y' stands for predicted value, Y is the field measured value, and n is the number of observations for a given genotype.

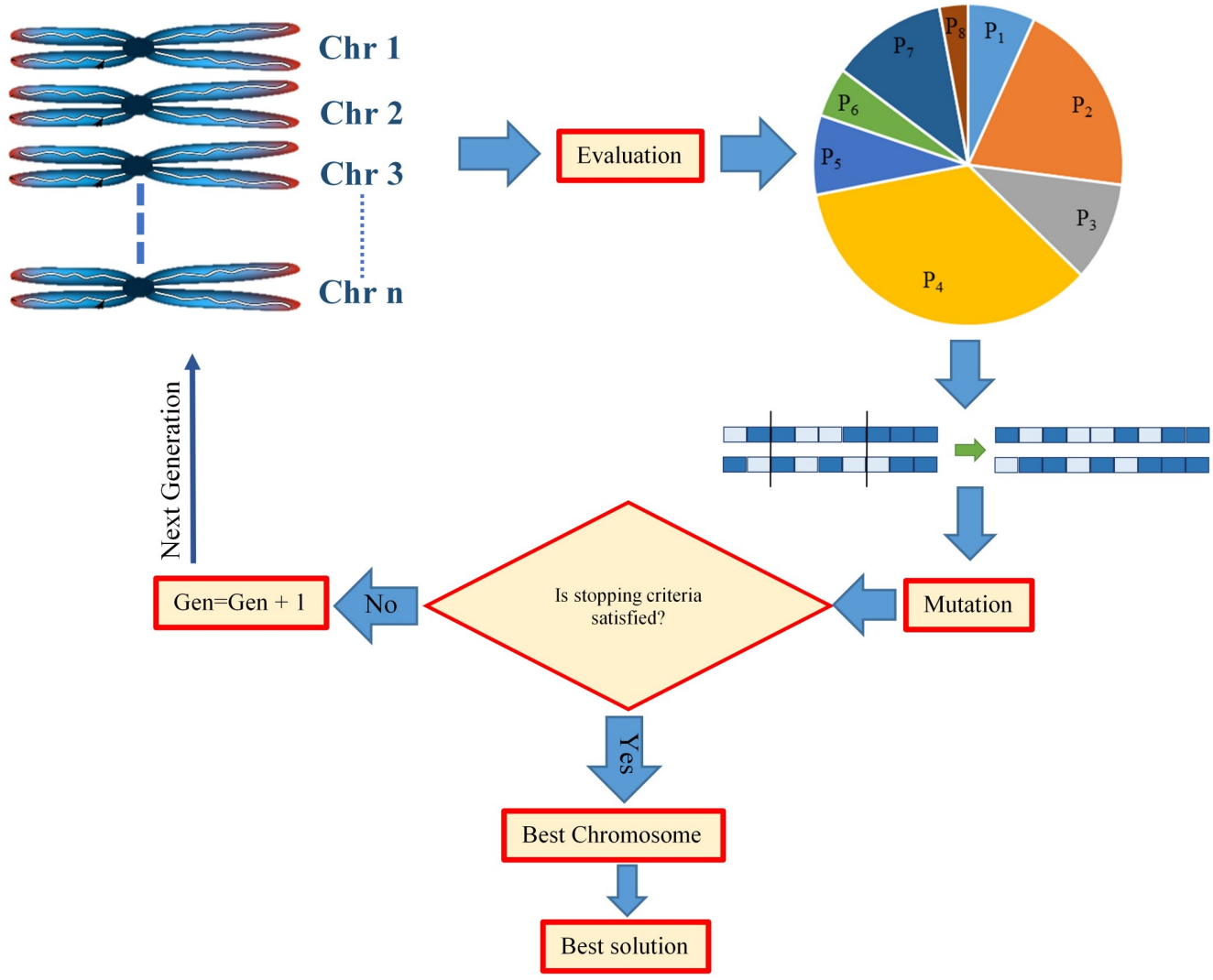

Fig 1. The schematic diagram of the genetic algorithm as the single objective evolutionary optimization algorithm.

https://doi.org/10.1371/journal.pone.0250665.g001

The analyses of the *goodness of fit* between observed and the predicted values were performed using the coefficient of determination (R^2, Eq 4). While we provide their definitions below, more comprehensive descriptions and definitions can be found in [71–73]:

$$R^2 = \frac{SST - SSE}{SST} \tag{4}$$

where *SST* stands for the sum of squares for total, and *SSE* stands for the sum of the squares for error.

Visualizing and statistical analyzing

The results were visualized using the Microsoft Excel software (2019), *ggvis* [74], and *ggplot2* [75] packages in the R software version 3.6.1. All statistical computational procedures were also conducted using *MASS* [76] package in R software.

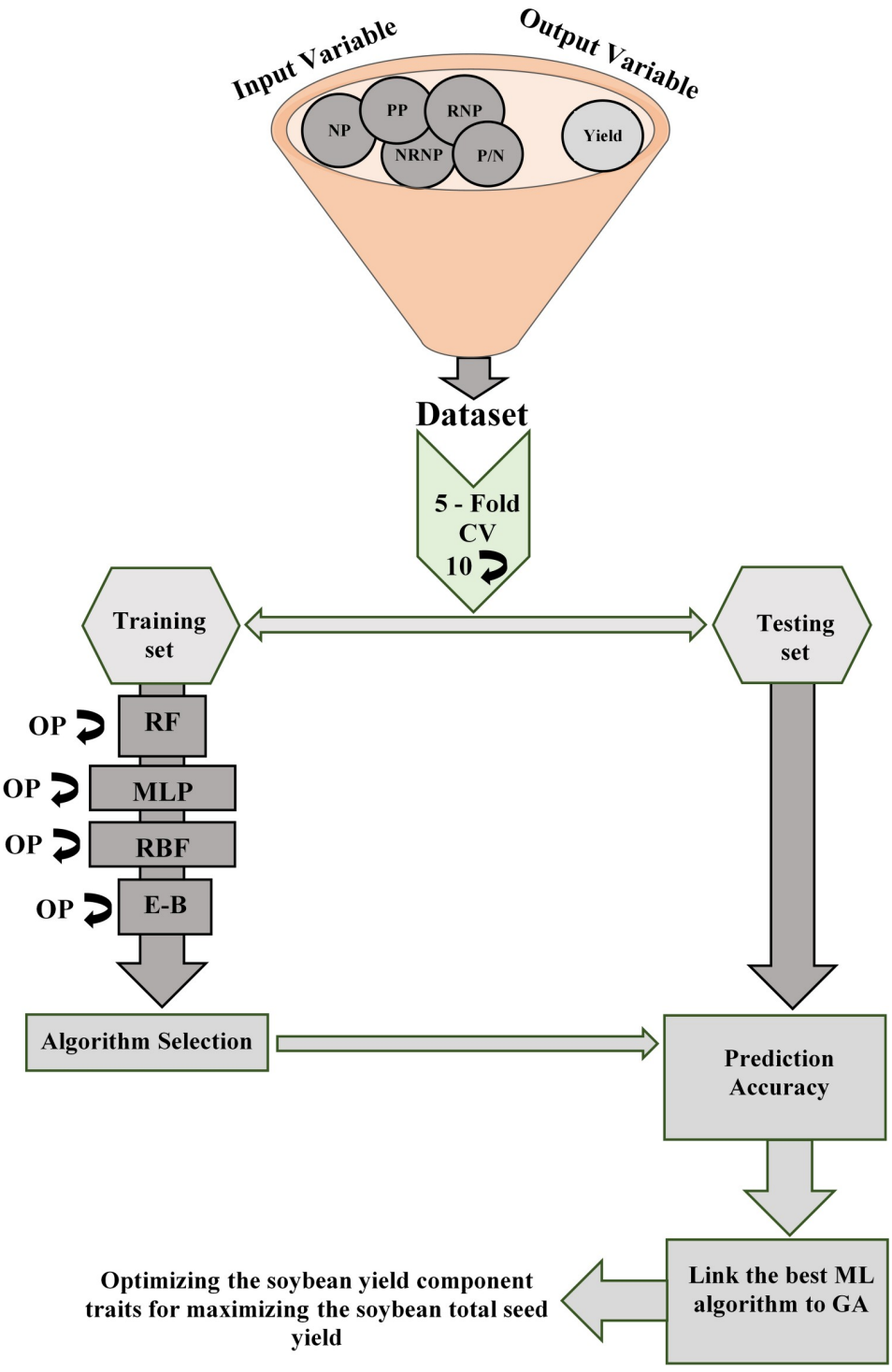

Fig 2. The experimental workflow of algorithm selection and validation for predicting the soybean seed yield. The Number of Nodes per plant (NP), the Number of Non-Reproductive Nodes per Plant (NRNP), the Number of Reproductive Nodes per Plant (RNP), and the Number of Pods per Plant (PP), the ratio of number of Pods to number of Nodes per plant (P/N), Genetic Algorithm (GA).

https://doi.org/10.1371/journal.pone.0250665.g002

Results

Pearson correlation analyses and individual ML evaluations

Based on the field data analyses, the average yield of 250 soybean genotypes was estimated to be between 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 potential benefits of using each soybean yield component for predicting the soybean seed yield was quantified using the Pearson coefficients of correlation among all the measured traits. Based on the correlation coefficients, all the yield components, except NRNP, were positively correlated with soybean seed yield. The linear correlation between soybean seed yield and PP (r = 0.71) was found to be the strongest followed by its correlation with NP (r = 0.68), RNP (r = 0.67), and P/N (r = 0.64). The negative correlation between soybean seed yield and NRNP was estimated as r = -0.29 (Fig 3).

Based on the results of correlation analyses, the top correlated variables were iteratively added to the algorithms and updated the algorithm training performances until all variables were included. The R^2 values of each model were then calculated (Fig 4). Among all the tested ML algorithms, the R^2 reached its maximum value of 0.81 in RBF taking into account PP, NP, and RNP. No changes in R^2 value were detected after adding P/N and NRNP in the RBF algorithm. The maximum R^2 value of 0.80 was obtained for both RF and MLP when all the yield components are added into the algorithms (Fig 4).

Model performance and evaluation

The full analysis of ML algorithms are presented in the S1 Table. Among the three tested ML algorithms, RBF had the highest value for R^2 (0.81) and the lowest values for MAE (148.61 kg. ha^{-1}) and RMSE (185.31 kg.ha^{-1}) (Fig 5A–5C). The R^2 values for MLP and RF were the same

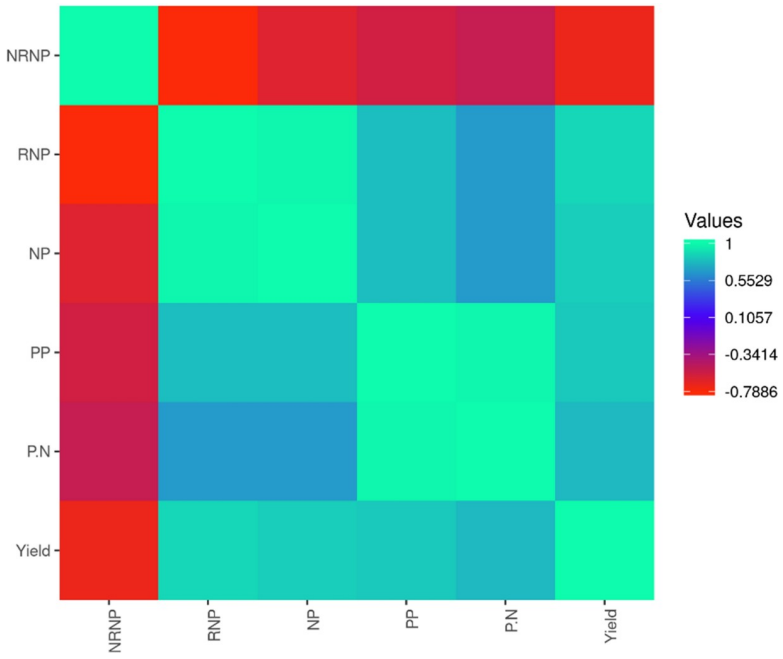

Fig 3. Pearson correlation analysis of soybean yield component traits. The Number of Nodes per plant (NP), the Number of Non-Reproductive Nodes per Plant (NRNP), the Number of Reproductive Nodes per Plant (RNP), and the Number of Pods per Plant (PP), the ratio of number of Pods to number of Nodes per plant (P.N).

https://doi.org/10.1371/journal.pone.0250665.g003

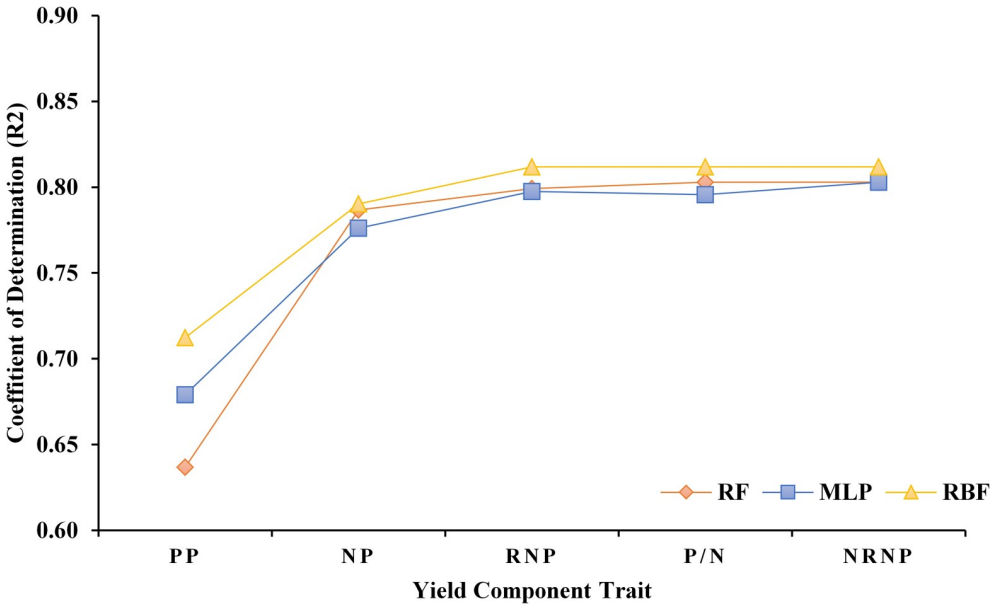

Fig 4. Model training accuracy for Random Forest (RF), Multilayer Perceptron (MLP), and Radial Basis Function (RBF) algorithms by adding variables based on the correlation results. The Number of Nodes per plant (NP), the Number of Non-Reproductive Nodes per Plant (NRNP), the Number of Reproductive Nodes per Plant (RNP), and the Number of Pods per Plant (PP), the ratio of number of Pods to number of Nodes per plant (P/N).

https://doi.org/10.1371/journal.pone.0250665.g004

(0.80); however, they had different values for MAE and RMSE. In comparison with the MLP algorithm, RF had the lower MAE and RMSE with 156.28 kg.ha^{-1} and 194.75 kg.ha^{-1}, respectively (Fig 5A–5C). MLP had the highest MAE (172.98 kg.ha^{-1}), and RMSE (211.57 kg.ha^{-1}) among the ML algorithms. In addition to individual evaluations of the three ML algorithms, an ensemble learning was also developed, which outperformed all the individual machine learning algorithms, attaining an R^2 value of 0.82. The E-B method had the acceptable RMSE and MAE with a value of 184.35 kg.ha^{-1} and 148.37 kg.ha^{-1}, respectively (Fig 5A–5C). In general, the R^2 value of E-B increased by 0.1, while the values of MAE and RMAS decreased by 0.24 kg.ha^{-1} and 0.96 kg.ha^{-1}, respectively, in comparison with RBF as the most accurate individual ML algorithm identified in this study.

Optimization of the soybean seed yield using E-B-GA

The aim of the current study, not only was to predict soybean seed yield using yield components, but also to estimate the optimum values of these traits, i.e., NP, PP, RNP, NRNP, and P/N, to maximize the final yield production in a given genotype. The results of the optimization process using GA, as the single objective evolutionary optimization algorithm, are presented in Table 1. Theoretically, the highest soybean seed yield production of 5.64 ton ha^{-1} can be achieved in an ideotype soybean genotype in which the values of NP, NRNP, RNP, PP, and P/N are 17.32, 3.07, 14.25, 48.98, and 2.83, respectively.

Discussion

One of the objectives of this study was to investigate the potential use of soybean yield components such as NP, PP, RNP, NRNP, and P/N for predicting the final seed yield production. Many studies reported the reliance of the final yield production on the performance of several yield-related traits [15, 77–80]. In soybean, PP and NP are considered as major players in

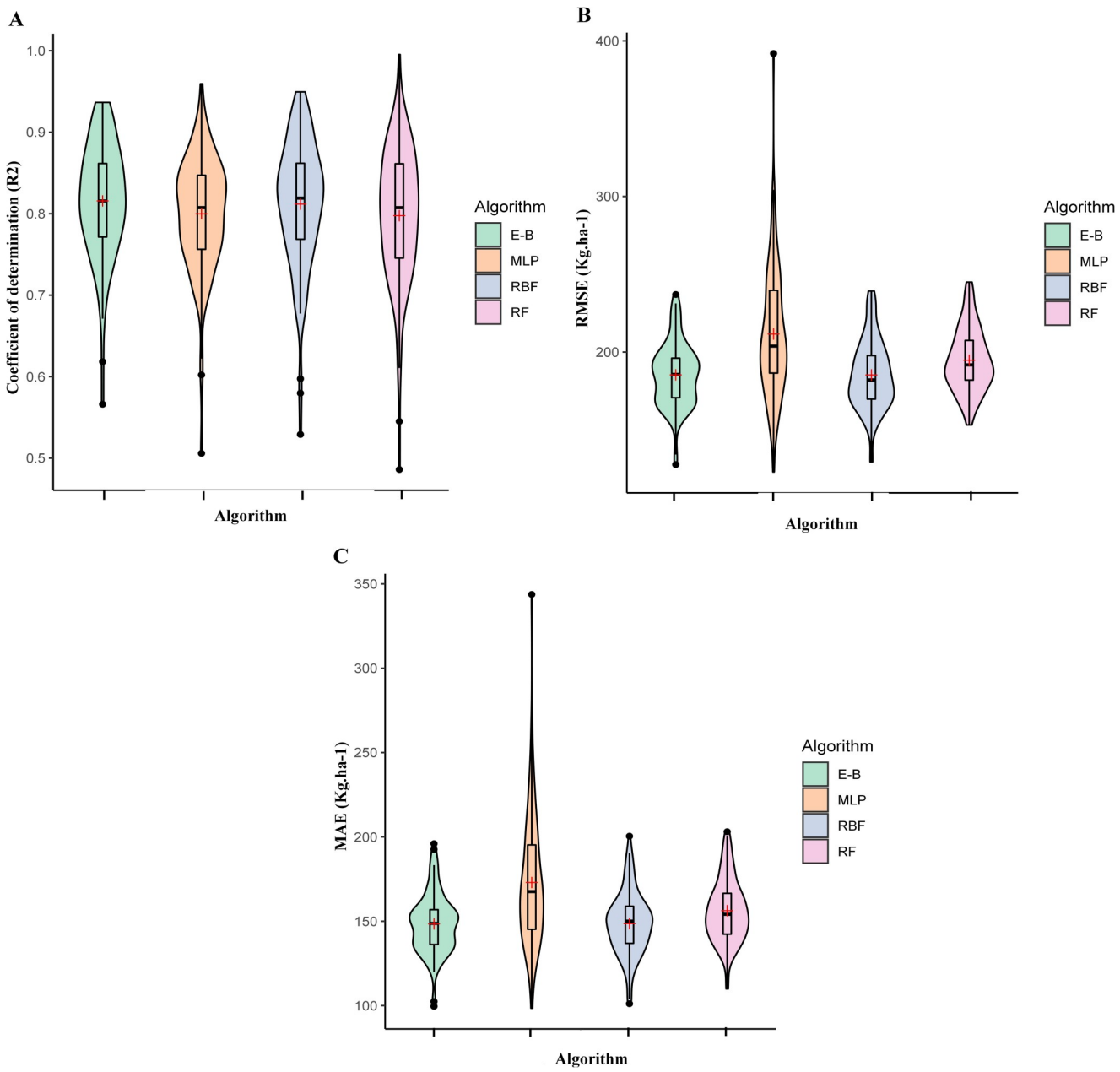

Fig 5. (A) coefficient of determination (R^2), (B) the Root Mean Square Error (RMSE) and (C) the Mean Absolute Errors (MAE) performance of Random Forest (RF), Multilayer Perceptron (MLP), and Radial Basis Function (RBF) algorithms, and the Ensemble-Bagging (E-B) strategy for soybean yield prediction using yield component traits. The mean performance is indicated with an + sign in each Figure.

https://doi.org/10.1371/journal.pone.0250665.g005

Table 1. Optimizing the number of nodes per plant (NP), the number of non-reproductive nodes per plant (NRNP), the number of reproductive nodes per plant (RNP), and the number of pods per plant (PP), the ratio of number of pods to number of nodes per plant (P/N) according to the E-B-GA for maximizing soybean seed yield.

Input variables					Predicted Yield (ton ha^{-1})
NP	NRNP	RNP	PP	P/N	
17.32	3.07	14.25	48.98	2.83	5.64

https://doi.org/10.1371/journal.pone.0250665.t001

determining the final seed yield [15, 81, 82]. In the current study, PP showed the highest linear correlation with the total seed yield. The direct impact of the number of pods per plant on the final soybean yield is also reported by Bastidas, Setiyono (83)]. Among all the tested yield component traits in this study, NP showed the second-highest linear correlation with total seed yield and showed a positive correlation with PP. Many studies reported that the variations in the number of nodes per plant is usually accounted for the changes in the number of pods per plant [81–84]. A negative correlation between the total soybean seed yield and NRNP was found in this study. It is in agreement with previous studies claimed that increasing the number of non-reproductive nodes decreases the reproductive potential of soybean seed yield [81, 85]. The results of linear correlation analyses in this study illustrated the importance of individual yield component traits in determining the total soybean seed yield.

Conventional statistical methods such as ANOVA and simple regression methods are typically recommended for small datasets with limited dimensions [86, 87]. However, soybean yield is a complex trait under controlled by different continuous variables called yield component traits. Therefore, more sophisticated methods are required for analyzing large data sets with high dimensions [88]. The successful uses of ML algorithms for analyzing big data with multi-collinearity among the variables have recently been reported in many plant species such as soybean [30], alfalfa [48], chrysanthemum [89], wheat [90], and lime [91]. The prediction performance of a given machine learning algorithm refers to the power of the model in predicting the values of a dependent variable when non-representative samples, or samples from a different population, are used as the test population [92]. The prediction performance of an ML algorithm is estimated using its R^2, RMSE, and MAE values [92–94]. In this study, the three common ML algorithms, RBF, MLP, and RF, were used to predict the soybean seed yield using its components and their prediction performance were estimated. RBF was found to be the most accurate ML algorithm for predicting the soybean seed yield from its component traits. In general, yield components in soybean are traits with low heritability that are influenced by environmental factors. The environmental factors can bring some levels of instability/noise in the results of all the ML analyses [95]. However, the structure of RBF (Fig 6) gives

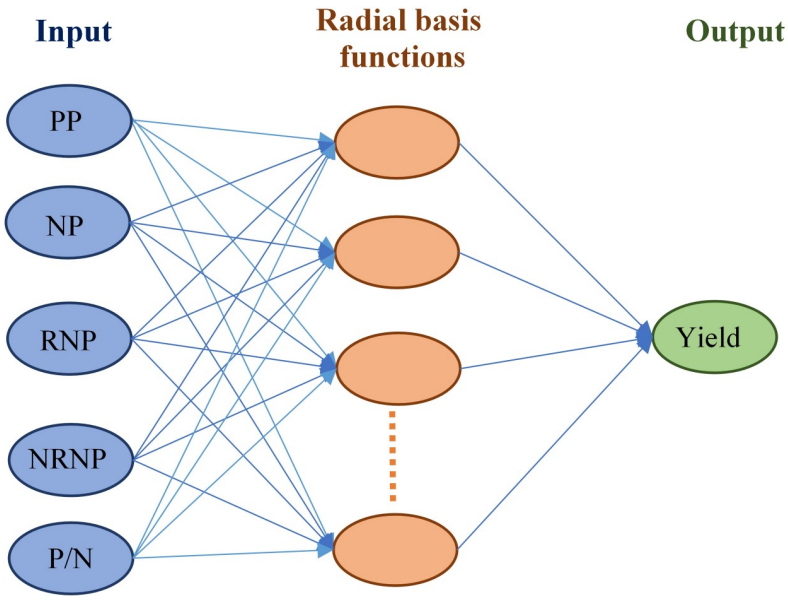

Fig 6. The schematic view of the Radial Basis Function (RBF) algorithm.

https://doi.org/10.1371/journal.pone.0250665.g006

some level of robustness against the adversarial noises, compared to other tested ML algorithms [96, 97]. The specific structure of RBF is the use of the transfer function of input variables to hidden layer name radial basis function [64, 98]. This function plays an important role in reducing noises of input variables resulted in more accurate prediction performance [99].

Although RF and MLP had the same R^2 values, the MAE and RMSE values were lower in RF. MLP, as one of the neural network algorithms, is highly susceptible to possible instabilities/noises caused by non-heritable factors. The MAE and RMSE values of this algorithm were the highest among all the tested ML algorithms that may indicate the sensitivity of the algorithm to noises. As a result, using this algorithm may not recommend for analyzing phenotypic data that are largely affected by environmental factors. RF with the R^2, MAE, and RMSE values of 0.80, 156.28 $kg.ha^{-1}$, and 194.75 $kg.ha^{-1}$, respectively, was selected as the second-best ML algorithm for predicting soybean seed yield in this study. Although the difference between RF and RBF in terms of R^2 values was not statistically significant (data are not shown), they were statistically different for their MAE and RMSE values. The performance of the RF algorithm relies on processing large dimensional data based on generalized error estimation [100, 101]. Also, there is no assumption requirement for RF about the distribution of data [102], and this algorithm can isolate outliers in a small region of the variable space resulted in acceptable performance against nonlinear environmental effects [102, 103].

In addition to individual comparison of the three tested ML algorithms, we developed a bagging ensemble model by combining the results of RBF, MLP, and RF in this study. Since the RBF had the highest performance in predicting the soybean seed yield, this algorithm was chosen as the metaClassifier for developing the E-B algorithm. Using the E-B model, a slight improvement was obtained in predicting the total soybean seed yield from its component traits. Diversity and sufficiency are two of the most important principles in selecting base learners for an ensemble model [67]. It means that the dependency among the used ML algorithms in the ensemble model should be minimized, while each based learner should have an acceptable predicting capability as well [104, 105]. Therefore, we selected different ML algorithms as the base learners for the E-B with different training data mechanisms. Also, the performance of the E-B was optimized by implementing RBF as the metaClassifier for this model. The effectiveness of using ensemble models was reported in different plant species such as chrysanthemum [106], sorghum [107], wheat [108], alfalfa [67], and brassicas [109]. This study demonstrates the benefit of using the RBF-based E-B approach to improve the soybean yield prediction accuracy using yield components.

Selecting high-yielding lines has always been one of the major goals in plant breeding programs that can be performed using either direct or indirect selection approaches [110]. Selecting superior genotypes based on the yield component traits can be considered as an indirect method. An analytical breeding strategy is an alternate breeding approach that is focused on the improvement of components of complex traits such as plant growth and development rates or yield components [111] rather than the traits *per se*. This strategy can improve genetic yield potential in varieties by setting up selection criteria on yield components and making the selection more efficient [112]. In order to move toward analytical breeding, it would be important to have the optimized level of each yield component traits in target populations. Genetic algorithm is commonly used in finding optimized solutions by searching problems through biological parameters such as selection, crossover, and mutation [53, 113]. After selecting E-B as the combined algorithm with the highest prediction ability in this study, GA was linked into this algorithm to estimate the optimum values of the yield component traits (Table 1). The successfulness of using the GA algorithm for estimating optimized solutions was reported previously in plant tissue culture [89], plant physiology [114], and remote sensing [115].

Conclusion

Efficient breeding approaches for improving the genetic potential of complex traits such as yield in soybean can be developed based upon secondary traits that govern the final performance of the complex traits. Thus, in order to increase the genetic yield potential in soybean, breeders can select soybean genotypes with improved yield component traits. However, measuring yield components in a large soybean breeding program is time-consuming and labor-intensive, which limit the implication of this approach in cultivar development programs. The main objective of this study was to evaluate the potential use of yield component traits for estimating final seed yield in soybean using different ML and E-B algorithms, which in turn can be used by breeders for selecting parental lines and designing promising crosses for developing cultivars with improved genetic yield potential. The results of the current study showed that RBF is a reliable solo ML algorithm for predicting the soybean seed yield from its component traits. However, an E-B algorithm that was built by combining the three ML and using RBF as its metaClassifier outperformed all individual ML algorithms and, therefore, it is recommended for predicting the soybean seed yield exploiting yield component traits. In the current study for the first time, we coupled E-B algorithm with GA in order to estimate optimum values of yield component traits in a theoretical genotype in which the yield is maximized using the real field data. The results seem to be promising but are recommended to be evaluated in new and independent breeding populations before using in cultivar development programs for selecting high-yielding potential genotypes. This information can be also used, in combined with molecular marker technology, for developing genomic-based toolkits that can be used for selecting genotypes with improved genetic yield potential at early generations.

Supporting information

S1 Table. Analysis performance of Random Forest (RF), Multilayer Perceptron (MLP), and Radial Basis Function (RBF) algorithms, and the Ensemble-Bagging (E-B) strategy for soybean yield prediction using yield component traits.
(DOCX)

Acknowledgments

The authors are grateful to the past and current members of Eskandari laboratory at the University of Guelph, Ridgetown Campus, including Dr. Sepideh Torabi, Mr. Bryan Stirling, Mr. John Kobler, and Mr. Robert Brandt for their technical support. We would also like to thank Mrs. Maryam Vazin for her assistance with the field data collection.

Author Contributions

Conceptualization: Milad Eskandari.

Formal analysis: Mohsen Yoosefzadeh-Najafabadi.

Funding acquisition: Milad Eskandari.

Methodology: Mohsen Yoosefzadeh-Najafabadi.

Supervision: Milad Eskandari.

Validation: Dan Tulpan, Milad Eskandari.

Writing – original draft: Mohsen Yoosefzadeh-Najafabadi.

Writing – review & editing: Dan Tulpan, Milad Eskandari.

References

1. Hashiguchi A, Komatsu S. Chapter 6—Proteomics of Soybean Plants. In: Colgrave ML, editor. Proteomics in Food Science: Academic Press; 2017. p. 89–105.
2. Miransari M. 11—Soybean production and N fertilization. In: Miransari M, editor. Abiotic and Biotic Stresses in Soybean Production. San Diego: Academic Press; 2016. p. 241–60.
3. Wilson RF. The role of genomics and biotechnology in achieving global food security for high-oleic vegetable oil. Journal of oleo science. 2012; 61(7):357–67. https://doi.org/10.5650/jos.61.357 PMID: 22790166
4. Ramasubramanian V, Beavis WD. Factors affecting Response to Recurrent Genomic Selection in Soybeans. BioRxiv. 2020.
5. Rebetzke G, Jimenez-Berni J, Fischer R, Deery D, Smith D. High-throughput phenotyping to enhance the use of crop genetic resources. Plant Science. 2019; 282:40–8. https://doi.org/10.1016/j.plantsci.2018.06.017 PMID: 31003610
6. Yuan J, Njiti V, Meksem K, Iqbal M, Triwitayakorn K, Kassem MA, et al. Quantitative trait loci in two soybean recombinant inbred line populations segregating for yield and disease resistance. Crop science. 2002; 42(1):271–7. https://doi.org/10.2135/cropsci2002.2710 PMID: 11756285
7. Tester M, Langridge P. Breeding technologies to increase crop production in a changing world. Science. 2010; 327(5967):818–22. https://doi.org/10.1126/science.1183700 PMID: 20150489
8. Araus JL, Cairns JE. Field high-throughput phenotyping: the new crop breeding frontier. Trends in plant science. 2014; 19(1):52–61. https://doi.org/10.1016/j.tplants.2013.09.008 PMID: 24139902
9. Qiu R, Wei S, Zhang M, Li H, Sun H, Liu G, et al. Sensors for measuring plant phenotyping: A review. International Journal of Agricultural and Biological Engineering. 2018; 11(2):1–17.
10. Kenga R, Tenkouano A, Gupta S, Alabi S. Genetic and phenotypic association between yield components in hybrid sorghum (Sorghum bicolor (L.) Moench) populations. Euphytica. 2006; 150(3):319–26.
11. Robbins MD, Staub JE. Comparative analysis of marker-assisted and phenotypic selection for yield components in cucumber. Theoretical and applied genetics. 2009; 119(4):621–34. https://doi.org/10.1007/s00122-009-1072-8 PMID: 19484431
12. Richards R. Selectable traits to increase crop photosynthesis and yield of grain crops. Journal of experimental botany. 2000; 51(suppl_1):447–58. https://doi.org/10.1093/jexbot/51.suppl_1.447 PMID: 10938853
13. Specht J, Hume D, Kumudini S. Soybean yield potential—a genetic and physiological perspective. Crop science. 1999; 39(6):1560–70.
14. Kumudini S, Hume DJ, Chu G. Genetic improvement in short season soybeans. Crop science. 2001; 41(2):391–8.
15. Xavier A, Rainey KM. Quantitative Genomic Dissection of Soybean Yield Components. G3: Genes, Genomes, Genetics. 2020; 10(2):665–75.
16. Sah R, Chakraborty M, Prasad K, Pandit M, Tudu V, Chakravarty M, et al. Impact of water deficit stress in maize: Phenology and yield components. Scientific reports. 2020; 10(1):1–15. https://doi.org/10.1038/s41598-019-56847-4 PMID: 31913322
17. Majhi PK, Mogali SC, Abhisheka L. Genetic variability, heritability, genetic advance and correlation studies for seed yield and yield components in early segregating lines (F3) of greengram [Vigna radiata (L.) Wilczek]. International Journal of Chemical Studies. 2020; 8(4):1283–8.
18. Jiang Y, Lindsay DL, Davis AR, Wang Z, MacLean DE, Warkentin TD, et al. Impact of heat stress on pod-based yield components in field pea (Pisum sativum L.). Journal of Agronomy and Crop Science. 2020; 206(1):76–89.
19. Jin J, Liu X, Wang G, Mi L, Shen Z, Chen X, et al. Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. Field Crops Research. 2010; 115(1):116–23.
20. Liu X, Jin J, Herbert S, Zhang Q, Wang G. Yield components, dry matter, LAI and LAD of soybeans in Northeast China. Field Crops Research. 2005; 93(1):85–93.
21. Ma B, Dwyer LM, Costa C, Cober ER, Morrison MJ. Early prediction of soybean yield from canopy reflectance measurements. Agronomy Journal. 2001; 93(6):1227–34.
22. Zeng Q, Huang H, Pei X, Wong S, Gao M. Rule extraction from an optimized neural network for traffic crash frequency modeling. Accident Analysis & Prevention. 2016; 97:87–95. https://doi.org/10.1016/j.aap.2016.08.017 PMID: 27591417
23. Zeng Q, Huang H, Pei X, Wong S. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Analytic methods in accident research. 2016; 10:12–25.

24. Maganathan T, Senthilkumar S, Balakrishnan V, editors. Machine Learning and Data Analytics for Environmental Science: A Review, Prospects and Challenges. IOP Conference Series: Materials Science and Engineering; 2020: IOP Publishing.

25. Sha W, Guo Y, Yuan Q, Tang S, Zhang X, Lu S, et al. Artificial Intelligence to Power the Future of Materials Science and Engineering. Advanced Intelligent Systems. 2020; 2(4):1900143.

26. Lee S, Liang X, Woods M, Reiner AS, Concannon P, Bernstein L, et al. Machine learning on genome-wide association studies to predict the risk of radiation-associated contralateral breast cancer in the WECARE Study. PloS one. 2020; 15(2):e0226157. https://doi.org/10.1371/journal.pone.0226157 PMID: 32106268

27. Duan K-B, Rajapakse JC, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE transactions on nanobioscience. 2005; 4(3):228–34. https://doi.org/10.1109/tnb.2005.853657 PMID: 16220686

28. Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, et al. Machine learning in genome-wide association studies. Genetic epidemiology. 2009; 33(S1):S51–S7. https://doi.org/10.1002/gepi.20473 PMID: 19924717

29. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: A review. Sensors. 2018; 18(8):2674. https://doi.org/10.3390/s18082674 PMID: 30110960

30. Yoosefzadeh-Najafabadi M, Earl HJ, Tulpan D, Sulik J, Eskandari M. Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean. Frontiers in Plant Science. 2021; 11(2169). https://doi.org/10.3389/fpls.2020.624273. PMID: 33510761

31. Crane-Droesch A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters. 2018; 13(11):114003.

32. McQueen RJ, Garner SR, Nevill-Manning CG, Witten IH. Applying machine learning to agricultural data. Computers and electronics in agriculture. 1995; 12(4):275–93.

33. Niazian M, Niedbała G. Machine Learning for Plant Breeding and Biotechnology. Agriculture. 2020; 10 (10):436. https://doi.org/10.3390/agriculture10100436

34. Zhang C, Pan X, Li H, Gardiner A, Sargent I, Hare J, et al. A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. ISPRS Journal of Photogrammetry and Remote Sensing. 2018; 140:133–44.

35. Wang Y, Gao W. Prediction of the water content of biodiesel using ANN-MLP: An environmental application. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects. 2018; 40(8):987–93.

36. Yilmaz I, Kaynar O. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. Expert systems with applications. 2011; 38(5):5958–66.

37. Bhojani SH, Bhatt N. Wheat crop yield prediction using new activation functions in neural network. Neural Computing and Applications. 2020:1–11.

38. Deore B, Kyatham A, Narkhede S, editors. A novel approach to ensemble MLP and random forest for network security. ITM Web of Conferences; 2020: EDP Sciences.

39. Araghinejad S. Data-driven modeling: using MATLAB® in water resources and environmental engineering: Springer Science & Business Media; 2013.

40. Hesami M, Naderi R, Tohidfar M. Modeling and Optimizing Medium Composition for Shoot Regeneration of Chrysanthemum via Radial Basis Function-Non-dominated Sorting Genetic Algorithm-II (RBF-NSGAII). Scientific Reports. 2019; 9(1):1–11. https://doi.org/10.1038/s41598-018-37186-2 PMID: 30626917

41. Heddam S, Bermad A, Dechemi N. Applications of radial-basis function and generalized regression neural networks for modeling of coagulant dosage in a drinking water-treatment plant: comparative study. Journal of Environmental Engineering. 2011; 137(12):1209–14.

42. Chouhan SS, Kaul A, Singh UP, Jain S. Bacterial foraging optimization based radial basis function neural network (BRBFNN) for identification and classification of plant leaf diseases: An automatic approach towards plant pathology. IEEE Access. 2018; 6:8852–63.

43. De Castro AI, Torres-Sánchez J, Peña JM, Jiménez-Brenes FM, Csillik O, López-Granados F. An automatic random forest-OBIA algorithm for early weed mapping between and within crop rows using UAV imagery. Remote Sensing. 2018; 10(2):285.

44. Alsahaf A, Azzopardi G, Ducro B, Hanenberg E, Veerkamp RF, Petkov N. Prediction of slaughter age in pigs and assessment of the predictive value of phenotypic and genetic information using random forest. Journal of animal science. 2018; 96(12):4935–43. https://doi.org/10.1093/jas/sky359 PMID: 30239725

45. Tulpan D. 311 A brief overview, comparison and practical applications of machine learning models. Journal of Animal Science. 2020; 98(Supplement_4):44–5.

46. Acharjee A, Prentice P, Acerini C, Smith J, Hughes IA, Ong K, et al. The translation of lipid profiles to nutritional biomarkers in the study of infant metabolism. Metabolomics. 2017; 13(3):25. https://doi.org/10.1007/s11306-017-1166-2 PMID: 28190990

47. Pal M. Random forest classifier for remote sensing classification. International Journal of Remote Sensing. 2005; 26(1):217–22.

48. Feng L, Zhang Z, Ma Y, Du Q, Williams P, Drewry J, et al. Alfalfa Yield Prediction Using UAV-Based Hyperspectral Imagery and Ensemble Learning. Remote Sensing. 2020; 12(12):2028.

49. Dietterich TG, editor Ensemble methods in machine learning. International workshop on multiple classifier systems; 2000: Springer.

50. Seni G, Elder JF. Ensemble methods in data mining: improving accuracy through combining predictions. Synthesis lectures on data mining and knowledge discovery. 2010; 2(1):1–126.

51. Aanonsen SI, Nævdal G, Oliver DS, Reynolds AC, Vallès B. The ensemble Kalman filter in reservoir engineering—a review. Spe Journal. 2009; 14(03):393–412.

52. Wang H, Khoshgoftaar TM, Napolitano A. Software measurement data reduction using ensemble techniques. Neurocomputing. 2012; 92:124–32.

53. Hesami M, Jones AMP. Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture. Applied Microbiology and Biotechnology. 2020. https://doi.org/10.1007/s00253-020-10888-2 PMID: 32984921

54. Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence: MIT press; 1992.

55. Yun Y, Chuluunsukh A, Gen M. Sustainable closed-loop supply chain design problem: A hybrid genetic algorithm approach. Mathematics. 2020; 8(1):84.

56. Hesami M, Naderi R, Tohidfar M, Yoosefzadeh-Najafabadi M. 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. 2020; 16(1):1–15. https://doi.org/10.1186/s13007-020-00655-9 PMID: 32817755

57. Hesami M, Naderi R, Tohidfar M. Introducing a hybrid artificial intelligence method for high-throughput modeling and optimizing plant tissue culture processes: the establishment of a new embryogenesis medium for chrysanthemum, as a case study. Applied Microbiology and Biotechnology. 2020; 104(23):10249–63. https://doi.org/10.1007/s00253-020-10978-1 PMID: 33119796

58. Stroup W, Mulitze D. Nearest neighbor adjusted best linear unbiased prediction. The American Statistician. 1991; 45(3):194–200.

59. Katsileros A, Drosou K, Koukouvinos C. Evaluation of nearest neighbor methods in wheat genotype experiments. Commun Biom Crop Sci. 2015; 10:115–23.

60. Robinson GK. That BLUP is a good thing: the estimation of random effects. Statistical science. 1991; 6(1):15–32.

61. Rossel RAV. ParLeS: Software for chemometric analysis of spectroscopic data. Chemometrics intelligent laboratory systems. 2008; 90(1):72–83.

62. Geetha M. Forecasting the Crop Yield Production in Trichy District Using Fuzzy C-Means Algorithm and Multilayer Perceptron (MLP). International Journal of Knowledge and Systems Science (IJKSS). 2020; 11(3):83–98.

63. Gardner MW, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment. 1998; 32(14–15):2627–36.

64. Orr MJ. Introduction to radial basis function networks. Technical Report, center for cognitive science, University of Edinburgh; 1996.

65. Wilamowski BM, Jaeger RC, editors. Implementation of RBF type networks by MLP networks. Proceedings of International Conference on Neural Networks (ICNN'96); 1996: IEEE.

66. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002; 2(3):18–22.

67. Feng L, Li Y, Wang Y, Du Q. Estimating hourly and continuous ground-level PM2. 5 concentrations using an ensemble learning algorithm: The ST-stacking model. Atmospheric Environment. 2020; 223:117242.

68. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009; 11(1):10–8.

69. Wang S-C. Genetic algorithm. Interdisciplinary Computing in Java Programming: Springer; 2003. p. 101–16.

70. Siegmann B, Jarmer T. Comparison of different regression models and validation techniques for the assessment of wheat leaf area index from hyperspectral data. International journal of remote sensing. 2015; 36(18):4519–34.

71. Farifteh J, Van der Meer F, Atzberger C, Carranza EJM. Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN). Remote Sensing of Environment. 2007; 110(1):59–78. https://doi.org/10.1016/j.rse.2007.02.005.

72. Cacuci DG, Ionescu-Bujor M, Navon IM. Sensitivity and uncertainty analysis, volume II: applications to large-scale systems: CRC press; 2005.

73. Taylor J. Introduction to error analysis, the study of uncertainties in physical measurements1997.

74. Chang W, Wickham H. ggvis: Interactive Grammar of Graphics. R package version0. 2016; 4.

75. Wickham H, Chang W, Wickham MH. Package 'ggplot2'. Create Elegant Data Visualisations Using the Grammar of Graphics Version. 2016; 2(1):1–189.

76. Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, et al. Package 'mass'. Cran R. 2013;538.

77. Ciampitti IA, Vyn TJ. Physiological perspectives of changes over time in maize yield dependency on nitrogen uptake and associated nitrogen efficiencies: A review. Field Crops Research. 2012; 133:48–67.

78. Cao S, Xu D, Hanif M, Xia X, He Z. Genetic architecture underpinning yield component traits in wheat. Theoretical and Applied Genetics. 2020:1–13. https://doi.org/10.1007/s00122-020-03562-8 PMID: 32062676

79. O'Connor K, Hayes B, Topp B. Prospects for increasing yield in macadamia using component traits and genomics. Tree genetics & genomes. 2018; 14(1):7.

80. Dutamo D, Alamerew S, Eticha F, Assefa E. Genetic variability in bread wheat (Triticum aestivum L.) germplasm for yield and yield component traits. Journal of Biology, Agriculture and Healthcare. 2015; 5(17):140–7.

81. Egli D. The relationship between the number of nodes and pods in soybean communities. Crop Science. 2013; 53(4):1668–76.

82. Egli D. Flowering, pod set and reproductive success in soya bean. Journal of Agronomy and crop science. 2005; 191(4):283–91.

83. Bastidas A, Setiyono T, Dobermann A, Cassman KG, Elmore RW, Graef GL, et al. Soybean sowing date: The vegetative, reproductive, and agronomic impacts. Crop Science. 2008; 48(2):727–40.

84. Wei MCF, Molin JP. Soybean Yield Estimation and Its Components: A Linear Regression Approach. Agriculture. 2020; 10(8):348.

85. Du Y, Zhao Q, Li S, Yao X, Xie F, Zhao M. Shoot/root interactions affect soybean photosynthetic traits and yield formation: a case study of grafting with record-yield cultivars. Frontiers in plant science. 2019; 10:445. https://doi.org/10.3389/fpls.2019.00445 PMID: 31024606

86. Rutherford A. Introducing ANOVA and ANCOVA: a GLM approach: Sage; 2001.

87. Homack SR. Understanding What ANOVA Post Hoc Tests Are, Really. 2001.

88. Vapnik V. The nature of statistical learning theory: Springer science & business media; 2013.

89. Hesami M, Naderi R, Tohidfar M, Yoosefzadeh-Najafabadi M. Application of Adaptive Neuro-Fuzzy Inference System-Non-dominated Sorting Genetic Algorithm-II (ANFIS-NSGAII) for Modeling and Optimizing Somatic Embryogenesis of Chrysanthemum. Frontiers in plant science. 2019; 10:869. https://doi.org/10.3389/fpls.2019.00869 PMID: 31333705

90. Montesinos-López OA, Montesinos-López A, Crossa J, de los Campos G, Alvarado G, Suchismita M, et al. Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. Plant methods. 2017; 13(1):4. https://doi.org/10.1186/s13007-016-0154-2 PMID: 28053649

91. Jafari M, Shahsavar A. The application of artificial neural networks in modeling and predicting the effects of melatonin on morphological responses of citrus to drought stress. Plos one. 2020; 15(10): e0240427. https://doi.org/10.1371/journal.pone.0240427 PMID: 33052940

92. Rocha A, Groen T, Skidmore A, Darvishzadeh R, Willemen L. Machine learning using hyperspectral data inaccurately predicts plant traits under spatial dependency. Remote sensing. 2018; 10(8):1263.

93. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: Springer; 2013.

94. Kuhn M, Johnson K. Applied predictive modeling: Springer; 2013.

95. Xavier A, Muir WM, Rainey KM. Assessing predictive properties of genome-wide selection in soybeans. G3: Genes, Genomes, Genetics. 2016; 6(8):2611–6.

96. Chenou J, Hsieh G, Fields T, editors. Radial Basis Function Network: Its Robustness and Ability to Mitigate Adversarial Examples. 2019 International Conference on Computational Science and Computational Intelligence (CSCI); 2019: IEEE.

97. Langenberg P, Balda E, Behboodi A, Mathar R, editors. On the Robustness of Support Vector Machines against Adversarial Examples. 2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS); 2019: IEEE.

98. Bawazeer SA, Baakeem SS, Mohamad AA. New Approach for Radial Basis Function Based on Partition of Unity of Taylor Series Expansion with Respect to Shape Parameter. Algorithms. 2021; 14(1):1.

99. Jiang Y, Wei G, Sun X, Zhang Y, editors. Predicting Noisy Data with an Improvement RBF Neural Network for Surrogate Models. 2016 4th International Conference on Machinery, Materials and Computing Technology; 2016: Atlantis Press.

100. Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews. 2015; 71:804–18.

101. Zhang P, Yin Z-Y, Jin Y-F, Chan TH. A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest. Engineering Geology. 2020; 265:105328.

102. Melesse AM, Khosravi K, Tiefenbacher JP, Heddam S, Kim S, Mosavi A, et al. River Water Salinity Prediction Using Hybrid Machine Learning Models. Water. 2020; 12(10):2951.

103. De'ath G, Fabricius KE. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology. 2000; 81(11):3178–92.

104. Araya DB, Grolinger K, ElYamany HF, Capretz MA, Bitsuamlak G. An ensemble learning framework for anomaly detection in building energy consumption. Energy and Buildings. 2017; 144:191–206.

105. Zhang Z, Jin Y, Chen B, Brown P. California almond yield prediction at the orchard level with a machine learning approach. Frontiers in plant science. 2019; 10:809. https://doi.org/10.3389/fpls.2019.00809 PMID: 31379888

106. Hesami M, Alizadeh M, Naderi R, Tohidfar M. Forecasting and optimizing Agrobacterium-mediated genetic transformation via ensemble model-fruit fly optimization algorithm: A data mining approach using chrysanthemum databases. Plos One. 2020; 15(9):e0239901. https://doi.org/10.1371/journal.pone.0239901 PMID: 32997694

107. Kosmowski F, Worku T. Evaluation of a miniaturized NIR spectrometer for cultivar identification: The case of barley, chickpea and sorghum in Ethiopia. PloS one. 2018; 13(3):e0193620. https://doi.org/10.1371/journal.pone.0193620 PMID: 29561868

108. Tian Y, Zhao C, Lu S, Guo X. Multiple classifier combination for recognition of wheat leaf diseases. Intelligent Automation & Soft Computing. 2011; 17(5):519–29.

109. Qi Y. Random forest for bioinformatics. Ensemble machine learning: Springer; 2012. p. 307–23.

110. Slinkard A, Solh M, Vandenberg A. Breeding for yield: the direct approach. Linking Research and Marketing Opportunities for Pulses in the 21st Century: Springer; 2000. p. 183–90.

111. Acevedo E, Aleppo S. Improvement of winter cereal crops in Mediterranean environments: Use of yield, morphological and physiological traits. Breeding for drought resistance in wheat. 1991; 12:188.

112. Reynolds M. Application of physiology in wheat breeding: Cimmyt; 2001.

113. Dasgupta K, Mandal B, Dutta P, Mandal JK, Dam S. A genetic algorithm (ga) based load balancing strategy for cloud computing. Procedia Technology. 2013; 10:340–7.

114. Halim AH, Ismail I, editors. Nonlinear plant modeling using neuro-fuzzy system with Tree Physiology Optimization. 2013 IEEE Student Conference on Research and Developement; 2013: IEEE.

115. Wu Y, Miao Q, Ma W, Gong M, Wang S. PSOSAC: particle swarm optimization sample consensus algorithm for remote sensing image registration. IEEE Geoscience and Remote Sensing Letters. 2017; 15(2):242–6.