scientific reports

Check for updates

OPEN

Machine learning models for predicting morphological traits and optimizing genotype and planting date in roselle (*Hibiscus Sabdariffa* L.)

Fazilat Fakhrzad[1]✉, Warqaa Muhammed ShariffAl-Sheikh[2], Mohammed M. Mohammed[3] & Heidar Meftahizadeh[4]✉

Accurate prediction and optimization of morphological traits in Roselle are essential for enhancing crop productivity and adaptability to diverse environments. In the present study, a machine learning framework was developed using Random Forest and Multi-layer Perceptron algorithms to model and predict key morphological traits, branch number, growth period, boll number, and seed number per plant, based on genotype and planting date. The dataset was generated from a field experiment involving ten Roselle genotypes and five planting dates. Both RF and MLP exhibited robust predictive capabilities; however, RF (R^2 = 0.84) demonstrated superior performance compared to MLP (R^2 = 0.80), underscoring its efficacy in capturing the nonlinear genotype-by-environment interactions. Permutation-based feature importance analysis further revealed that planting date had a more significant impact on trait variation than genotype. To identify optimal combinations of genotype and planting date for maximizing morphological traits, the RF model was integrated with the Non-dominated Sorting Genetic Algorithm II (NSGA-II). According to the RF–NSGA-II optimization results, the optimal values, including 26 branches per plant, a growth period of 176 days, 116 bolls per plant, and 1517 seed numbers per plant, were achieved with the *Qaleganj* genotype planted on May 5. Collectively, these findings highlight the potential of integrating machine learning and evolutionary optimization algorithms as powerful computational tools for crop improvement and agronomic decision-making.

Keywords Machine learning techniques, Prediction, Optimization algorithm, Multi-layer perceptron, Random forest

Roselle (*Hibiscus sabdariffa* L.) is an annual herbaceous plant belonging to the Malvaceae family that is widely cultivated in tropical and subtropical regions due to its high adaptability and drought tolerance. Various parts of the plant, including the flowers, stem fiber, leaves, seeds, fruits, and roots, are multi-purpose in the food and medicinal industries. Among various parts of the plant, the bright red fleshy calyx holds the most significant economic value, which contains vitamin C, iron, beta-carotene, anthocyanins, and phenolic compounds[1]. Additionally, roselle seeds contain high levels of protein, vitamin E, and unsaturated fatty acids such as oleic and linoleic acids[2]. In recent years, global demand for roselle and its processed products has increased steadily, particularly in the health food and natural product markets[3].

Plant growth and development are influenced by a variety of environmental and genetic factors. Among these, the selection of suitable genotypes and the determination of optimal planting dates are critical agronomic decisions that directly impact plant performance[3]. Genotype selection allows breeders to identify plant varieties with favorable characteristics such as early maturity, high yield, and resistance to environmental stress[4]. The

[1]Department of Horticultural Science, College of Agriculture, Shiraz University, Shiraz, Iran. [2]Faculty of Basic Science Branch, Faculty of Dentistry, University of Al- Qadisiyah, Diwaniyah, Iraq. [3]Department of Horticulture and landscape gardening, Colleague of agricultural engineering sciences, University of Baghdad, Baghdad, Iraq. [4]Department of Horticultural Science, Faculty of Agriculture & Natural Resources, Ardakan University, Ardakan, Iran. ✉email: ffakhrzad@Shirazu.ac.ir; Hmeftahizade@ardakan.ac.ir

results of research conducted on different genotypes of roselle indicate that these traits have high heritability and are largely influenced by genetic factors[5]. It has been reported that there are significant differences among genotypes in terms of flower and leaf production and nutrient density[6]. Multiple studies have shown that delayed sowing can significantly reduce crop yield, including plant height, number of bolls, biomass, calyx, and seed yield. Also, early planting leads to a longer growth period and more desirable growth traits[7]. Planting date has also been reported to significantly affect the growth and yield of Roselle, with July sowing leading to earlier flowering, a higher number of flower buds, and longer calyxes[8]. Furthermore, planting on May 15 has been associated with the best performance in terms of vegetative growth, such as plant height, stem diameter, number of branches, and number of fruits, and yield components including fresh and dry weights of shoots, calyces, and seeds[7].

However, the interaction between genotype and planting date is complex and non-linear, requiring advanced modeling approaches to accurately predict outcomes and make informed recommendations. Traditional statistical methods, while useful, often fall short in capturing the non-linear interactions and high-dimensional relationships among multiple plant traits and environmental variables. In contrast, machine learning (ML) techniques offer powerful alternatives by learning patterns from historical data without relying on predefined models[9]. Among these, Random Forest (RF) and Multi-layer Perceptron (MLP) are popular ML algorithms used in agricultural modeling due to their flexibility and robustness. MLP, as a type of feed-forward artificial neural network (ANN), is particularly known for its ability to model highly nonlinear functions due to its deep architecture composed of interconnected neurons and hidden layers[9]. RF is a widely used ensemble machine learning algorithm that offers a balance between simplicity, accuracy, and robustness. It has gained popularity across various domains, including biological and agricultural research, due to its reliable predictive capabilities and flexibility in handling both classification and regression tasks. Structurally, an RF model consists of an ensemble of decision trees, each trained on a random subset of the original dataset using the bootstrap aggregation (bagging) technique. At each node of every tree, a random subset of features is considered for splitting, which helps to reduce correlation between trees and improve generalization. The final prediction of the RF model is obtained by aggregating the outputs of all individual trees through majority voting for classification or averaging for regression. This ensemble approach enhances accuracy, reduces variance, and mitigates overfitting, making RF particularly suitable for modeling complex, noisy, high-dimensional, and imbalanced datasets[10,11].

In biosystems, prediction alone is insufficient, and decision-making often involves balancing multiple conflicting objectives. Among the various optimization approaches, evolutionary algorithms, particularly the Non-dominated Sorting Genetic Algorithm II (NSGA-II), address this challenge by identifying a set of optimal solutions known as the Pareto front[12,13]. NSGA-II, first introduced by Deb et al.[14]has become one of the most popular and widely applied algorithms for solving multi-objective optimization problems due to its simplicity, efficiency, and ability to maintain solution diversity through crowding distance calculation. Unlike single-objective genetic algorithms, NSGA-II can handle multiple objectives simultaneously and identify a set of Pareto-optimal solutions, representing the best trade-offs between conflicting goals. These solutions offer decision-makers a spectrum of options, enabling tailored strategies based on specific priorities and constraints[12,14]. The integration of ML models with NSGA-II provides a synergistic framework that leverages the strengths of both approaches[9,15]. The integration of ML with optimization algorithms has been successfully applied in various domains of plant science. For instance, ML–NSGA-II frameworks have been employed to optimize growth regulator combinations in plant tissue culture systems[9,11] enhance secondary metabolite production in plants[16] predict plant responses to drought and salinity stress[15,17] and estimate crop yield and agronomic traits[18]. Recently, the efficiency of the ANN-NSGA II model was used to predict and optimize pomegranate morphological traits under salinity and drought stress influenced by γ-aminobutyric acid. This approach helped identify optimal treatment conditions and provided insights into improving stress tolerance[15]. Moreover, in a study, the combination of ML algorithms and genetic optimization was employed to model and optimize soybean yield based on its component traits. This integrated approach provided a better understanding of the relationships between yield and its morphological components. It can be effectively used in selecting parental lines and designing crosses aimed at improving the genetic yield potential of soybean cultivars[19]. These applications underscore the growing relevance of data-driven modeling approaches in addressing complex, multi-objective problems in plant biology.

The present study aims to develop an integrated framework for predicting and optimizing the main agronomic traits of Roselle, focusing on four key traits, including the number of branches, growth period, number of bolls, and seeds per plant. Using a comprehensive dataset collected from a field experiment, we trained and compared the performance of two ML models, followed by the application of NSGA-II for multi-objective optimization. While ML models ensure the accurate prediction of morphological traits based on genotype and planting date, NSGA-II enables the optimization of these traits under a multi-objective setting, assisting in the selection of the best genotype–planning date combinations that meet multiple morphological targets simultaneously. The novelty of this study lies in its data-driven approach to simultaneously address traits prediction and optimal genotype-date selection. Specifically, the objectives are to:

(1) examine how different planting dates and genotype combinations influence roselle's morphological performance,

(2) identify stable and high-performing genotypes across diverse environmental conditions,

(3) develop predictive ML models that can accurately forecast key traits based on genotype and planting date, and.

(4) employ cutting-edge multi-objective optimization algorithms to recommend the most effective genotype–planting date combinations that maximize crop yield. The insights derived from this model can inform breeding programs and cultivation planning, contributing to sustainable and high-yielding Roselle production systems.

Materials and methods

Plant materials and experimental design

The field experiment was conducted in Dalgan (27° 28′ N, 59° 27′ E; 389 m a.s.l.), located in Sistan and Baluchestan province, southeast Iran. The region has a hot-arid climate, with minimal rainfall and high summer temperatures. A detailed climatic condition has been illustrated in Fig. S1. The soil texture was loam with a pH of 7.6 and EC of 2.12 dS m^{-1}. A detailed chemical and physical soil characteristics has been mentioned in Table S1. After land preparation, before sowing, 25 kg/ha of triple superphosphate fertilizer was added to the soil and mixed in by a light disc. Nitrogen fertilization was performed at the six-leaf stage by applying urea as a foliar solution at a rate of 30 kg ha^{-1}. The study employed a factorial experimental design based on a randomized complete block design (RCBD) with three replications. The experimental treatments consisted of ten roselle genotypes including eight native accessions *Jiroft*, *Dalgan*, *Bampoor*, *Iranshahr*, *Nikshahr*, *Roodbar*, *Saravan*, *and Qaleganj* were collected from different agro-ecological zones of Iran, along with two exotic landraces: *HA* (originating from Ghana) and *HS-24* (from Bangladesh), both supplied by the Jiroft Agricultural Research Station. The ten genotypes were sown under five different planting dates: March 6, April 6, May 5, June 5, and July 1. Each genotype × planting date combination was replicated three times, resulting in a total of 150 experimental units.

Morphological trait measurement

Morphological trait measurements were performed at the physiological maturity stage, defined as the point when approximately 70% of the seeds within each flower capsule had browned and the sepals had reached their maximum development. The following traits were recorded for each genotype and planting date treatment: plant height, number of branches per plant, fruit length, number of bolls per plant, fresh sepal weight per plant, number of seeds per capsule, seed weight per plant, 1000-seed weight, biomass yield, harvest index, calyx yield, and growth period (life cycle duration). The morphological variation of sepals among the studied genotypes is shown in Supplementary Fig. S2. This visual comparison clearly highlights the differences in sepal size, shape, and pigmentation among the studied accessions.

Data Pre-processing, and statistical analyses

The dataset included ten genotypes and five planting dates, with four primary output traits: number of branches, growth period, number of bolls, and seed yield per plant. Input features were encoded using one-hot encoding, and the output variables were normalized using z-score standardization. Outlier detection and removal were conducted to enhance data quality using two complementary techniques: the interquartile range (IQR) method and the Z-score method. In the IQR method, any data point lying outside the range of $Q1 - 1.5 \times IQR$ or $Q3 + 1.5 \times IQR$ was considered an outlier. In the Z-score method, samples with absolute Z-score values greater than 3.0 were flagged and removed. Prior to model training, a two-step statistical analysis was conducted to evaluate the relevance of each phenotypic trait with respect to the input variables (genotype and planting date). First, Pearson correlation analysis was conducted using Python (version 3.11.12) to assess the linear relationship between inputs and each target. Subsequently, a two-way analysis of variance (ANOVA) was conducted to determine the statistical significance of the main factors (genotype and planting date) and their interaction on each trait. The two-way ANOVA was performed using the statsmodels Python library (version 0.14.0), specifically employing the ols function from the statsmodels. formula.api module to fit the linear model and the anova_lm function from statsmodels.api to compute the ANOVA table based on Type II sum of squares. In Type II ANOVA, each main factor (such as genotype or planting date) is evaluated separately while considering the influence of the other factors, but without including their interaction. This approach makes it easier to understand the individual effect of each factor and leads to clearer and more straightforward interpretation of the results. The dataset was split into 80% (120 samples) for training and 20% (30 samples) for testing.

Hyperparameter optimization in ML models

In machine learning, prior optimization and tuning of model hyperparameters are critical steps that significantly affect predictive performance and generalization ability. In this study, a structured grid search algorithm, Grid Search Cross Validation (GridSearchCV) was used for hyperparameter tuning of two ML models, MLP and RF, in combination with 10-fold cross-validation. During the K-fold cross-validation process ($K = 10$), the dataset was randomly partitioned into 10 subsets. Each subset served once as a validation set while the remaining subsets were used to train the model. This approach ensured that every data point contributed to both training and validation, thereby reducing the risk of overfitting or underfitting. The hyperparameter combinations yielding the highest cross-validated R^2 scores and lowest generalization errors on the test set were selected as the optimal configuration. The hyperparameters and their corresponding search spaces used in the GridSearchCV procedure are summarized in Table S2.

Description of ML models and optimization algorithm

Model development

In this study, a supervised ML framework was developed to predict agronomic traits of roselle as influenced by genotype and planting date. Two models, MLP and RF, were implemented within a unified scikit-learn[20] pipeline structure. Each model was trained and evaluated based on multi-output regression performance (Fig. 1).

RF model

For a single decision tree with L terminal leaves, the input feature space is partitioned into R non-overlapping regions R_m such that the prediction function is defined as:

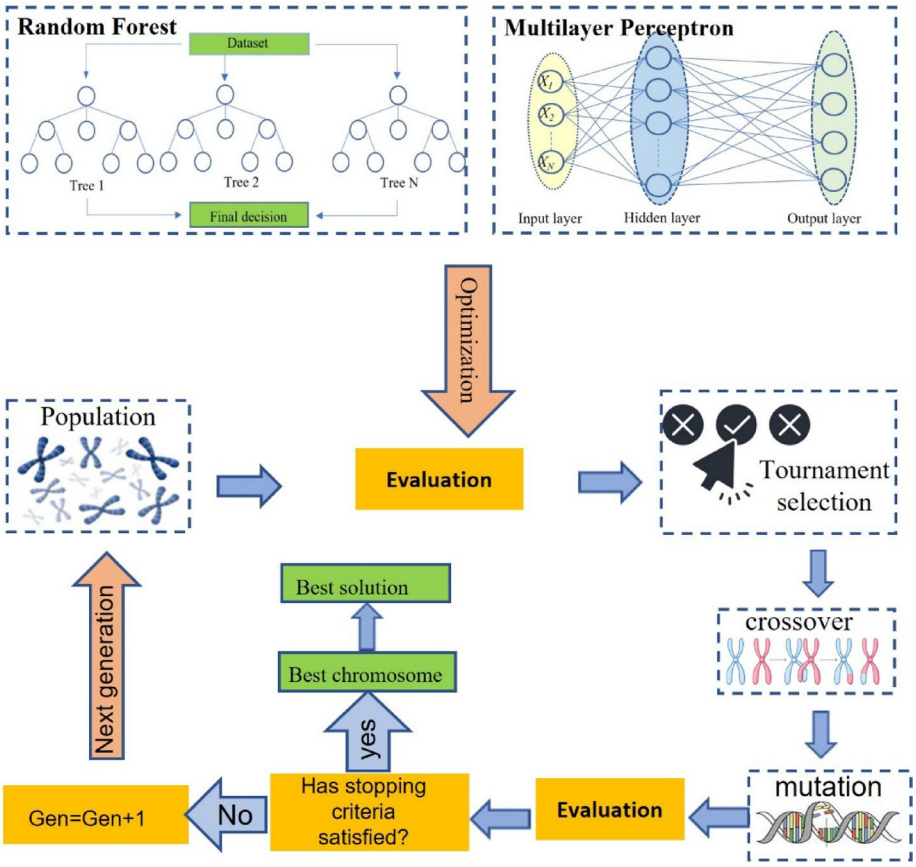

Fig. 1. Schematic diagram of the procedure used in this study, modeling morphological traits based on two input variables, including roselle genotypes and planting dates, using Random Forest (RF) and multilayer perceptron (MLP), and the step-by-step optimization process of morphological traits via non-dominated sorting genetic algorithm-II (NSGA-II).

$$f\,(x) = \sum {}^{R}_{m=1} C_m . \Pi\,(x, R_m\,) \tag{1}$$

where c_m represents the predicted constant value in region R_m, and $\Pi\,(x,R_{m)}$ is an indicator function defined by:

$$\Pi\;\;(x, R_m\,) = \left\{ \begin{array}{c} x\;\in\;\; R_m \\ 0,\; otherwise \end{array} \right. \tag{2}$$

The final RF prediction for an input sample x is obtained by averaging the outputs of all T individual trees. The final prediction $\hat{y}$ for an input x is the average of predictions from all trees:

$$? = \left(\frac{1}{T}\right) \sum {}^{T}_{t=1} f_t\,(x) \tag{3}$$

where $f_t(x)$ is the output of the tth tree. This ensemble averaging mechanism enhances the predictive performance by reducing variance, improving robustness across different training subsets, and minimizing the risk of overfitting. In this study, the RF model was configured with 50 estimators (n_estimators = 50) and a maximum depth of 5 per tree (max_depth = 5). At each node, a random subset of predictors was selected to determine the best split, which encourages diversity among the trees and improves generalization.

MLP model

In this study, MLP model was implemented as a supervised learning algorithm consisting of three layers: an input layer, two hidden layers with five neurons each (5, 5), and a single output layer. The input vector $x = [x_1, x_2, \ldots, x_n]$, representing genotype and planting date encoded via one-hot encoding, was passed through the hidden layers, each employing the hyperbolic tangent activation function (tanh):

$$h_j = tanh(\sum\; w_{ji}\;\cdot\; x_i\;+\; b_j) \tag{4}$$

*i*th input and the *j*th hidden neuron, b_j is the bias for the *j*th neuron, and tanh() is the hyperbolic tangent activation function. The final output $\hat{y}$, representing predicted plant traits (branch number, growth period, boll number, and seed yield), was computed through a linear activation function:

$$\hat{y} = \sum w_j \cdot h_j + b_0 \tag{5}$$

w_j is the weight connecting the *j*th hidden neuron to the output, and b_0 is the output bias.

The model was trained using the Adam optimizer with a maximum of 1000 iterations. To improve generalization and prevent overfitting, L2 regularization (weight decay) was incorporated via the alpha parameter, set to 1. Accordingly, the loss function minimized during training was defined as the regularized mean squared error (MSE):

$$MSE = \left(\frac{1}{K}\right) \sum (y_k - \hat{y}_k)^2 + a \sum w_j^2 \tag{6}$$

where K is the number of training samples, $\hat{y}_k$ is the predicted output for the *k*th sample, and $\alpha = 1$ is the regularization parameter controlling the penalty applied to large weights. This formulation penalizes large weights and enhances the model's robustness against noise and overfitting. The architecture and number of hidden neurons were selected based on a combination of empirical trial-and-error and structured grid search optimization, ensuring the best trade-off between complexity and predictive accuracy.

Model performance evaluation

To evaluate the performance of the developed ML models, several statistical indicators were employed, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute percentage error (MAPE), and mean bias error (MBE). These metrics provide a comprehensive assessment of the model's accuracy, precision, and bias. These quantitative indicators can be found in Table 1. They were computed for both training and testing subsets to ensure the generalization capability of the models across unseen data. The prediction results for each target trait were subsequently compared and visualized through scatter plots of observed vs. predicted values (Fig. 3a-h). The two implemented models were then compared based on these performance metrics in the investigated targets.

Permutation-based feature importance

To assess the relative contribution of each input feature to the model's predictive performance, permutation-based feature importance analysis was performed for the RF and MLP models[21]. For each model, a MultiOutputRegressor was trained and evaluated, and permutation importance was computed separately for each target trait: number of branches, growth period, number of bolls, and seed yield per plant. The permutation importance method involves randomly shuffling the values of each feature and observing the resulting change in prediction error[21]. This process was repeated 50 times to obtain a stable estimate of feature relevance. The procedure was applied to the test set after transforming it using the model's preprocessing pipeline (including standardization and one-hot encoding). The importance scores were averaged across all repetitions, and the results were visualized using bar plots for each target trait and also aggregated to assess global feature influence. The implementation was carried out using the Scikit-learn library (version 1.3.2)[20] via the permutation_importance function. The results were visualized using bar plots to highlight the most influential genotype and planting date combinations (Fig. 4).

Optimization of ML model via NSGA-II

To identify optimal genotype and planting date combinations for maximizing agronomic performance, the best-performing ML model, RF, was employed as a fitness function and integrated with the NSGA-II for multi-objective optimization (Fig. 1). The initial population of candidate solutions was randomly generated, and the Tournament Selection method with Crowding Distance Comparison (selTournamentDCD), implemented in the DEAP library, was used to select elite individuals. In this method, two individuals are randomly selected and compared first based on their Pareto rank, and if they belong to the same non-dominated front, the one with the higher crowding distance is selected[22]. This mechanism promotes both convergence and diversity by preferring individuals located in sparsely populated areas of the objective space. The concept of Pareto dominance states that a solution A dominates solution B if it is no worse than B in all objectives and strictly better in at least one,

Metric	Formula	Description
R^2 (Coefficient of Determination)	$R^2 = 1 - (\Sigma (y_i - \hat{y}_i)^2) / (\Sigma (y_i - \bar{y})^2)$	R^2 indicates the proportion of the variance in the observed values explained by the model. A higher R^2 value (closer to 1) reflects better model performance and fit.
RMSE (Root Mean Squared Error)	$RMSE = \sqrt{[(1/n) \Sigma (y_i - \hat{y}_i)^2]}$	RMSE measures the standard deviation of prediction errors. It quantifies the model's ability to predict observed values close to actual ones. Lower values indicate higher accuracy.
MAPE (Mean Absolute Percentage Error)	$MAPE = (100/n) \Sigma \lvert(y_i - \hat{y}_i) / y_i\rvert$	MAPE expresses the average absolute error as a percentage of actual values. It allows for easy interpretation and comparison across different scales or units.
MBE (Mean Bias Error)	$MBE = (1/n) \Sigma (\hat{y}_i - y_i)$	MBE quantifies the average bias in predictions, indicating whether a model tends to overestimate (positive MBE) or underestimate (negative MBE) values.

Table 1. Description of regression evaluation metrics. Where y_i: observed value, $\hat{y}_i$: predicted value, $\bar{y}$: mean of observed values, n: number of samples.

forming the basis for constructing non-dominated fronts[12,14,23]. Within each front, the crowding distance of a solution is computed as the average normalized distance to its adjacent neighbors across all objectives, serving as an estimate of local solution density. Higher crowding distance values indicate greater isolation, which helps in maintaining diversity and avoiding premature convergence during the evolutionary process[22]. To improve the quality of the optimization, key parameters, including population size, number of generations, crossover rate, and mutation rate, were optimized through trial and error[22,23]. The final configuration included a population size of 100, 200 generations, a crossover probability of 0.9, and a mutation probability of 0.1. The distribution indices for crossover and mutation operators were set to 15 and 20, respectively, to control the extent of variation introduced in offspring.

All analyses and model implementations were conducted in Python (version 3.11.12)[20]. Key libraries utilized include scikit-learn (version 1.3.2)[20] for data preprocessing, hyperparameter tuning, model development, performance evaluation, and feature importance analysis (via the permutation_importance function). Moreover, the multi-objective optimization algorithm NSGA-II was implemented using the deap library (version 1.4.2)[22]. To ensure the reproducibility of the results, a fixed random seed (random_state = 42) was applied consistently across all stages of the analysis, including data splitting, model training, optimization, and hyperparameter tuning.

Results

Correlation coefficient, and statistical analyses

Comprehensive correlation and two-way ANOVA analyses were conducted to evaluate the influence of genotype and planting date on a broad set of morphological and yield traits in Roselle. The correlation results (Fig. 2a) revealed that most traits were more strongly associated with planting date than with genotype, highlighting the dominant role of temporal factors in trait variation. Notably, branch number exhibited the highest correlation with planting date ($r = 0.63$), followed by sepal weight ($r = 0.49$), calyx size ($r = 0.47$), harvest index ($r = 0.34$), and boll number ($r = 0.31$). In contrast, correlations with genotype were consistently weak, with the highest observed for seed per plant ($r = 0.082$) and negligible values for other traits, including a slight negative correlation for plant height ($r = -0.12$). Based on these findings, traits with minimal correlation to either input, such as plant height, calyx, harvest index, and sepal weight, were excluded from further modeling to enhance predictive clarity and reduce noise. Complementing the correlation analysis, two-way ANOVA was performed to assess the statistical significance of genotype, planting date, and their interaction (G×E) on each trait (Fig. 2b). Our findings highlight that planting date consistently emerged as the most influential factor across most traits, underscoring the critical role of environmental conditions and sowing time in shaping Roselle performance. Traits such as branch number, boll number, and seed yield per plant demonstrated highly significant responses ($p < 0.01$) to all three factors, genotype, planting date, and their interaction, suggesting that these traits are highly responsive to both genotype and planting date. In contrast, traits such as harvest index, growth period, sepal weight, and calyx size were mainly influenced by planting date and G×E interaction, but not by genotype ($p > 0.05$), highlighting their strong environmental dependence. Overall, among all the morphological traits measured in this study, branch number, boll number, seed per plant, and growth period were selected as the final targets for predictive modeling and optimization.

Model performance evaluation

In this study, RF and MLP models were used to predict morphological parameters based on genotype and planting date. The performances of both models are presented in Table 2. The accuracy of each model was evaluated by R^2, RMSE, MBE, and MAPE. The results show that two models provided acceptable levels of accuracy. however, among the models tested, the RF algorithm ($R^2 = 0.84$) outperformed the MLP ($R^2 = 0.80$) in most of the target

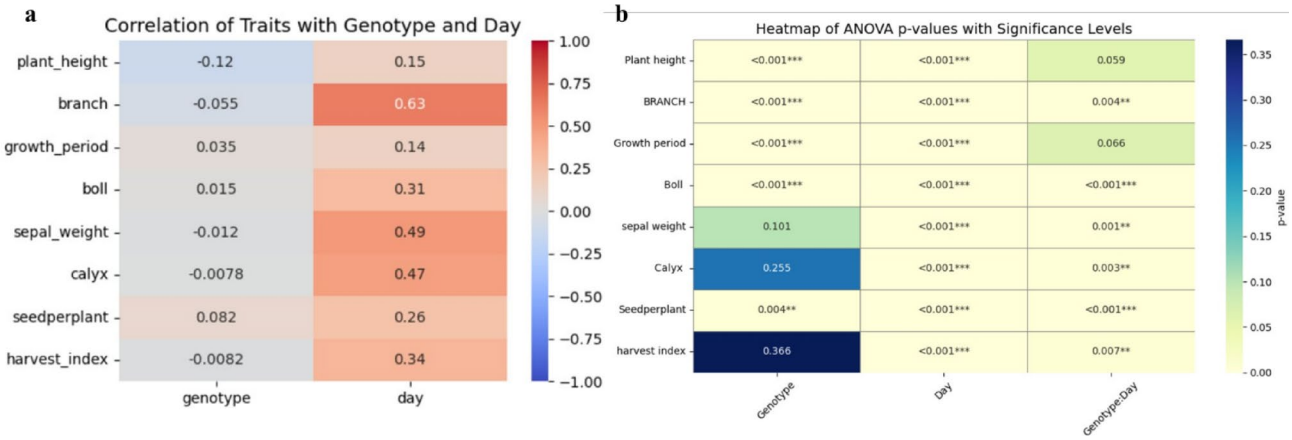

Fig. 2. (**a**) The correlation heatmap between plant traits (targets) and two input variables: genotype and planting date. (**b**) The heatmap of ANOVA p-values showing the significance of genotype, planting date, and their interaction on each trait. The significance level of each factor is represented both numerically and by asterisks: * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

Model	Subset	Criterion	branch	growth_period	boll	seedperplant
RF	Training	R^2	0.857	0.969	0.856	0.702
		RMSE	1.90	5.59	8.33	190.15
		MBE	0.007	0.020	0.141	-3.856
		MAPE	0.097	0.025	0.083	0.181
	Testing	R^2	0.845	0.956	0.861	0.681
		RMSE	1.88	6.72	8.49	199.83
		MBE	-0.21	-0.33	-1.17	17.45
		MAPE	0.093	0.031	0.089	0.198
MLP	Training	R^2	0.871	0.967	0.860	0.755
		RMSE	1.81	5.81	8.23	172.66
		MBE	-0.005	-0.002	0.004	0.758
		MAPE	0.092	0.026	0.077	0.160
	Testing	R^2	0.860	0.953	0.821	0.739
		RMSE	1.99	6.83	8.98	210.12
		MBE	-0.356	-0.259	-1.285	28.62
		MAPE	0.105	0.032	0.096	0.220

Table 2. Comparison statistics of MLP and RF models for various morphological traits of *H. sabdariffa* under training and testing conditions. Traits include branch number (branch), growth period (growth_period), number of bolls per plant (boll), and seeds per plant (seedperplant). R^2, coefficient of determination; RMSE, root mean square error; MBE, mean bias error; MAPE, mean absolute percentage error.

traits. Specifically, the RF model yielded the highest R^2 and lowest error values in the majority of traits, indicating robust generalization ability and a well-fitted predictive structure. The regression lines demonstrated a good fit correlation between the observed and predicted data for all growth parameters during the training and testing processes of the RF model (Fig. 3a-h).

Feature importance evaluation

The permutation importance analysis revealed clear and consistent patterns regarding the relative influence of each feature, including planting date and genotype, on the prediction of Roselle morphological traits (Fig. 4a-d). Among the evaluated features, including five planting dates and ten genotypes, planting date emerged as the dominant factor across most traits. According to the results, planting dates (particularly May, April, and March) showed the highest importance scores, indicating their dominant role in predicting Roselle morphological traits (Fig. S3). Traits such as branch number, growth period, and boll number demonstrated high sensitivity to planting date. In terms of genotypic influence, although generally less impactful than planting date, the genotypes *Iranshahr*, *HA*, and *Qaleganj* showed relatively higher importance in specific traits (Fig. S3). For branch number and boll count, '*Iranshahr*' and '*Qaleganj*' emerged as the most influential genotypes, contributing noticeably to model predictions. In the case of the growth period, '*HA*' was the most prominent genotype, though its influence remained modest compared to planting date. For seed per plant, both planting date and genotype exhibited very low importance, indicating limited predictability and a likely dependence on unmeasured physiological or environmental factors. Overall, despite their lower importance in morphological trait prediction in Roselle, genotypes may still hold distinct value for targeted breeding programs, particularly when aligned with optimized planting schedules.

Model optimization using NSGA-II

The NSGA-II algorithm was integrated with the RF model, as the most accurate predictive algorithm, in this study. The combined RF-NSGA-II model effectively determined the optimal values of the input variables (genotype and planting date) to simultaneously maximize four agronomic traits: branch number, growth period, boll number, and seeds per plant. The results of the multi-objective optimization process using the NSGA-II algorithm are summarized in Table 3. The theoretically optimal performance can be achieved with the genotype *Qaleganj* and the planting date of May, yielding the predicted trait values of branch number, growth duration, boll number, and seed production per plant, 26.009, 175.872, 116.078, and 1517.165, respectively.

Discussion

Yield prediction has an important role in crop farming aimed at efficient and sustainable production. Accurate and timely predictions are important for farmers' decision-making regarding genotype selection, planting date, irrigation, fertilization, harvesting, and trading[24]. Yield prediction in crop science is inherently challenging because of the multifaceted interactions among genetic factors (G), environmental conditions (E), and management practices[25]. Traditional linear models, including multiple linear regression and correlation-based approaches, often fall short in capturing these nonlinear, dynamic relationships, particularly under variable environmental contexts[10]. In contrast, ML models such as RF, support vector regression (SVR), ANNs, and deep

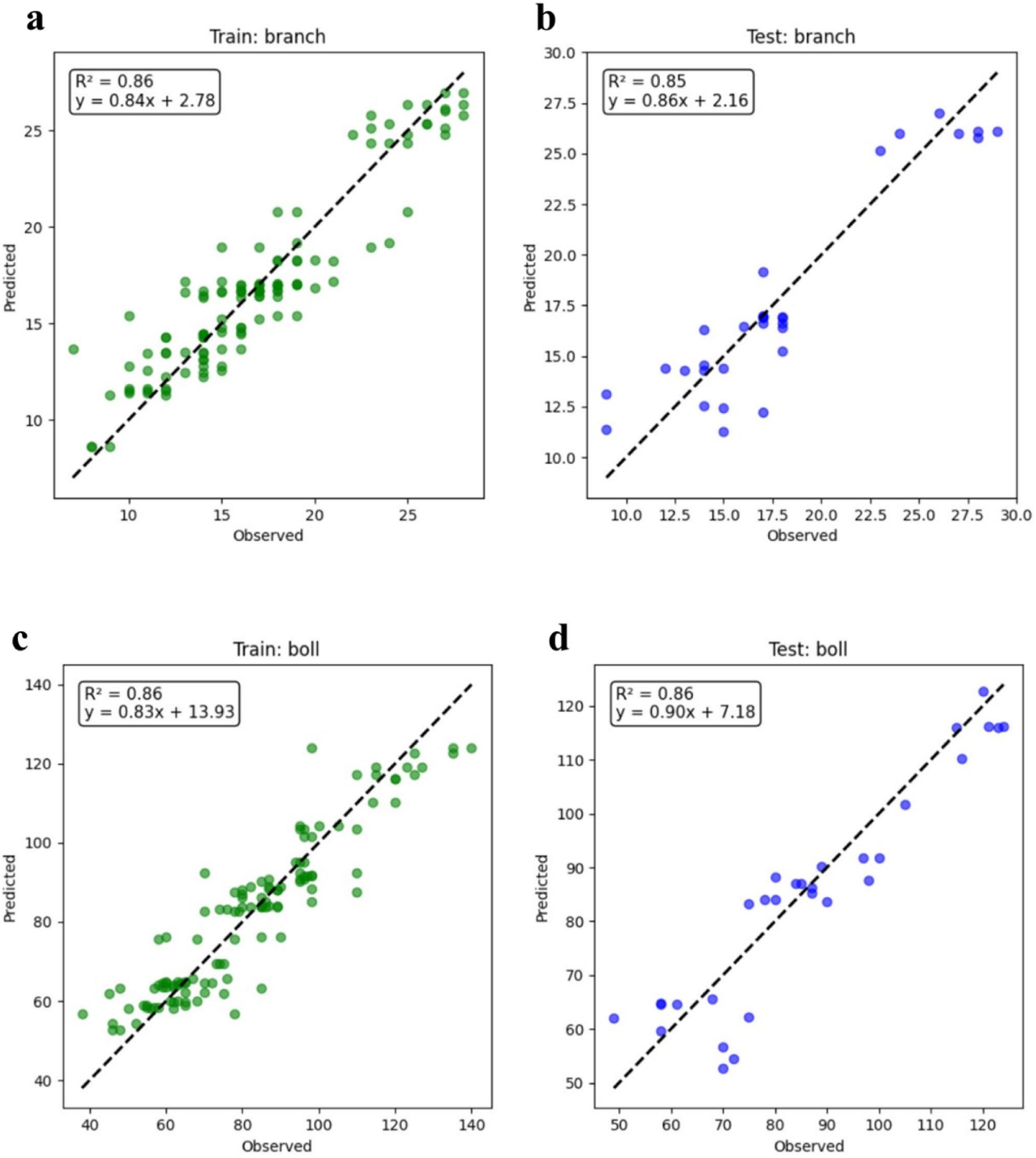

Fig. 3. The scatter plot of observed values vs. predicted values of (**a**, **b**) branch number (Training) and (Testing), (**c**, **d**) boll number (Training) and (Testing), (**e**, **f**) seed per plant (Training) and (Testing), (**g**, **h**) growth period (Training) and (Testing) obtained by the Random Forest (RF) model. The dotted line represents the fitted simple linear regression line across the scatter points, highlighting the relationship between observed and predicted values for each trait and reflecting the model's predictive trend.

learning (DL) architectures have emerged as robust alternatives, offering improved accuracy and adaptability in diverse agricultural scenarios[25].

Beyond predictive accuracy, model complexity and computational efficiency are important considerations in agricultural ML applications. In this study, both the RF and MLP models were evaluated not only in terms of prediction accuracy but also in terms of training time and resource demands. Although the difference in predictive performance between RF ($R^2 \approx 0.84$) and MLP ($R^2 \approx 0.80$) was relatively small, RF consistently outperformed

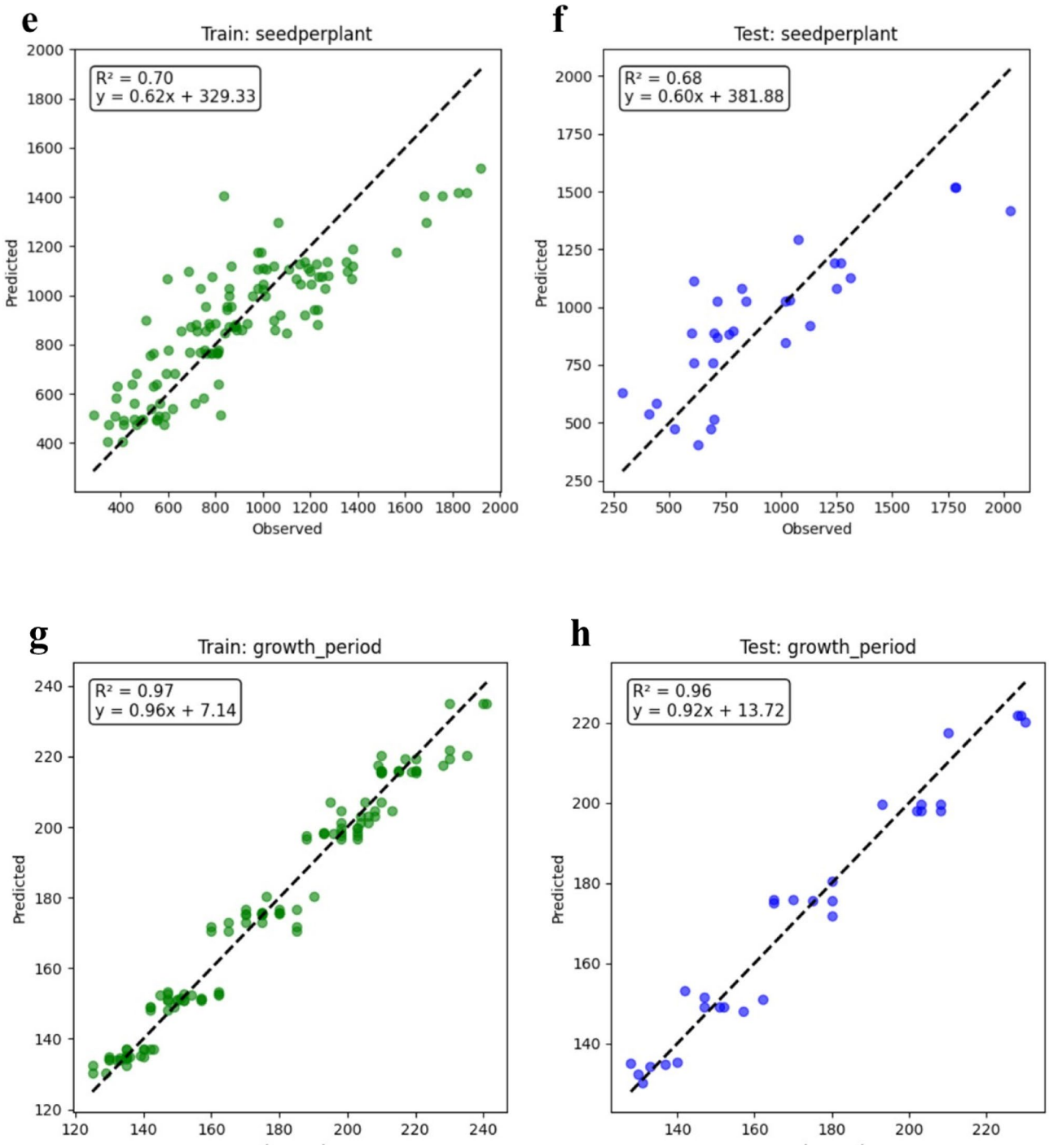

Fig. 3. (continued)

MLP across all yield-related traits. The ensemble structure of RF, which aggregates predictions from multiple decorrelated decision trees, helps mitigate overfitting and improves generalizability, especially when dealing with noisy or limited agricultural datasets[26,27]. In contrast, MLP models require extensive hyperparameter tuning (e.g., hidden layer architecture, learning rates, activation functions) and iterative training using gradient-based optimization. This complexity not only increases the risk of overfitting but also significantly extends training and tuning times. From a computational standpoint, RF was faster to train and tune using GridSearchCV with 10-fold cross-validation, while MLP incurred notably higher computational costs due to its sensitivity to network configuration and the need for convergence over many iterations. This makes RF more suitable for rapid deployment and real-time use in field settings or resource-constrained environments[26,27]. Therefore, while MLP may offer additional flexibility in modeling complex nonlinear interactions, RF was selected as a

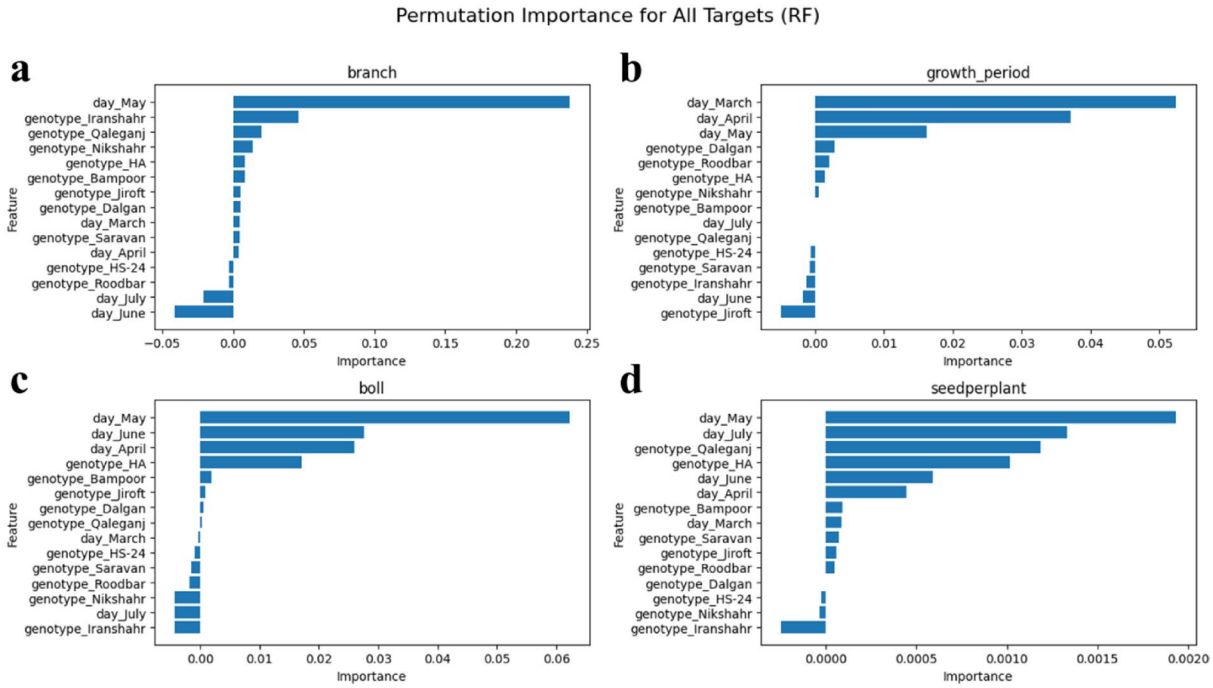

Fig. 4. Bar plots of feature importance in the RF model showing the influence of planting dates and different genotypes in predicting Roselle yield traits: (**a**) branch number; (**b**) growth period; (**c**) boll number and (**d**) seed per plant. To improve clarity and avoid overloading the plots, only the top-ranked input features were visualized in the feature importance graphs, although all encoded genotype × planting date combinations were included in the analysis.

Input Items		Output Items			
Genotype	Planting date	Predicted Branch number	Predicted growth period	Predicted boll number	Predicted seed per plant
Qaleganj	May	26.009	175.872	116.078	1517.165

Table 3. Optimization of genotype and planting date of roselle according to the RF-NSGA-II algorithm to obtain the best morphological traits, including number of branches, growth period, number of bolls, and number of seeds per plant.

more interpretable, computationally efficient, and robust model in the context of Roselle yield prediction under diverse genotype × environment interactions.

Permutation importance analysis identified planting date as the most influential factor across most traits, overshadowing the contributions of genotype. This aligns with findings of previous studies, which have consistently highlighted early planting as a critical determinant of biomass accumulation and overall productivity in Roselle[8,28]. Early planting dates (March to May) likely provide more favorable thermal and photoperiodic conditions, enhancing growth rates and extending the vegetative phase, and eventually improving the yield[7].

Although genotype had a comparatively lesser effect, its role was not negligible. Specific genotypes such as *Iranshahr* and *Qaleganj* demonstrated higher predictive importance for branch number and boll count, while the *HA* genotype showed prominence in extending the growth period. These findings corroborate earlier observations, which identified substantial genetic variability in Roselle for key yield-contributing traits, including plant height, branching, calyx yield, fiber production, and nutritional composition[29,30]. Collectively, these studies highlight the considerable potential for selective breeding and genetic improvement in Roselle, emphasizing the importance of exploiting this variability to develop high-yielding and resilient cultivars. High heritability observed in traits such as plant height and number of branches suggests that these are primarily governed by genetic factors and can be effectively improved through selection. The interactions between genotype and planting date observed in this study suggest that while environmental timing is the primary driver of trait variability, genotype selection can further refine yield performance when optimized within suitable environmental windows. Moreover, our study revealed that seed yield per plant consistently exhibited low predictive importance for both planting date and genotype, suggesting a reliance on factors not directly captured in this dataset. This finding underscores the complexity and multifaceted nature of seed production in Roselle. Specifically, it highlights that while other morphological traits (such as branch number and boll count) are primarily shaped by genotype and planting date, seed yield is more strongly influenced by unmeasured physiological and environmental interactions. Our

findings are consistent with earlier studies that have underscored the pivotal role of micro-environmental and physiological factors in determining seed production in Roselle[29,31]. These factors include variations in plant morphology, soil moisture levels, and physiological responses to environmental stressors, all of which contribute significantly to yield differences across genotypes[30]. In the permutation feature importance plots (Fig. 4), it is observed that some features have negative values. In permutation importance analysis, the value of a feature is randomly permuted in the dataset to assess its impact on model performance. Typically, a decrease in model performance after permuting a feature indicates a high importance of that feature in the prediction process. However, in some cases, permuting a feature leads to a slight improvement in model performance, which appears as a negative importance value. This phenomenon often occurs when the feature in question contains noise, redundant information, or weak content, and has harmed the model's predictions[21]. Such features may cause a decrease in accuracy by perturbing the patterns learned by the model, and therefore, removing or ignoring them can slightly improve model performance. Additionally, negative significance values can be an indication of multicollinearity, a condition in which multiple features are highly correlated and the model is overfitted. In such cases, removing one of the correlated features can reduce the model's complexity and increase its generalizability. Also, statistical fluctuations can cause false negatives in significance analysis when the test set is small or variable[32]. In general, the presence of negative values in the permutation test is not necessarily a cause for concern, but can provide useful information about features that not only do not play a useful role in prediction, but may also weaken the generalizability of the model. These findings suggest opportunities for dimensionality reduction and improving generalizability without sacrificing model accuracy.

After diagnosing the RF model as the best model based on the highest accuracy, the NSGA-II algorithm was linked to it. The results of the RF-NSGA-II algorithm highlighted the combination of the *Qaleganj* genotype and a planting date of May 5 as the best compromise or optimal point for maximizing morphological traits, among the studied scenarios. This was not the result of a single-objective maximization but rather the consequence of a multi-objective search that balanced conflicting agronomic goals, maximizing branch number, boll number, and seed yield per plant, while simultaneously minimizing growth duration. The NSGA-II framework enabled the identification of a diverse set of trade-off solutions using the principles of Pareto dominance and crowding distance[15]. In each generation, individuals (i.e., candidate genotype × planting date combinations) were ranked based on non-dominated sorting, where a solution is considered superior if it is not outperformed across all objectives by any other solution[15,17]. Among solutions of the same Pareto rank, selection was guided by crowding distance, which estimates the density of neighboring solutions in objective space. Individuals with higher crowding distances were favored to maintain population diversity and avoid premature convergence to suboptimal regions[13,23]. The final solution displayed in Table 3 was selected from this Pareto-optimal set as it offered one of the most balanced compromises across all four traits. Its emergence as an elite solution can be attributed to its ability to remain non-dominated throughout the evolutionary process while maintaining a high crowding distance, which ensured its persistence through TournamentDCD-based selection. This dual mechanism, favoring solutions that are both non-dominated and well-distributed, allowed the optimization process to explore a wide solution space and converge toward an efficient frontier of agronomic performance. Therefore, the application of NSGA-II in this study was not only instrumental in identifying optimal input combinations but also provided biological interpretability by illustrating the inherent trade-offs between vegetative and reproductive growth parameters under varying temporal and genetic conditions[17]. While Table 3 highlights the *Qaleganj* genotype as having optimal predicted trait values under the May planting scenario, it is important to note that the *HA* genotype also consistently exhibits strong performance across all key traits. Specifically, permutation importance analysis revealed that the *Iranshahr* and *HA* genotypes had higher relevance scores for growth period, boll number, and branch number, and its seed per plant value was comparable to *Qaleganj*. Moreover, *Iranshahr* showed the highest importance for the branch number, indicating its potential value in enhancing vegetative growth. However, the selection of *Qaleganj* was not based solely on individual trait performance or feature importance. Rather, it was the result of a multi-objective optimization process using the NSGA-II algorithm, which identifies genotypes that provide the best trade-offs among multiple traits. In this context, *Qaleganj* emerged as a Pareto-optimal solution offering a balanced combination of vegetative growth and reproductive potential when planted in May. This does not diminish the potential of genotype *HA*; in fact, our findings suggest that *HA* is another promising candidate with high trait performance. Future field validation may further confirm whether *HA* can match or outperform *Qaleganj* under varying environmental and planting conditions.

To the best of our knowledge, this is the first study to apply ML-NSGA-II for modeling and optimizing morphological traits of Roselle in response to various genotypes and planting dates. In conclusion, our study highlights the applicability and reliability of ML models, particularly RF, for analyzing complex genotype-environment interactions in Roselle cultivation.

While this study provides valuable insights into the prediction and optimization of agronomic traits using ML techniques, several limitations should be acknowledged. First, the experimental dataset was collected from a single geographical location. As a result, the models developed and optimized in this study may not generalize well to other environmental conditions or geographic regions with differing climatic and soil characteristics. Future studies should consider multi-location trials to capture broader environmental variability. Second, although the dataset included ten genotypes and five planting dates, the scope of genetic and temporal diversity remains limited. Consequently, extrapolating the results to other cultivars or planting schedules should be approached with caution. Validation with independent datasets from other seasons or genetic backgrounds is recommended to confirm the model's robustness. Third, the study relied on limited categorical traits as input variables without incorporating additional physiological or environmental covariates that may influence yield-related responses. Factors such as soil moisture content, light intensity, nutrient availability, or hormonal levels were not measured and may have contributed to unexplained variability in plant performance. Including such variables in future

models could enhance prediction accuracy and biological interpretability. Addressing these limitations in future research will support the development of more robust and generalizable predictive frameworks for plant trait modeling and agricultural decision-making.

Conclusion

This study aimed to predict and understand how genotype and planting date affect the morphological traits of Roselle by leveraging advanced ML models, specifically RF and MLP, for the first time. The comparative analysis demonstrated that despite the relatively small difference in predictive performance, the RF model consistently outperformed MLP, confirming its robustness in capturing the complex genotype-by-environment interactions typical in agricultural systems. By integrating RF with the NSGA-II algorithm, we successfully identified the optimal combination of the *Qaleganj* genotype and the planting date of May 5 to maximize morphological traits. The RF-NSGA-II hybrid algorithm proved highly effective for multi-objective optimization, offering a powerful tool to identify ideal input combinations under varying conditions. These findings emphasize the potential of advanced ML-based approaches as robust alternatives to traditional statistical methods, offering new avenues for optimizing and improving morphological traits in Roselle and other crops in future studies. Future research can expand on these findings by incorporating additional environmental variables, exploring genomic selection approaches, and validating the models across multiple growing seasons and agroecological zones.

Data availability

The authors confirm that the datasets analyzed during the current study are available from the corresponding author on request.

Received: 17 June 2025; Accepted: 7 August 2025
Published online: 09 August 2025

References

1. Ali, B. H., Wabel, N. A. & Blunden, G. Phytochemical, Pharmacological and toxicological aspects of *Hibiscus Sabdariffa* L.: a review. Phytotherapy research. *Int. J. Devoted Pharmacol. Toxicol. Evaluation Nat. Prod. Derivatives*. **19**, 369–375. https://doi.org/10.1002/ptr.1628 (2005).
2. Zand-Silakhoor, A., Madani, H., Sharifabad, H. H., Mahmoudi, M. & Nourmohammadi, G. Influence of different irrigation regimes and planting times on the quality and quantity of calyx, seed oil content and water use efficiency of roselle (*Hibiscus Sabdariffa* L). *Grasas Y Aceites*. **73**, e472–e472. https://doi.org/10.3989/gya.0564211 (2022).
3. Mahunu, G. K. Roselle (*Hibiscus sabdariffa*): Botany, Production, and Uses. In Roselle (*Hibiscus sabdariffa*) 1–14 (Elsevier, 2021). https://doi.org/10.1016/B978-0-12-822100-6.00005-7
4. Mulyaningsih, E. S. et al. Morpho genetic variability and anthocyanine (Cyanidin-3-OGlucoside) Concent of Indonesia roselle (*Hibiscus Sabdariffa* L). *Int. J. Adv. Sci. Eng. Inform. Technol.* **15** (1), 240. https://doi.org/10.18517/ijaseit.15.1.19871 (2025).
5. Ibrahim, M. & Hussein, R. Variability, heritability and genetic advance in some genotypes of roselle (*Hibiscus Sabdariffa* L). *World J. Agricultural Sci.* **2**, 340–345 (2006). http://www.idosi.org/wjas/wjas2(3).htm
6. Richardson, M. L. & Arlotta, C. G. Differential yield and nutrients of *Hibiscus Sabdariffa* L. genotypes when grown in urban production systems. *Sci. Hort.* **288**, 110349. https://doi.org/10.1016/j.scienta.2021.110349 (2021).
7. El-Sagher, M., Mostafa, G. G., El-Ghadban, E. M. A., Soliman, W. S. & Gahory, A. A. Sowing date as a determining factor for roselle, *Hibiscus sabdariffa*, production: I. Effect on vegetative and yield components. *Aswan Univ. J. Sci. Technol.* **4**, 28–37. https://doi.org/10.21608/aujst.2024.337887 (2024).
8. Aung, C. & Uape, M. The effect of different planting dates on the growth and yield of roselle (*Hibiscus Sabdariffa* L.) during the rainy season. *Univ. Yangon Res. J.* **11**, 59–68 (2022). https://meral.edu.mm/records/9019
9. Fakhrzad, F., Jowkar, A. & Hosseinzadeh, J. Mathematical modeling and optimizing the in vitro shoot proliferation of wallflower using multilayer perceptron non-dominated sorting genetic algorithm-II (MLP-NSGAII). *PLoS One*. **17**, e0273009. https://doi.org/10.1371/journal.pone.0273009 (2022).
10. Yoosefzadeh-Najafabadi, M., Tulpan, D. & Eskandari, M. Application of machine learning and genetic optimization algorithms for modeling and optimizing soybean yield using its component traits. *Plos One*. **16**, e0250665. https://doi.org/10.1371/journal.pone.0250665 (2021).
11. Zarbakhsh, S., Shahsavar, A. R. & Soltani, M. Optimizing PGRs for in vitro shoot proliferation of pomegranate with bayesian-tuned ensemble stacking regression and NSGA-II: a comparative evaluation of machine learning models. *Plant. Methods*. **20**, 82. https://doi.org/10.1186/s13007-024-01211-5 (2024).
12. Verma, S., Pant, M. & Snasel, V. A comprehensive review on NSGA-II for multi-objective combinatorial optimization problems. *IEEE Access*. **9**, 57757–57791. https://doi.org/10.1109/ACCESS.2021.3070634 (2021).
13. Kukkonen, S. & Deb, K. A fast and effective method for pruning of non-dominated solutions in many-objective problems. *Int. Conf. Parallel Problem Solving Nat. (Springer)*. **4193**, 553–562. https://doi.org/10.1007/11844297_56 (2006).
14. Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**, 182–197. https://doi.org/10.1109/4235.996017 (2002).
15. Zarbakhsh, S. & Shahsavar, A. R. Artificial neural network-based model to predict the effect of γ-aminobutyric acid on salinity and drought responsive morphological traits in pomegranate. *Sci. Rep.* **12**, 16662. https://doi.org/10.1038/s41598-022-21129-z (2022).
16. Eftekhari, M., Yadollahi, A., Ahmadi, H., Shojaeiyan, A. & Ayyari, M. Development of an artificial neural network as a tool for predicting the targeted phenolic profile of grapevine (*Vitis vinifera*) foliar wastes. *Front. Plant Sci.* **9**, 837. https://doi.org/10.3389/fpls.2018.00837 (2018).
17. 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*. **15**, e0240427. https://doi.org/10.1371/journal.pone.0240427 (2020).
18. Azrai, M. et al. Optimizing ensembles machine learning, genetic algorithms, and multivariate modeling for enhanced prediction of maize yield and stress tolerance index. *Front. Sustainable Food Syst.* **8**, 1334421. https://doi.org/10.3389/fsufs.2024.1334421 (2024).
19. Yoosefzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J. & Eskandari, M. Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. *Front. Plant Sci.* **11**, 624273. https://doi.org/10.3389/fpls.2020.624273 (2021).
20. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
21. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340–1347. https://doi.org/10.1093/bioinformatics/btq134 (2010).

22. Fortin, F. A., De Rainville, F. M., Gardner, M. A. G., Parizeau, M. & Gagné, C. Evolutionary algorithms made easy. *J. Mach. Learn. Res. 13*. **DEAP**, 2171–2175 (2012).
23. Hesami, M. & Jones, A. M. P. Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture. *Appl. Microbiol. Biotechnol.* **104**, 9449–9485. https://doi.org/10.1007/s00253-020-10888-2 (2020).
24. Khaki, S. & Wang, L. Crop yield prediction using deep neural networks. *Front. Plant Sci.* **10**, 621. https://doi.org/10.3389/fpls.2019.00621 (2019).
25. Leukel, J., Zimpel, T. & Stumpe, C. Machine learning technology for early prediction of grain yield at the field scale: A systematic review. *Comput. Electron. Agric.* **207**, 107721. https://doi.org/10.1016/j.compag.2023.107721 (2023).
26. Asamoah, E., Heuvelink, G. B., Chairi, I., Bindraban, P. S. & Logah, V. Random forest machine learning for maize yield and agronomic efficiency prediction in Ghana. *Heliyon* **10**, e37065 (2024).
27. Gupta, I. et al. Innovations in agricultural forecasting: A multivariate regression study on global crop yield prediction. *ArXiv Preprint arXiv*. 231202254. https://doi.org/10.48550/arXiv.2312.02254 (2023).
28. Parsa Motlagh, B., Rezvani Moghaddam, P. & Azami Sardooei, Z. Responses of calyx phytochemical characteristic, yield and yield components of roselle (*Hibiscus Sabdariffa* L.) to different sowing dates and densities. *Int. J. Hortic. Sci. Technol.* **5**, 241–251. https://doi.org/10.22059/ijhst.2018.258629.243 (2018).
29. Ibrahim, E. B., Abdalla, A. W. H., Ibrahim, E. A. & El Naim, A. M. Variability in some roselle (*Hibiscus Sabdariffa* L.) genotypes for yield and its attributes. *Int. J. Agric. Forestry*. **3**, 261–266. https://doi.org/10.5923/j.ijaf.20130307.02 (2013).
30. Tetteh, A. Y., Ankrah, N. A., Coffie, N. & Niagiah, A. Genetic diversity, variability and characterization of the agro-morphological traits of Northern Ghana roselle (*Hibiscus Sabdariffa* var. altissima) accessions. *Afr. J. Plant Sci.* **13**, 168–184. https://doi.org/10.5897/AJPS2019.1783 (2019).
31. Atta, S. et al. Yield character variability in roselle (*Hibiscus Sabdariffa* L). *Afr. J. Agric. Res.* **6**, 1371–1377. https://doi.org/10.5897/AJAR10.334 (2011).
32. Khan, A., Ali, A., Khan, J., Ullah, F. & Faheem, M. Using Permutation-Based feature importance for improved machine learning model performance at reduced costs. *IEEE Access.* **13** https://doi.org/10.1109/ACCESS.2025.3544625 (2025).

Author contributions

Fazilat Fakhrzad: Conceptualization, Writing – original draft, Formal analysis, software, Review & editing, Visualization. Warqaa Muhammed ShariffAl-Sheikh: Investigation. Mohammed M. Mohammed: Investigation. Heidar Meftahizade: Conceptualization, Project administration, Review & editing.

Funding

No specific financial credit was used in this experiment.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-15373-2.

Correspondence and requests for materials should be addressed to F.F. or H.M.