 doi:10.20944/preprints202208.0522.v1

Towards modeling and predicting the yield of oilseed crops: Multi-machine learning techniques approach

Mahdieh Parsaeian [1], Mohammad Rahimi [2], Abbas Rohani [2], Shaneka S. Lawson[3]*

1. Department of Agronomy and Plant Breeding, Shahrood University of Technology, Shahrood, Iran.

2. Department of Biosystems Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

3. USDA Forest Service, Northern Research Station, Hardwood Tree Improvement and Regeneration Center (HTIRC), Purdue University, Department of Forestry and Natural Resources, 715 West State Street, West Lafayette, USA

*Corresponding author, Email Address: arohani@um.ac.ir

Abstract

The modeling and prediction of crop seed yield can be a vital improvement in the precision agriculture industry as it provides reliable assessments of the effectiveness of agro-traits. Here, multiple machine learning (ML) techniques are established for predicting sesame (*Sesamum indicum* L.) seed yield (SSY) and incorporating agro-morphological features. Models utilized for coupled PCA-ML (Principal component analysis-Machine Learning) methods were compared with original ML models to evaluate predicted efficiency. The Gaussian process regression (GPR) and Radial basis function neural network (RBF-NN) models exhibited the most accurate SSY predictions with determination coefficients or R^2 values of 0.99 and 0.91, respectfully. The root-mean-square error (RMSE) for the ML models fluctuated between 0 to 0.30 t/ha (metric tons/hectare) for the varied modeling process phases. Estimation of sesame seed yield with coupled PCA-ML models improved performance accuracy. The K-fold process suggested the utilization of datasets with the lowest error rates to ensure the continued accuracy of GPR and RBF models. Sensitivity analysis revealed the capsule number per plant (CPP), seed number per capsule (SPC), and 1000-seed weight (TSW) were the most significant seed yield determinants.

Keywords: Agro-morphological; Data-driven; Machine Learning; Seed yield; Sensitivity Analysis

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

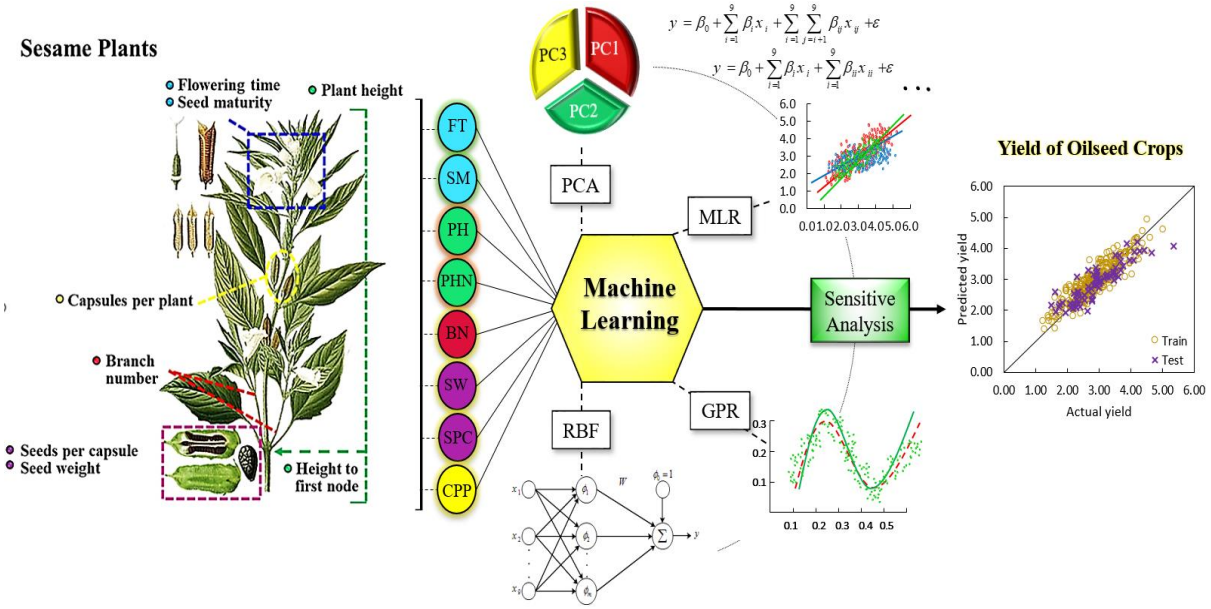
Nomenclature			
ANN	Artificial neural network	RBF	Radial Basis Function
GPR	Gaussian Process Regression	RMSE	Root-Mean-Square Error
MLP	Multilayer Perceptron Neural Network	EF	Efficiency Factor
MLR	Multiple Linear Regression	TSSE	Total Sum Squared Error
PCA	Principal Component Analysis	MAPE	Mean Absolute Percentage Error
SM	Seed maturity	PH	Plant height
CPP	Capsule number per plant	SSY	Seed yield of sesame
BN	Branch number	SPC	Seed number per capsule
TSW	1000-seed weight	FT	Flowering time

Highlights

- Four ML models were employed to aid in predicting sesame seed yield (SSY).
- First use of coupled PCA-ML models for in-depth predictions of SSY.
- Use of the GPR, RBF, and MLR models led to greater accuracy of SSY predictions.
- The primary agro-morphological features for predicting SSY were revealed with sensitivity analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

Graphical Abstract



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

1. Introduction

The continued expansion of industrialization and quickening depletion of resources has led to increased demands for green energy sources. Fossil fuels (e.g., oil, coal, natural gas) are non-renewable resources and, despite being the leading contributors to rising CO_2 emission levels, have long been utilized by industries as fuel [1–3]. Oilseeds are regarded as one of the most important energy sources with diverse industrial and medicinal applications. Therefore, the precise prediction of crop yield is a principal objective for agricultural and industries applications [4,5]. Forecasting crop yields prior to harvest can help identify optimal reaping times and ameliorate concerns by farmers regarding field conditions and management [6,7]. Thus, it is vital to improve planting methodologies for oilseed species and to produce new cultivars with greater potential yields. Sesame (*Sesamum indicum* L.) is one of the oldest oilseeds with nutritious seeds containing oil (34.4-63.2%), proteins (17-32%), minerals, and fat-soluble vitamins [8,9]. Sesame oil is the most stable and high-quality edible oil due to a unique combination of fatty acids and natural antioxidants. However, little research exists regarding the planting and development of adaptable, high-yield cultivars [10–12]. A crucial breeding objective, yield is a complex, quantitative, polygenic trait primarily influenced by several factors underpinning production. The phenotypic representation of this trait is typically impacted by environment and environment × genotype interaction. Hence, it is seldomly heritable and the effectiveness and efficiency of long-term direct selection for this trait is limited [13]. In contrast, selecting seed yield-related traits that are heritable is a promising route to improve seed crop yield. These traits are relatively insensitive to the environment and often highly heritable [14–16]. Yield components can indirectly affect seed yield through their positive or negative interactions. Thus, figuring out the relationships between seed yield and agro-traits is regarded as an impressive approach to trait enhancement.

Statistical modeling applications have been widely used to explain the relationship between morphological and agronomic traits for sesame seed yield. Other yield prediction methods such as quadratic, pure-quadratic, interactions (2FI), and polynomial have previously been used for cotton, maize, and wheat crops. The optimal regression model for this study was selected based on assessment criteria values [17]. In another study, regression analyses were adapted for survey of major environmental factors and their impact on crop yield. Yield predictions were thought to provide substantial benefits to farmers while reducing crop loss and increasing earnings [18]. Alternatively, the Multiple Linear Regression (MLR) technique was employed in the East Godavari district of Andhra Pradesh in India to predict crop yields. Those findings, in comparison with the dataset currently available, will aid efforts to evaluate the efficacy of the proposed technique [19]. A regression model was also manipulated to reveal the relative importance of agronomic traits and genetic correlations with sesame seed yield. The model conveyed data to support CPP having stronger

direct and positive effects on seed yield than most other traits [20]. Another study heralded the significance of CPP and deemed it important for yield selections [21], while other efforts reported greater plant heights combined with higher CPP could increase overall sesame seed yield [22,23]. Similarly, seeds per capsule (SPC), thousand-seed weight (TSW), and the number of capsules per branch (CPB) positively and significantly correlated with seed yield [22,24]. Additional studies in this area uncovered multiple linear regression (MLR) models developed to assess crop yield traits. Independent variables (inputs) affecting seed yield were identified and considered however, CPP was the first variable required for the best results [25]. Similar results were achieved when fitting a predictor equation for seed yield [26,27]. Although traditional statistical methods (i.e. regression analyses) are widely used to derive plant seed yield prediction equations, assumptions such as dependent variables normality, homogeneity of error variance, and inefficient representation of the nature of complex and nonlinear relations in empirical phenomenon represent substantial drawbacks [28].

Machine learning (ML) techniques have attracted extensive attention because they can be easily used in fields such as agriculture, chemical, and energy for a variety of applications [29–35] Consequently, agronomists have shifted to machine learning methods like Artificial Neural Networks (ANNs) and Gaussian Process Regression (GPR) models in recent years [36–40]. ML models are especially effective in agricultural fields and have been used for product image processing [41], separation of weeds and vegetative cover in remote sensing [42], prediction of solar radiation [43], flood forecasting [44], hydrogen storage on bio-carbon [45], biomass estimation [46], and estimation of soil erosion rate [47]. As shown in **Table 1**, numerous studies have expounded upon the usefulness of ML in investigations of seed and crop yields. Moreover, predictions of agro-product constituents such as oil or nitrogen contents or disease diagnosis and plant classifications are most often accomplished with ML models. These intelligent models use numerous interconnected processing elements to solve problems and can be modified to perform specific functions including pattern identification, data classification, and prediction and modeling of processes through a reliable learning process [48]. ANNs are characterized by suitable error tolerance, direct learning from data, and a lack of a need for statistical quantity estimation [49,50]. Predictions of output correspond to a set of inputs where parameter relationships serve different functions based on study goals [16]. In the agriculture field, ML is most often tasked with investigating multi-objective concerns such as crop yield estimation and quality control. A selection of agricultural research studies devoted to crop yield, plant classification, seed assessments, and crop quality control where ML was incorporated is illustrated in **Fig. 1**. Radial basis function neural networks (RBF-NN) and regression models have been adapted for the prediction of tree trunk volume. The RBF neural network has been operationally more reliable than regression models in completing this task [51].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

Table 1. A selection of previous studies using ML to advance agricultural crop research.

Application	Performance prediction	Model	RMSE (R^2)*	*Ref.*
Seed and Crop Yield	Prediction of oilseed rape yield with alternative planting styles and varied nitrogen fertilizer application.	SVR ANN PLSR	-	[52]
	Estimation of soybean seed yield using collected multispectral images for prediction.	MLP	-	[53]
	Incorporation of multi-qualitative and quantitative features for estimation of wheat yield.	ANN	-	[54]
	ML model comprised of high dimensional phenotypic trait data to carry out in-season seed yield predictions.	RF	- (0.83)	[55]
	Ten agro-morphological and phenological traits (plant height, number of branches per plant, number of capsules per plant number of days to flowering, number of days to maturity, thousand seed weight, etc.) were used as the basis for a predictive seed yield model.	ANN MLP RBF PCA	0.87 (0.92)	[16,56]
	Prediction of crop yields in mustard and potato with models using soil elemental properties, physicochemical features, pH, electrical conductivity, organic carbon, and others for training and test datasets.	ANN SVR KNN	- 4.62 (0.72)	[57,58]
	Prediction of corn crop yield by careful climate change factor (temperature and moisture) evaluation to compile an impact assessment for corn fields.	ANN	1.5	[59]
	ML to aid predictive estimates of maize crop yield using topography, land use, soil data, and multiple other parameters.	-	(0.96)	[60]
	Yield predictions in rice paddies using climate-based factors (rainfall, morning and evening relative humidity, minimum and maximum temperature).	ANN	31	[61]
	Utilization of fertilizer volume in tandem with general atmospheric conditions to predict maize yield.	ANN MR	30	[62]
	Predictions of rice paddy yields based on environmental features (area, number of open wells, tanks, maximum temperature, etc.) as independent variables.	ANN MLR SVR RF	0.05-0.1 (0.8)	[63]
	Construct several distinct ML models to predict winter rapeseed yield at specific timepoints from six agro-morphological traits (oil and protein content, seed yield, oil and protein yield, and thousand seed weight) inputs.	ANN RF	- (0.944)	[64,65]
	Examination of micro-topographic attributes related to growth found in agronomic crops from analyses of vegetation indices, lidar derivatives, and crop type.	ANN	-	[66]
	Investigation of available water holding capacity of soil coupled with climate data and used to estimate average wheat yield within a region.	ANN	-	[67]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

	Cotton lint yield derived from a remote sensing ANN model evaluating eight phenological crop indices.	ANN	-	[68]
	Prediction of seed yield with ML and coupled PCA-ML models accompanied by sensitivity analysis based on agro-morphological features of sesame plants.	MLR PCA RBF GPR	0.00-0.36 (0.88-0.99)	*This work*
Nitrogen and Oil Concentration	Nitrogen prediction in oilseed rape leaves based on ten spectral features from both barley and oilseed rape.	ANN	0.30 (0.9)	[69,70]
	The merger of bio physiochemical and spectral features in leaves for further in-depth studies.	GP	2.2 - 5.8	[71]
	Prediction of sesame oil content from eighteen agro-morphological and phonological traits using ML in efforts to prevent marginal effects.	ANN MLR PCA	0.56 (0.86)	[72]
Disease and Quality Diagnoses for use in Classification	Integration of ML models coupled with hyperspectral imaging to detect disease in pre-symptomatic tobacco plants.	LS-SVM	-	[73]
	Conduction of oilseed disease analysis with directed surveys of ten common oilseed disease classes.	DT RF MLP	-	[74]
	Investigation of physicochemical properties (fatty acid, mineral profile) and additional physical attributes from six sunflower varieties for use in classification, grading, and quality assessment studies.	SVM, RF, MLR etc.	0.21 (0.81)	[75]

**Column indicates the best RMSE value or R^2 value (in parenthesis) obtained in the referenced study. DT (decision tree); GP (Gaussian process); k-NN (k-nearest neighbor); LS-SVM (Least-squares support-vector machines); PLSR (partial least squares regression); RF (radio frequency); SVR (support vector regression); PHN (pruning hidden nodes).*

Similarly, the RBF-NN is more efficient than the Multilayer Perceptron Neural (MLP) network for the prediction of rice yield in terms of training time, precision, and the number of neurons in the hidden layer [76]. Efficiency comparisons of the RBF and MLP neural networks with the MLR model indicated ANN models more accurately estimated biological and grain yield in barley [77]. Earlier works corroborated the greater ability of ANNs to predict wheat performance and to map and determine rice yield [67,78]. GPR is also used for yield prediction in agricultural products. Applying a set of random variables, GPR is capable of solving nonlinear problems using a novel data mining method [79]. Previous studies with MLR models are often centralized around model evaluation without checking generalizability. To overcome such drawbacks, this study aimed to evaluate the ability of some well-known machine learning models (e.g., ANNs and GPR) and coupled models such as PCA-ML to estimate plant yields, a gap currently found in the literature. As such, the main objectives of the present work are to estimate the agronomical yield of sesame using ML and coupled PCA-ML models to predict SSY and compare resultant data from each model. The primary motivation for and contributions of this study include: (i) Use the Principal Component Analysis (PCA) approach to simplify calculations and reduce the number of sesame production input variables,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

(ii) Assess ML and PAC-ML model generalizability and identify an optimal train and test dataset utilizing the K-fold approach, (iii) Employ MLR training and testing procedures to provide a comparison with ML models, (iv) Predict sesame yields with RBF-NN and GPR models for comparison to MLR model outputs, and (v) Apply the sensitivity analysis approach to uncover the features essential for the sesame seed yield.

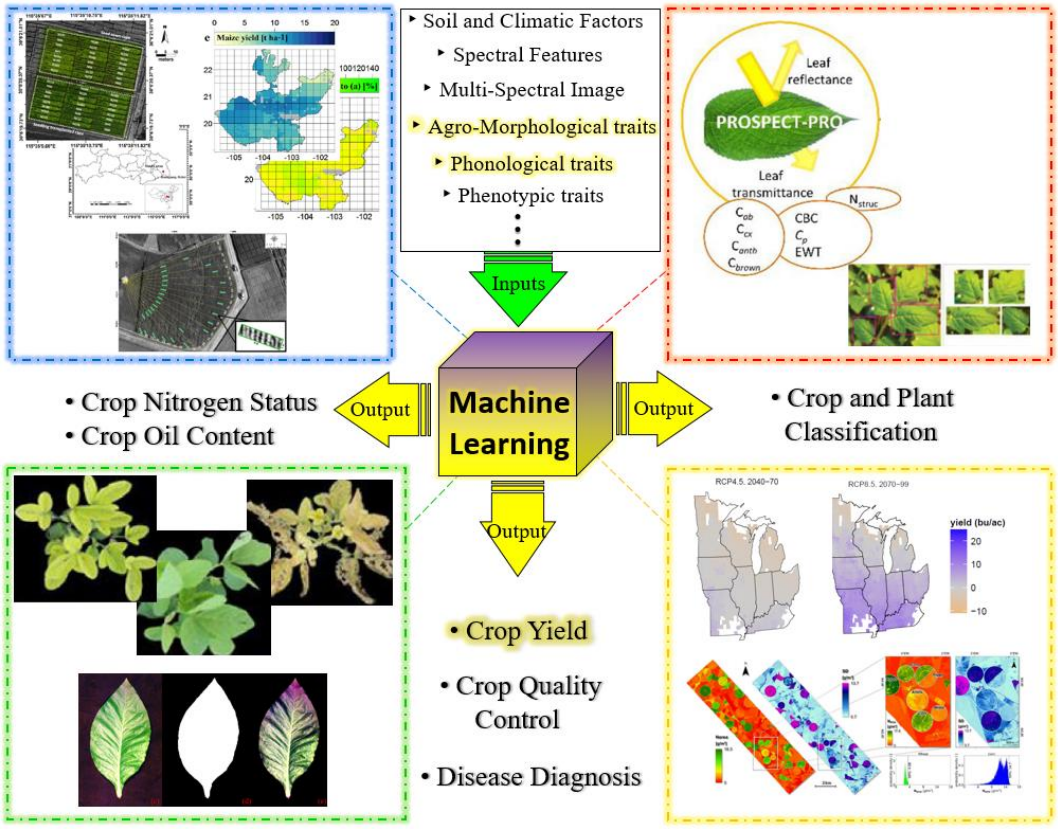


Fig. 1. Overview of the potential agricultural research questions that have been investigated using ML modeling methods [52,53,60,71,73,80–83].

2. Material and method

2.1. Field experiments

In this study, 135 sesame genotypes were derived from the five genotypes (selected from nine diverse genotypes) representing various sesame growing zones within Iran (Varamin 2822, Borazjan 1, Darab 1, Tn_{240}, and Tn_{234}). The selected genotypes were produced from various Iranian landraces planted on a research farm at the Shahrud University of Technology (36.39°N, 54.94°E) during 2018-2019. Each experimental plot consisted of two 1.5 m rows with 50 cm spacing between rows and 7 cm spacing between plants. The plots were fertilized with 80 kg ha^{-1} N and 100 kg ha^{-1} P before sowing and 40 kg ha^{-1} N at flower initiation. The crops were grown in a clay loam Typic Haplargid aridisol at pH 7.5 with 1% organic matter. Agro-morphological data were amassed from the F1 and F2 progeny (81 F1 and 45 F2) arranged in a randomized complete block design (RCBD). As

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

illustrated in **Fig. 2**, the provided datasets are applied to construct the ML models to predict the seed yields and importance level of input features.

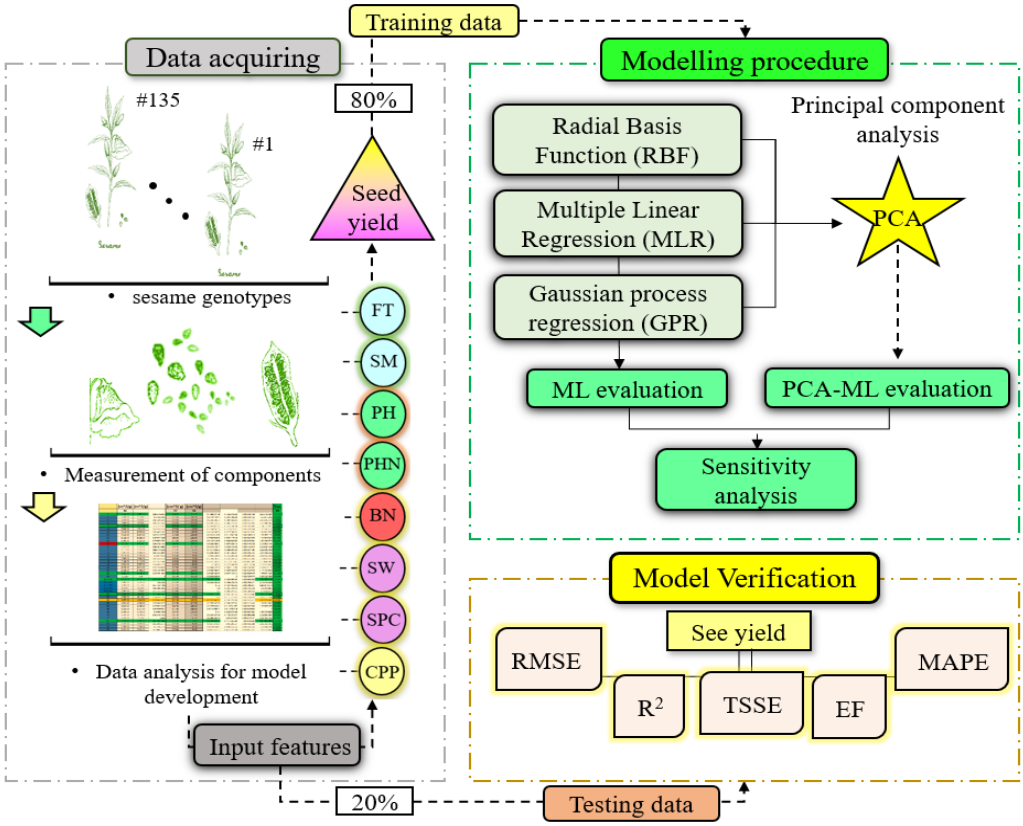


Fig. 2 The schematic of datasets provision, modeling, and model verification processes for seed yields prediction.

The experiment was repeated in triplicate and the agro-morphological traits (Independent variables) are illustrated in **Table 2.** These variables include flowering time at 10% (FT-10) and 100% (FT-100), seed maturity (SM), plant height (PH), the height of the first fruit-bearing node (PHN), number of fruit-bearing branches (BN), capsule number per plant (CPP), seed number per capsule (SPC), and 1000-seed weight (TSW). These input features were measured on 10 plants randomly from each sampled plot. The remaining plants were used after eliminating marginal effects to determine output as sesame seed yield (SSY). Input variables are defined $x_{1\ to\ 9}$ while (y), sesame seed yield, is designated as output. A summary of the measurements obtained is presented in **Table 2**. Visual data were collected daily from each plot for FT10, FT100, and SM. SSY data were obtained by harvesting two rows from the middle of the experimental plot.

2.2. Machine learning algorithms and models

2.2.1. Multiple linear regression (MLR) models

The study assesses four MLR models include linear, interaction (2FI), a reduced quadratic and a quadratic models as can be seen in equations 1 to 4, respectively. [84]. The ANOVA regression coefficients were determined by MATLAB software 2019a (MathWorks Inc., Natick, Massachusetts, USA).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

$$y = \beta_0 + \sum_{i=1}^{9} \beta_i x_i + \varepsilon \quad (1)$$

$$y = \beta_0 + \sum_{i=1}^{9} \beta_i x_i + \sum_{i=1}^{9} \sum_{j=i+1}^{9} \beta_{ij} x_{ij} + \varepsilon \quad (2)$$

$$y = \beta_0 + \sum_{i=1}^{9} \beta_i x_i + \sum_{i=1}^{9} \beta_{ii} x_{ii} + \varepsilon \quad (3)$$

$$y = \beta_0 + \sum_{i=1}^{9} \beta_i x_i + \sum_{i=1}^{9} \sum_{j=i+1}^{9} \beta_{ij} x_{ij} + \sum_{i=1}^{9} \beta_{ii} x_{ii} + \varepsilon \quad (4)$$

where y is crop yield (t/ha^{-1}), x represents the independent variables (inputs), β_0 is the intercept, β_i is the linear regression coefficient, β_{ij} is the interactions regression coefficient, and β_{ii} is the quadratic regression coefficient.

2.2.2. Principal component analysis (PCA)

PCA is a tool for reducing model calculations and input vector dimensions. The basis of PCA is the transformation of vast amounts of model input data to a smaller set of new variables (principal component) with lower autocorrelations [85,86]. The procedure is as follows:

$$X = X - \bar{X} \quad (5)$$

$$C = \frac{1}{n} X X^T \quad (6)$$

$$D = V^{-1} C V \quad (7)$$

$$Z = XV \quad (8)$$

where X is the input vector matrix, C is the covariance matrix, V is the eigenvector of C, D is the vector of eigenvalues C, and Z is the eigenvector of X.

2.2.3. Gaussian process regression (GPR) model

The GPR model is a non-parametric kernel-based probabilistic regression framework, which infers functions from a set of training data $D = \{(y_n, x_n), n = 1, 2, 3, \ldots, N\}$ of N vector input pairs, $x_n \hat{I} R^L$ and output y_n within a noisy scalar field. Effective for smaller datasets, this Bayesian model effectively generalizes output distribution at unrevealed input zones. Output noise, or model uncertainty, is often caused by external factors unrelated to x such as observation errors. The model noise assumption is zero-mean and defined:

$$y = f(x) + \varepsilon, \quad \varepsilon \gg N(0, \sigma_{noise}^2) \quad (9)$$

where σ_{noise}^2 equals the noise variance. GPR utilizes the Gaussian process (GP) to describe a latent variable function, referred to as f, with x used to describe and index-related latent variables in a finite

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

collection $\{f(x_1), \ldots, f(x_k)\}$ where the aforementioned indices constitute a consistent normal or "Gaussian" distribution. This allows for nonlinear regression between latent variable pairs. There are several advantages to GPR such as the ability to estimate model uncertainty the ability to use estimations to specify function types. The mean function *m(x)* and k(x,x') which equals the kernel, or covariance function where *E* equals expectation is defined:

$$m(x) = E[f(x)], k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))] \quad (10)$$

Typically defined as either zero or the dataset mean, the mean function is typically constant. The mean function is significant only about the average behavior of the model over time while the covariance function, a more comprehensive value, includes all procedure observations. The covariance function most often uses a hierarchical model, where covariance parameters called hyperparameters define the distribution *f*(*x*). The squared exponential covariance function employed to generate a smooth path is defined:

$$K(x, x') = K(x, x') = q1\exp - (\|x - x'\|)2(\mathrm{q2}) \quad (11)$$

A stationary covariance function and Euclidian norm, $\|.\|$, is a function of $x - x'$ and is invariant to changes in the input or x-space. With an increase in x-space distance, or the space between x and x', decay in covariance escalates exceptionally quickly. This implies correlations between *f*(*x*) and *f*(*x'*) are negligible. The hyperparameter $q1$ specifies the maximum permissible covariance while $q2$ defines the rate of decay as correlation distance increases. The covariance matrix represents the relatedness of one observation to another based on a set of kernel parameters. It can be defined:

$$K(x_i, x_j) = \sigma_f^2 \exp\left(-(x_i - x_j)^2 / 2l^2\right) + \sigma_n^2 \delta(x_i, x_j) \quad (12)$$

where σ_f^2 is the maximum acceptable covariance, l is the covariance matrix length parameter, and $\delta(x_i, x_j)$ is the Kronecker delta function. The covariance matrix is assessed during the GPR training process and then, then the training dataset output is estimated.

2.2.4. Radial basis function (RBF) neural network

The RBF model is a flexible feed-forward network able to automatically predict and classify new output patterns after the training phase [87,88]. The structure of the RBF model applied in the present study is illustrated in **Fig. 3**. The independent variables serve as first layer inputs while second (hidden) layer inputs are subjected to nonlinear activator function $\phi(r)$ before all are tallied together in the third or output layer. RBF is effective at approximating non-linear input-output mapping and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

has a strong tolerance to input noise. The optimum values of the matrix for the weight (W) and other model parameters are acquired during the training stage by minimizing the sum of squared error (SSE). RBF model output, (y), is defined as:

$$y(x) = \sum_{k=1}^{m} w_k \emptyset(\|x - c_k\|) \quad (13)$$

where, w_k is the linking weight of k^{th} neuron from the hidden layer to the output layer, and c_k is the pre-pattern center of k^{th} neuron from the hidden layer. The Gaussian Radial Basis Function is denoted:

$$\emptyset(r) = e^{-\frac{r^2}{2\sigma^2}} \quad (14)$$

where, r is the distance between input and pattern center (c), and σ is a parameter controlling the smoothness of the interpolation function [89,90].

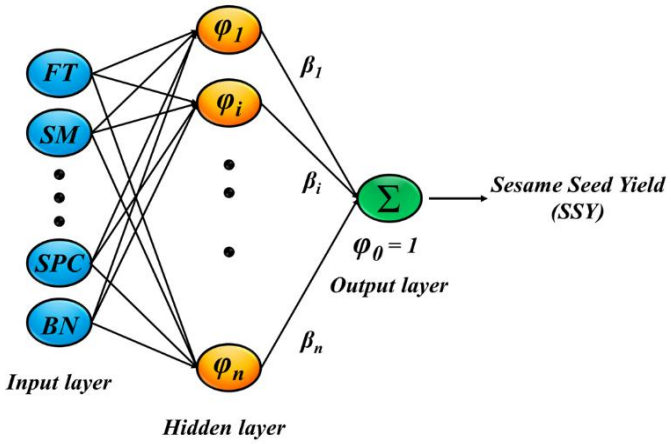


Fig. 3. The RBF neural network structure is applied in the present study.

2.3. K-fold cross-validation

K-fold cross-validation is used as a reliable approach to rule out model bias. To train MLR, GPR, and RBF models, the dataset is randomly divided into subsets for training (80%) and testing (20%). Since the input data partitioning is randomly done, the model gives different results for each training and testing run. Types of cross-validation differ such as k-fold cross-validation, K×2 cross-validation, leave-one-out cross-validation, repeated random sub-sampling validation, etc. [91]. Cross-validation involves repeated random subsampling procedures, where no overlapping occurs between the test datasets. During the process, the learning set is divided into equally sized k subgroups. The "fold" refers to the number of resulting sub-samples. Subsequently, one of the sub-samples is designated k and used as the test, or validation, dataset while the remaining k-1 sub-samples are utilized as training data. The first subsample selected becomes the first fold, and serves as a validation sample $D_{val,1}$ while successive sub-samples serve as the training set $D_{train,1}$. The ensuing result with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

the least error is named E_i. The outcomes for each k-fold are averaged or combined to propose a single estimation. The turning parameter for k-fold cross-validation was defined [92]. Each subgroup or $i = 1, 2, 3, \ldots, k$ help establish the fit model with γ or k-1 parts. Finally, $\alpha^{-k}(\gamma)$ is combined with additional computation k^{th} to identify prediction error and shown as:

$$E_k(\gamma) = \sum_{i \,\epsilon\, kth\ part} [y_i - x_i \alpha^{-k}(\gamma)]^2 \tag{15}$$

This procedure continues for numerous γ cycles and the value of γ displaying the smallest error will be selected.

2.4. Model assessment criteria

Models are assessed during the training and testing steps by multiple criteria including, Mean Absolute Percentage Error (MAPE), Root-Mean-Square Error (RMSE), Efficiency Factor (EF), and Total Sum Squared Error (TSSE). The linear relationship between actual (y) and predicted values ($\hat{Y}$) including their coefficients of determination (R^2) are defined as [93]:

$$MAPE = \left(\sum_{r=1}^{N} \frac{|Y_r - \hat{Y}_p|}{Y_r} \right) / N \tag{16}$$

$$RMSE = \sqrt{\left(\sum_{r=1}^{N} |Y_r - \hat{Y}_p|^2 \right) / (N-1)} \tag{17}$$

$$TSSE = \sum_{r=1}^{N} |Y_r - \hat{Y}_p|^2 \tag{18}$$

$$EF = 1 - \left(\sum_{j=1}^{n} (y_j - \hat{y}_j)^2 \Big/ \sum_{j=1}^{n} (y_j - \bar{y}_j)^2 \right) \tag{19}$$

$$R^2 = 1 - \left(\sum_{1}^{N} (Y_r - Y_p)^2 / \sum_{1}^{N} (Y_r - \hat{Y}_r)^2 \right) \tag{20}$$

where Y_r and Y_p represent the real and predicted values; and $\hat{Y}_r$ indicates the average of the real values. The MAPE, RMSE, and TSSE values closest to zero indicate the best ML model performance and provide accurate predictions with allowable estimation errors. EF values ascend as the model approaches an optimal state. For additional validation, the line spanning actual and predicted values (see Section 3.3) designates the best result is achieved when the slope and intercept approach 1 and 0, respectively, and the coefficient of determination (R^2) nears 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

3. RESULTS AND DISCUSSIONS

3.1. Primary statistical analysis on datasets

The generated datasets provided input for analysis of variance (ANOVA) with a general linear model. The ML models were configured using the QNet v2000 software package (QNet Ltd., Hong Kong, China) to define SSY as output and the other agro-morphological traits as inputs. The ML models were trained and tested using 135 samples from the study area. The corresponding statistical indices for each variable are illustrated in **Table 2**.

Table 2. Summary of measured statistical parameters for agronomic traits data.

Variables	Symbols	Input	Min	Max	Mean	Stdv
Flowering time 10% (days)	FT10	x1	41	55	47.25	2.57
Flowering time 100% (days)	FT100	x2	41	60	52.17	2.73
Seed maturity (days)	SM	x3	102	155	130.63	13.05
Plant height (cm)	PH	x4	100.86	199.2	143.45	16.14
Plant height to first fruiting node (cm)	PHN	x5	26.15	87.92	58.26	10.65
Capsule number per plant	CPP	x6	24.68	99.21	50.83	11.94
Thousand seed weight (g)	TSW	x7	1.98	4.22	3.41	0.36
Seed number per capsule (g)	SPC	x8	27.74	102.61	54.52	13.47
Branch number	BN	x9	0	5.3	1.84	1.08
Seed yield of sesame (t/ha)	SSY	y	1.19	4.18	2.52	0.58

3.2. Statistical processes of PCA and MLR algorithms

The present study utilized nine agro-morphological traits to estimate yield in sesame seeds. PCA was employed to improve speed, reduce calculation complexities, and decrease the number of independent variables by presenting the principal components. Five of the nine independent variables in **Fig. 4(a)** explained 98.99% of the total variance. Therefore, only five traits PH (x_4), SPC (x_8), CPP (x_6), SM (x_3), and PHN (x_5) were needed as principal components. **Fig. 4(b)** showed that two variables explained 69.50% of the total variance and were more influential than other variables as principal components. The morphological traits of 250 sesame genotypes have been evaluated with PCA. Shim et al. (2016) reported the first through fifth components accounted for 29%, 16%, 14%, 13%, and 10% of morphological trait variance. The primary traits of the first and second components were FT, SM, and SPC [94]. Furthermore, Baraki et al. (2015) used a group of 30 African sesame genotypes to construct a PCA. Three principal components explained 88.49% of the variance in the measured agronomic traits. Their study concluded seed yield and oil percentage have the greatest influence on principal components formation [95]. Ismaila and Usman (2014) also reported three components applied PCA, an indication that those three components were accounted for 86.73% of the variance [96].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

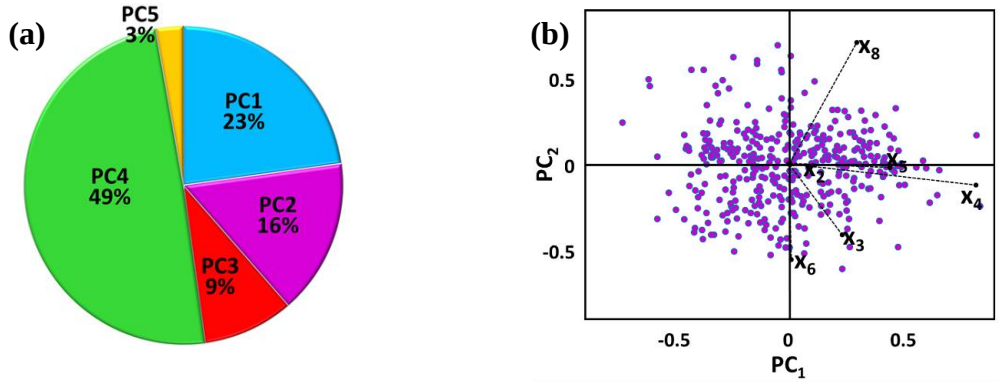


Fig. 4 (a) Variance of the five most influential principal components and (b) the relationship between the two first principal components PC_1, and PC_2.

The results from four regression models (linear, 2FI, quadratic, and reduced quadratic) are presented in **Table 3** to present the error rates for the training and testing stages. These error rates helped inform determinations of model efficiency after a thorough evaluation of all 50 training and testing datasets produced using the *k-fold* method. From these data, it is apparent the 2FI regression model was best used for all nine independent variables. The quadratic model performed best with principal components (PC_1 to PC_5) and represented the lowest mean and standard deviation of RMSE, MAPE, and EF. Limiting models to principal components PC_1 to PC_5 did not improve MLR model yield prediction. Model accuracy decreased in the testing rather than training phases (Table 3). Therefore, MLR models have an acceptable level of generalizability. Emamgholizadeh et al. (2015) applied an MLR model in an investigation of agronomic traits and yield in sesame [16].

Table 3. Comparison of yield prediction in four MLR model types before and after PCA exclusion.

Phase		Train			Test		
Model		RMSE	MAPE	EF	RMSE	MAPE	EF
Linear	no-PCA	0.29±0.01	8.04±0.80	0.82±0.01	0.29±0.03	8.36±0.80	0.80±0.04
	PCA	0.36±0.01	11.21±0.26	0.72±0.01	0.37±0.02	11.43±1.02	0.71±0.05
2FI	no-PCA	**0.21±0.01**	**5.82±0.15**	**0.90±0.01**	**0.25±0.02**	**7.02±0.68**	**0.86±0.02**
	PCA	0.35±0.01	10.27±0.32	0.74±0.02	0.38±0.04	11.11±1.14	0.67±0.10
Quadratic	no-PCA	0.21±0.01	5.68±0.15	0.90±0.01	0.26±0.02	7.15±0.67	0.84±0.33
	PCA	**0.31±0.00**	**8.97±0.26**	**0.80±0.01**	**0.34±0.03**	**9.89±1.03**	**0.75±0.05**
Reduced	no-PCA	0.25±0.01	6.52±0.22	0.86±0.01	0.26±0.03	6.96±0.80	0.84±0.03
quadratic	PCA	0.32±0.01	9.37±0.24	0.78±0.01	0.34±0.03	8.87±0.80	0.76±0.03

Mokarram and Bijanzadeh (2016) predicted barley (*Hordeum vulgare*) grain yield using percentage soil organic and grain/spike ratio as independent variables for MLR model inputs. They declared the machine learning model ANN (R^2 = 0.922) performed more accurately than MLR (R^2 = 0.784) [77]. The four transformation types shown in **Table 4** were assessed for their ability to improve MLR model performance. Comparative results between achieved results (Table 3 and 4), it is revealed that utilization of response variable transformation (y) did not enhance model prediction efficiency. Thus, the final MLR model is created without transformation. The *k-fold* method generated 50 different datasets from which to select the "optimal" dataset for assessment of model training and validation

steps. Results from subjecting this dataset to ANOVA with 2FI were presented in **Table 5**. Correlation tests between independent variables and yield noted an insignificant correlation between yield and the number of fruit-bearing branches (x_9) leading to its exclusion from the model. A p-value cutoff of 0.10 was employed to aid the selection of the final model and effective traits. This stringency level resulted in 11 traits (x_2, x_3, x_6, x_8, x_{12}, x_{16}, x_{18}, x_{37}, x_{67}, x_{68}, and x_{78}) selected of 36 using a stepwise regression process. Among main variables, days to 100% flowering (x_2), days to maturity (x_3), CPP (x_6), and SPC (x_8) are directly incorporated into the model. Other variables are included after viewing interactions of main independent variables. El-Mohsen (2013) discovered a relationship between yield and agro-morphological traits with a stepwise regression process that showed 77.25% of the variance in SSY was explained by the days to flowering and CPP [27]. Also, Parimala and Mathur (2006) indicated CPP was the most effective factor for predicting yield [25]. Yol et al. (2010) defined selection inputs based on PH, CPP, BN, and TSW to elucidate SSY [97].

Table 4. MLR model performance parameters for four different transformation types.

Phase	Train			Test		
TV	EF	MAPE	RMSE	EF	MAPE	RMSE
$\sqrt{y}$	0.90±0.00	5.99±0.17	0.22±0.00	0.84±0.03	7.32±0.70	0.27±0.02
lny	0.88±0.00	6.34±0.26	0.23±0.00	0.79±0.05	7.86±0.81	0.31±0.04
$1/y$	0.06±3.09	7.87±0.79	0.49±0.46	-0.85±5.80	11.11±4.40	0.92±1.42
y^2	0.84±0.01	6.65±0.12	0.27±0.01	0.75±0.20	9.58±0.91	0.28±0.32

**TV, transformation variable*

Table 5. MLR regression model ANOVA for the optimal training and validation dataset.

Source	DF	SS	P	Source	DF	SS	P	Source	DF	SS	P
Model	*36*	*135*	*0*	FT10*PP	**1**	**0.34**	**0.1**	SM*SPC	1	1.03	0.9
FT10	1	7.3	0.48	FT10*TSW	1	0	0.65	PH*PHN	1	0.01	0.47
FT100	**1**	**0.55**	**0.07**	FT10*SPC	**1**	**0.56**	**0.1**	PH*CPP	1	0.27	0.02
SM	**1**	**5.19**	**0.09**	FT100*SM	1	0	0.8	PH*TSW	1	0.04	0.93
PH	1	19.52	0.54	FT100*PH	1	0	0.79	PH*SPC	1	0.27	0.66
PHN	1	0.23	0.69	FT100* PHN	1	0	0.42	PHN*CPP	1	1.08	0.51
CPP	**1**	**36.36**	**0**	FT100*CPP	1	0.04	0.63	PHN*TSW	1	0.3	0.18
TSW	1	1.9	0.5	FT100*TSW	1	0	0.68	PHN*SPC	1	0	0.73
SPC	**1**	**51.08**	**0.02**	FT100*SPC	1	0.35	0.11	CPP*TSW	**1**	**0**	**0**
FT10*100	**1**	**1.05**	**0.01**	SM*PH	1	0	0.49	CPP*TSW	**1**	**3.95**	**0**
FT10*SM	1	0.17	0.32	SM*PHN	1	0.02	0.27	TSW*SPC	**1**	**1.43**	**0**
FT10*PH	1	0.31	0.37	SM*CPP	1	0.1	0.9	Residual	265	14.61	-
FT10*PHN	1	0.03	0.51	SM*TSW	**1**	**1.17**	**0.05**	Total	301	149.3	-

**Interaction function between the two independent variables; P-value (P); Sum of squares (SS)*

The percentage contribution (PC) for features applied of estimations of SSY is generated by dividing the sum of squares of each feature by the total sum of squares. Presented in **Fig. 5**, the PC of error is 9.84% in the training step with CPP (x_6) and SPC (x_8) displaying the highest PC. As interaction factors such as x_{12} (FT10*FT100), x_{16} (FT10*CPP), and x_{67} (CPP*TSW) have the lowest PC, it can be assumed the x_6 and x_8 variables are very effective parameters for SSY estimation. The FT10*SPC (x_{18}), SM*TSW (x_{37}), CPP*SPC (x_{68}), and TSW*SPC (x_{78}) interaction variables also contributed significantly to determining SSY.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

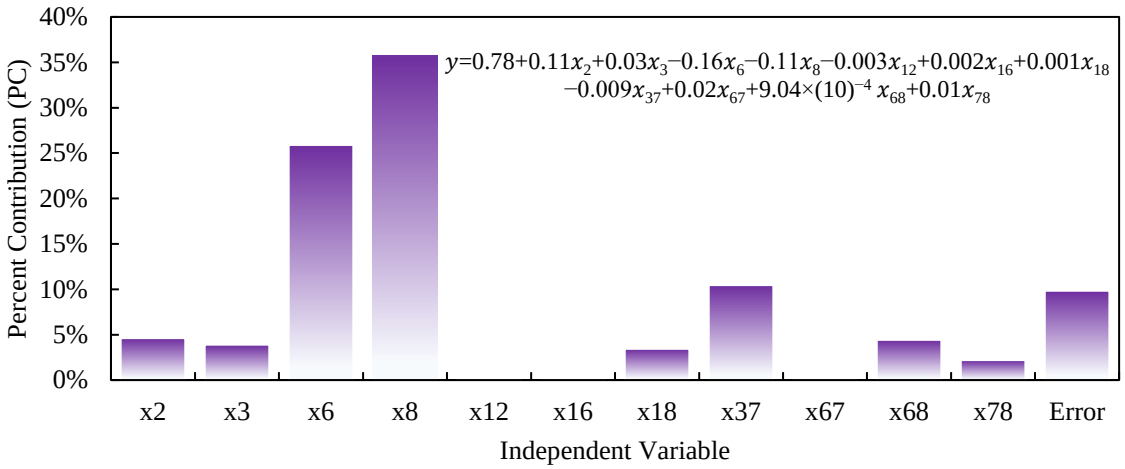


Fig. 5. Percent contribution (PC) for each MLR model factor and PC of error in the training phase.

3.3. ML models evaluation

Evaluation of model variables with and without PCA allowed for effective observation of input accuracy in predicting SSY. In addition to pinpointing specific input sets which can influence the effectiveness of ML model (MLR, GPR, and RBF) prediction, honing the hidden size number is crucial for RBF. Determining hidden layer neuron numbers (i.e., hidden layer size) is a key first step in reliable neural network model design. The trial-and-error process enables adjustments to neuron numbers by providing data akin to factorial analysis for optimization of hidden layer size (Table 6). The results of RMSE, MAPE, and EF at the training and test steps with the 50 unique datasets indicated the highest predicted performance of RBF was with 20 neurons in the hidden layer. The complete assessment of the MLR, GPR, and RBF models during the train, test, and combined phases are presented in **Table 6**. These data are displayed without PCA using the two classes of the original factors (see Equation 18) and with PCA incorporating principal components PC_1 to PC_5. The addition of PCA-based variables did not enhance MLR, GPR, and RBF model performance. Dataset performance in **Table 7** was gauged by considering the lowest RMSE, TSSE, and MAPE values and the highest EF value.

Table 6. RBF neural network performance factorial with varied hidden layer neuron numbers.

Phase	Criteria	Hidden layer size							
		5	10	15	**20**	25	30	35	40
Train	RMSE	0.22	0.21	0.21	**0.2**	0.2	0.2	0.2	0.2
	MAPE	6.31	5.99	5.85	**5.74**	5.81	5.75	5.8	5.81
	EF	0.89	0.9	0.9	**0.91**	0.91	0.91	0.91	0.91
Test	RMSE	0.26	0.25	0.26	**0.26**	0.26	0.26	0.26	0.26
	MAPE	7.31	7.09	7.24	**7.15**	7.14	7.17	7.16	7.21
	EF	0.86	0.85	0.86	**0.85**	0.85	0.85	0.85	0.85
Total	RMSE	0.23	0.22	0.22	**0.22**	0.22	0.22	0.22	0.22
	MAPE	6.51	6.21	6.13	**6.02**	6.08	6.04	6.07	6.09
	EF	0.89	0.89	0.89	**0.9**	0.9	0.9	0.9	0.9

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

Table 7. MLR, GPR, and RBF model assessments are based on initial and PCA-derived input variables.

-PCA*	Train			Test			Total		
Model	MLR	GPR	RBF	MLR	GPR	RBF	MLR	GPR	RBF
RMSE	0.23	**0.05±0.06**	0.20±0.01	0.23	0.26±0.03	0.26±0.02	0.23	0.13±0.03	0.22±0.01
TSSE	16.25	**2.17±3.48**	13.10±1.74	3.90	5.46±1.27	5.27±1.05	20.15	7.63±3.45	18.38±1.30
MAPE	6.29	**1.37±1.83**	5.74±0.38	5.89	7.11±0.71	7.15±0.76	6.21	2.52±1.44	6.02±0.25
EF	0.89	**0.98±0.02**	0.91±0.01	0.90	0.85±0.03	0.85±0.03	0.89	0.95±0.01	0.90±0.01

+PCA*	Train			Test			Total		
Model	MLR	GPR	RBF	MLR	GPR	RBF	MLR	GPR	RBF
RMSE	0.36	**0.21±0.09**	0.30±0.01	0.32	0.34±0.02	0.30±0.01	0.35	0.25±0.04	0.30±0.01
TSSE	38.92	**16.89±8.77**	27.88±2.95	7.77	**9.16±1.49**	27.88±2.95	46.69	26.06±8.17	27.88±2.95
MAPE	11.01	**6.31±2.84**	8.89±0.42	80.55	9.97±0.91	8.89±0.42	10.52	7.05±2.22	8.89±0.42
EF	0.74	**0.88±0.05**	0.81±0.01	0.80	0.74±0.03	0.81±0.01	0.75	0.86±0.04	0.81±0.01

**-, excluding PCA; +, including PCA*

Results of GPR and RBF model performance in **Table 8** were more accurate than GPR-PCA and RBF-PCA in both training and testing steps. Since the training step performance of GRP is reliable, no overfitting problems are observed. Although the previous results (Table 7) are acceptable, hidden layer optimization procedures improve the performance accuracy of GPR and RBF models. Also, as reported with the previous models, RBF model prediction performance is improved with the use of the most influential variables than with principal components PC_1 to PC_5 (RBF-PCA) as neural network inputs.

Table 8. RBF and GPR model performance with the optimal hidden layer neuron number.

Phase	Train				Test				Total			
Model	RBF	RBF - PCA	GPR	GPR - PCA	RBF	RBF - PCA	GPR	GPR - PCA	RBF	RBF - PCA	GPR	GPR - PCA
RMSE	0.21	0.30	0.00	0.29	0.23	0.36	0.21	0.30	0.21	0.31	0.09	0.29
TSSE	12.75	27.17	0.00	26.01	4.08	9.77	3.22	6.75	16.83	36.94	3.22	32.76
MAPE	5.72	8.75	0.02	8.60	6.39	9.89	5.48	9.09	5.86	8.98	1.12	8.70
EF	0.91	0.81	0.99	0.83	0.91	0.78	0.90	0.81	0.91	0.80	0.98	0.82

3.4. Prediction of sesame seed yield-based ML models

The distribution of predicted and actual SSY values in **Fig. 6** between ML and ML-PCA models reveals similar performances between the training and testing steps. Fit results for the MLR and model are illustrated in **Fig. 6(a and b)** with an R^2 value of 0.89 and 0.90 in the training and testing steps, respectively. Distribution patterns of actual and predicted data in **Fig. 6(c and d)** for the GPR and GPR-PCA models were comparable to those of MLR for the test phase (R^2 = 0.89), but the test phase was ten percent higher (R^2 = 0.99) training phase than the test phase. An R^2 value of 0.91 was presented in **Fig. 6(e and f)** for the train and test phases of the RFP model. The ML-PCA models all exhibited lower R^2 values in the training and testing phases for the prediction of SSY, an indication that the ML models were less precise when combined with the PCA technique. Based on the R^2 values reported here, the use of agro-morphological features is an acceptable method for use in models to predict SSY. The frequency distributions in **Fig. 7** aid evaluations of ML and ML-PCA model performance across several error ranges.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

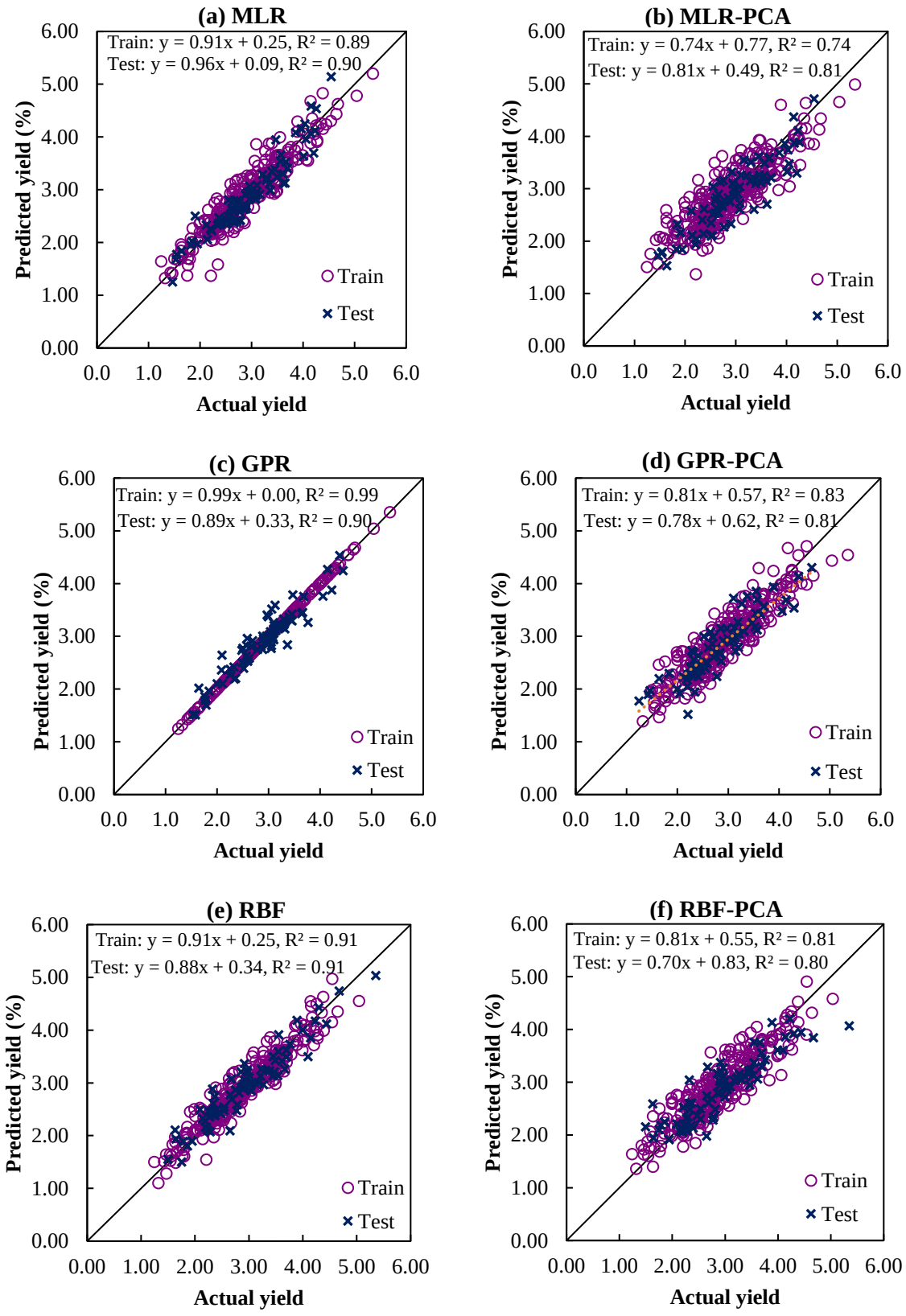


Fig. 5. Correlations between actual and predicted SSY in three ML and ML-PCA models.

MLR model error ranged from -0.24 to 0.84 with 78% of error falling between -0.24 and 0.30. Error for the MLR-PCA model ranged from -0.30 to 1.02 with 60% of error between -0.30 to 0.36. Error distribution for the GPR and GPR-PCA models implies the GPR model is superior to the GPR-PCA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

model. Approximately 94% of errors in the GPR model are between -0.19 to 0.17 while the range for all errors is from -0.19 to 0.53. Additionally, error distribution of the RBF and RBF-PCA models are provided in three intervals like other MLs as abovementioned. Approximately 71% of errors in the RBF model fall between -0.15 and 0.26 while 74% of errors fall between -0.21 to 0.54 in the RBF-PCA model.

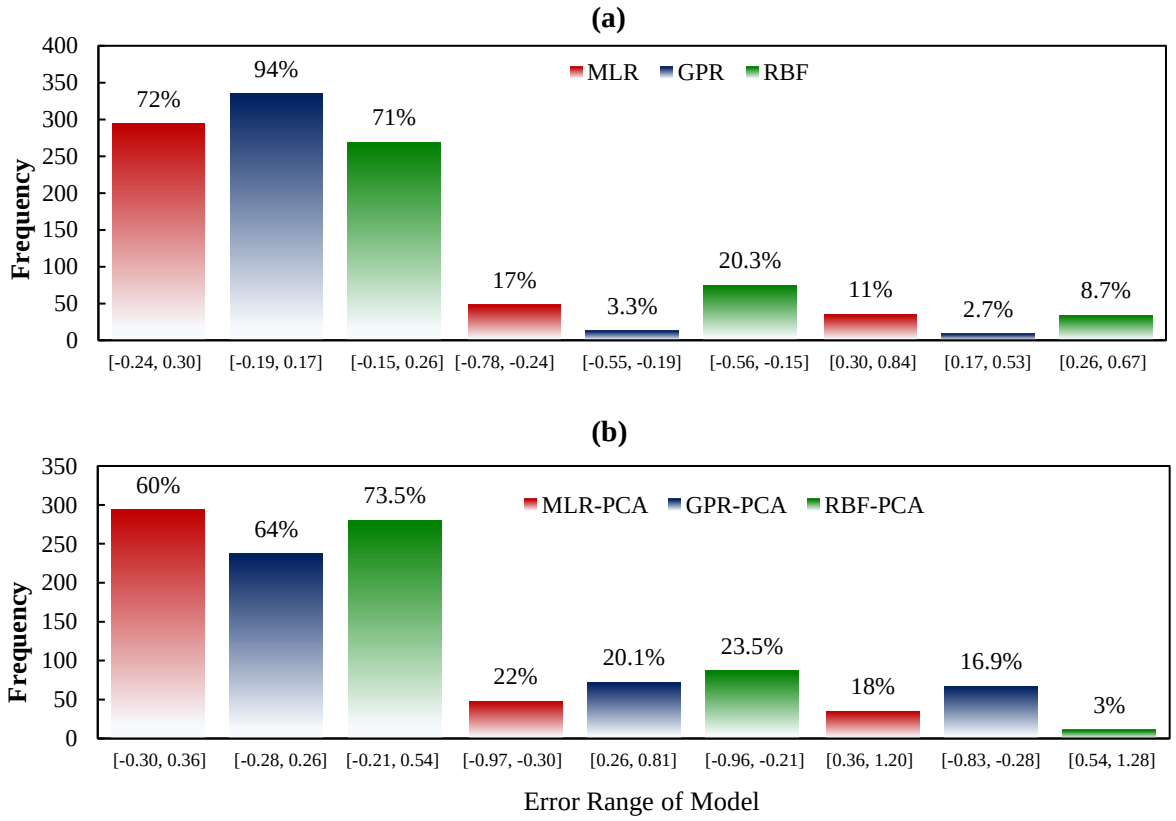


Fig. 7. Frequency distribution of MLR, GPR, and RBF error for (a) initial and (b) PCA independent variables.

3.5. Sensitivity analysis

Uncertainty in ML modeling is best visualized with a sensitivity analysis. Conventionally categorized into one of two classes, global or local, sensitivity analyses determine how data results change in response to alterations in ML models or methods [98,99]. A local sensitivity analysis only surveys a specific area while leaving the remaining inputs or independent variables unexplored. Whereas, global sensitivity analyses simultaneously incorporate variance from all independent variables to note a nonlinear trend across a range of variables [100]. These tests allowed for ease in distinguishing the comparative importance of each input variable on SSY. As illustrated in **Fig. 8,** the GPR, MLR, and RBF model sensitivity indices are indicative of model performance after the exclusion of agro-morphological variables as inputs. Measurements of sensitivity analysis data were collected and presented with RMSE, MAPE, and EF. Each factor (x_1 to x_9) was individually excluded to isolate the level of sensitivity for each variable. Manipulation of the CPP (x_6), SPC (x_8), and TSW

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2022 doi:10.20944/preprints202208.0522.v1

(x_7) variables were the most influential on SSY precision, with error rates between 0.29 and 0.55 for RMSE in **Fig. 8(a)**.

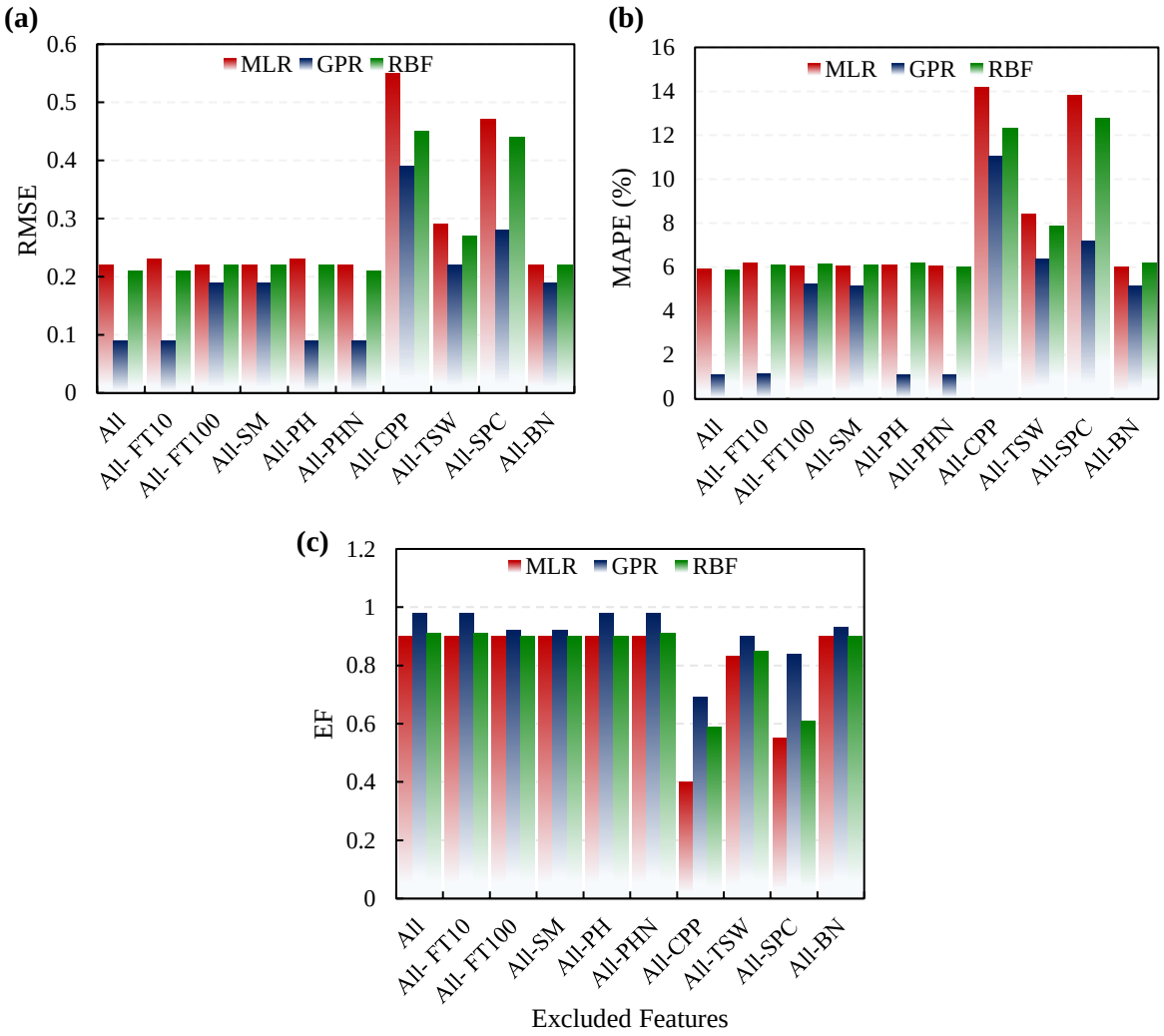


Fig. 8. Sensitivity analysis results for MLR, GPR, and RBF models on SSY.

The most sensitive ML models for each of these variables were MLR, RBF, and GPR, respectively. When these ML models were assessed based on the MAPE index, **Fig. 8(b)** revealed similar results. The highest MAPEs, those from CPP (14.2-11.06%), SPC (13.81-7.18%), and TSW (8.43-6.37%); were recorded after MLR, RBF, and GPR model exclusion. The results of EF in **Fig. 8(c)** exposed a trending decrease when CPP, SPC, and TSW were removed from model input. Notable errors were revealed in other agro-morphological indicators providing further evidence of the necessity for careful variable selection. The obtained model errors illustrate the fact that their exclusion reduces the efficiency of the ML models including MLR, GPR, and RBF by 50%.

4. Conclusion

Here, multi-machine learning (ML) approaches have been applied to generate predictive seed yield in sesame. Initial study began with the MLR model and PCA analysis combined with data from

existing study efforts. Nine agro-morphological factors contributed to the establishment of multi-ML techniques including MLR, GPR, and RBF for prediction of SSY. Moreover, the coupled PCA-ML models incorporated five principal components from nine primary variables and, for prediction accuracy, were compared with the original ML models. The results obtained here suggest SSY can be effectively estimated by the GPR, RBF, and MLR models. The SSY estimated by GPR, RBF and MLR had a RMSE of 0 to 0.05 t/ha, 0.20 to 0.23 t/ha, and 0.23 to 0.36 t/ha, respectively. The best performance was displayed by the MLR model subject to 2FI of inputs. RMSE declined in the GPR and RBF models after the k-fold method was employed to assist in dataset selection for the training and validation steps. The GPR model predicted SSY more accurately and precisely than MLR or RBF. Use of global sensitivity analysis indicated CPP, SPC, and TSW were the most sensitive factors for sesame yield estimation. These findings could be vital to efforts to promote productivity by planting more robust sesame species. The agro-morphological characteristics investigated here can be coupled with phenolic and spectral data analyses in future research studies for a multi-objective modeling approach. Additional studies of traits such as yield and quality level can provide data necessary to formulate a solid and sustainable plan to overcome several primary in oilseed cropping systems.

Acknowledgment

The authors acknowledge and appreciate the funding and technical supports provided by the Ferdowsi University of Mashhad, Iran for this project.

References

[1] Bassegio D, Zanotto MD, Santos RF, Werncke I, Dias PP, Olivo M. Oilseed crop crambe as a source of renewable energy in Brazil. Renew Sustain Energy Rev 2016;66:311–21. https://doi.org/https://doi.org/10.1016/j.rser.2016.08.010.

[2] Mousavi-Avval SH, Shah A. Techno-economic analysis of hydroprocessed renewable jet fuel production from pennycress oilseed. Renew Sustain Energy Rev 2021;149:111340. https://doi.org/https://doi.org/10.1016/j.rser.2021.111340.

[3] Ikegami M, Wang Z. Does energy aid reduce CO2 emission intensities in developing countries? J Environ Econ Policy 2021:1–16.

[4] Agidew MG, Dubale AA, Atlabachew M, Abebe W. Fatty acid composition, total phenolic contents and antioxidant activity of white and black sesame seed varieties from different localities of Ethiopia. Chem Biol Technol Agric 2021;8:14. https://doi.org/10.1186/s40538-021-00215-w.

[5] Muthulakshmi C, Sivaranjani R, Selvi S. Modification of sesame (Sesamum indicum L.) for

Triacylglycerol accumulation in plant biomass for biofuel applications. Biotechnol Reports 2021;32:e00668. https://doi.org/https://doi.org/10.1016/j.btre.2021.e00668.

[6] Benami E, Jin Z, Carter MR, Ghosh A, Hijmans RJ, Hobbs A, et al. Uniting remote sensing, crop modelling and economics for agricultural risk management. Nat Rev Earth Environ 2021;2:140–59.

[7] WANG Y, LI X, Lee T, PENG S, DOU F. Effects of nitrogen management on the ratoon crop yield and head rice yield in South USA. J Integr Agric 2021;20:1457–64.

[8] Hiremath SC, Patil CG, Patil KB, Nagasampige MH. Genetic diversity of seed lipid content and fatty acid composition in some species of Sesamum L.(Pedaliaceae). African J Biotechnol 2007;6.

[9] Uzun B, Arslan Ç, Furat Ş. Variation in fatty acid compositions, oil content and oil yield in a germplasm collection of sesame (Sesamum indicum L.). J Am Oil Chem Soc 2008;85:1135–42.

[10] Han L, Li J, Wang S, Cheng W, Ma L, Liu G, et al. Sesame oil inhibits the formation of glycidyl ester during deodorization. Int J Food Prop 2021;24:505–16.

[11] Karrar E, Ahmed IAM, Manzoor MF, Ammar A-F, Wei W, Albakry Z, et al. Effect of roasting pretreatment on fatty acids, oxidative stability, tocopherols, and antioxidant activity of gurum seeds oil. Biocatal Agric Biotechnol 2021;34:102022.

[12] Mahmood T, Mustafa HS Bin, Aftab M, Ali Q, Malik A. Super canola: newly developed high yielding, lodging and drought tolerant double zero cultivar of rapeseed (Brassica napus L.). Genet Mol Res 2019;18.

[13] Tadesse T, Singh H, Weyessa B. Correlation and path coefficient analysis among seed yield traits and oil content in Ethiopian linseed germplasm. Int J Sustain Crop Prod 2009;4:8–16.

[14] Solanki ZS, Gupta D. Inheritance studies for seed yield in sesame. Sesame Safflower Newsl 2003:25–8.

[15] Khan MA, Mirza MY, Akmal M, Ali N, Khan I. Genetic parameters and their implications for yield improvement in sesame. Sarhad J Agric 2007;23:623.

[16] Emamgholizadeh S, Parsaeian M, Baradaran M. Seed yield prediction of sesame using artificial neural network. Eur J Agron 2015;68:89–96.

[17] Shastry A, Sanjay HA, Bhanusree E. Prediction of crop yield using regression techniques. Int J Soft Comput 2017;12:96–102.

[18] Sellam V, Poovammal E. Prediction of crop yield using regression analysis. Indian J Sci Technol 2016;9:1–5.

[19] Ramesh D, Vardhan BV. Analysis of crop yield prediction using data mining techniques. Int J Res Eng Technol 2015;4:47–473.

[20] Chowdhury S, Datta AK, Saha A, Sengupta S, Paul R, Maity S, et al. Traits influencing yield in sesame (Sesamum indicum L.) and multilocational trials of yield parameters in some desirable plant types. Indian J Sci Technol 2010;3:163–6.

[21] Sengupta S, Datta AK. Genetic studies to ascertain selection criteria for yield improvement in sesame. J Phytol Res 2004;17:163–6.

[22] Shim KB, Kang CW, Lee SW, Kim DH, Lee BH. Heritabilities, genetic correlations and path coefficients of some agronomic traits in different cultural environments in sesame. Sesame Safflower Newsl 2001:16–22.

[23] Boureima S, Diouf S, Amoukou M, Van Damme P. Screening for sources of tolerance to drought in sesame induced mutants: Assessment of indirect selection criteria for seed yield. Int J Pure Appl Biosci 2016;4:45–60.

[24] Ganesh SK, Sakila M. Association analysis of single plant yield and its yield contributing characters in sesame (Sesamum indicum L.). Sesame Safflower Newsl 1999:16–9.

[25] Parimala K, Mathur RK. Yield component analysis through multiple regression analysis in sesame. Int J Agric Res 2006;2:338–40.

[26] Shim K-B, Kang C-W, Seong J-D, Hwang C-D, Suh D-Y. Interpretation of relationship between sesame yield and it's components under early sowing cropping condition. 한국작물학회지 2006;51:269–73.

[27] Abd El-Mohsen AA. Comparison of some statistical techniques in evaluating Sesame yield and its contributing factors. Scientia 2013;1:8–14.

[28] Soltanali H, Rohani A, Tabasizadeh M, Abbaspour-Fard MH, Parida A. An improved fuzzy inference system-based risk analysis approach with application to automotive production line. Neural Comput Appl 2020;32:10573–91.

[29] Shin M, Ithnin M, Vu WT, Kamaruddin K, Chin TN, Yaakub Z, et al. Association mapping analysis of oil palm interspecific hybrid populations and predicting phenotypic values via machine learning algorithms. Plant Breed 2021;140:1150–65.

[30] Wen G, Ma B-L, Vanasse A, Caldwell CD, Earl HJ, Smith DL. Machine learning-based canola yield prediction for site-specific nitrogen recommendations. Nutr Cycl Agroecosystems 2021;121:241–56.

[31] Sharma R, Kamble SS, Gunasekaran A, Kumar V, Kumar A. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. Comput Oper Res 2020;119:104926. https://doi.org/10.1016/j.cor.2020.104926.

[32] Rahimi M, Abbaspour-Fard MH, Rohani A. A multi-data-driven procedure towards a comprehensive understanding of the activated carbon electrodes performance (using for supercapacitor) employing ANN technique. Renew Energy 2021;180:980–92.

https://doi.org/10.1016/J.RENENE.2021.08.102.

[33] Xu H, Zhang X, Ye Z, Jiang L, Qiu X, Tian Y, et al. Machine learning approaches can reduce environmental data requirements for regional yield potential simulation. Eur J Agron 2021;129:126335. https://doi.org/https://doi.org/10.1016/j.eja.2021.126335.

[34] Wang X, Miao Y, Dong R, Zha H, Xia T, Chen Z, et al. Machine learning-based in-season nitrogen status diagnosis and side-dress nitrogen recommendation for corn. Eur J Agron 2021;123:126193. https://doi.org/https://doi.org/10.1016/j.eja.2020.126193.

[35] Rahimi M, Abbaspour-Fard MH, Rohani A, Yuksel Orhan O, Li X. Modeling and Optimizing N/O-Enriched Bio-Derived Adsorbents for CO2 Capture: Machine Learning and DFT Calculation Approaches. Ind Eng Chem Res 2022.

[36] Soltanali H, Nikkhah A, Rohani A. Energy audit of Iranian kiwifruit production using intelligent systems. Energy 2017;139:646–54.

[37] Nikkhah A, Rohani A, Rosentrater KA, El Haj Assad M, Ghnimi S. Integration of principal component analysis and artificial neural networks to more effectively predict agricultural energy flows. Environ Prog Sustain Energy 2019;38:13130.

[38] Taki M, Mehdizadeh SA, Rohani A, Rahnama M, Rahmati-Joneidabad M. Applied machine learning in greenhouse simulation; new application and analysis. Inf Process Agric 2018;5:253–68.

[39] Bolandnazar E, Rohani A, Taki M. Energy consumption forecasting in agriculture by artificial intelligence and mathematical models. Energy Sources, Part A Recover Util Environ Eff 2020;42:1618–32.

[40] Rahimi M, Abbaspour-Fard MH, Rohani A. Synergetic effect of N/O functional groups and microstructures of activated carbon on supercapacitor performance by machine learning. J Power Sources 2022;521:230968. https://doi.org/https://doi.org/10.1016/j.jpowsour.2021.230968.

[41] Jayas DS, Paliwal J, Visen NS. Review paper (AE—automation and emerging technologies): multi-layer neural networks for image analysis of agricultural products. J Agric Eng Res 2000;77:119–28.

[42] Karimi Y, Prasher SO, McNairn H, Bonnell RB, Dutilleul P, Goel PK. Classification accuracy of discriminant analysis, artificial neural networks, and decision trees for weed and nitrogen stress detection in corn. Trans ASAE 2005;48:1261–8.

[43] Elizondo D, Hoogenboom G, McClendon RW. Development of a neural network model to predict daily solar radiation. Agric For Meteorol 1994;71:115–32.

[44] Mukerji A, Chatterjee C, Raghuwanshi NS. Flood forecasting using ANN, neuro-fuzzy, and neuro-GA models. J Hydrol Eng 2009;14:647–52.

[45] Rahimi M, Abbaspour-Fard MH, Rohani A. Machine learning approaches to rediscovery and optimization of hydrogen storage on porous bio-derived carbon. J Clean Prod 2021;329:129714. https://doi.org/https://doi.org/10.1016/j.jclepro.2021.129714.

[46] Jin Y-Q, Liu C. Biomass retrieval from high-dimensional active/passive remote sensing data by using artificial neural networks. Int J Remote Sens 1997;18:971–9.

[47] Kim M, Gilley JE. Artificial Neural Network estimation of soil erosion and nutrient concentrations in runoff from land application areas. Comput Electron Agric 2008;64:268–75.

[48] Fakoor Sharghi AR, Makarian H, Derakhshan Shadmehri A, Rohani A, Abbasdokht H. Predicting Spatial Distribution of Redroot Pigweed (Amaranthus retroflexus L.) using the RBF Neural Network Model. J Agric Sci Technol 2018;20:1493–504.

[49] Vakil-Baghmisheh M-T, Pavešić N. Premature clustering phenomenon and new training algorithms for LVQ. Pattern Recognit 2003;36:1901–12.

[50] Azadeh A, Ghaderi SF, Sohrabkhani S. Forecasting electrical consumption by integration of neural network, time series and ANOVA. Appl Math Comput 2007;186:1753–61.

[51] Bayati H, Najafi A. Performance comparison artificial neural networks with regression analysis in trees trunk volume estimation 2013.

[52] Peng Y, Zhu T, Li Y, Dai C, Fang S, Gong Y, et al. Remote prediction of yield based on LAI estimation in oilseed rape under different planting methods and nitrogen fertilizer applications. Agric For Meteorol 2019;271:116–25. https://doi.org/10.1016/j.agrformet.2019.02.032.

[53] Eugenio FC, Grohs M, Venancio LP, Schuh M, Bottega EL, Ruoso R, et al. Estimation of soybean yield from machine learning techniques and multispectral RPAS imagery. Remote Sens Appl Soc Environ 2020;20. https://doi.org/10.1016/j.rsase.2020.100397.

[54] Niedbała G, Nowakowski K, Rudowicz-Nawrocka J, Piekutowska M, Weres J, Tomczak RJ, et al. Multicriteria prediction and simulation of winter wheat yield using extended qualitative and quantitative data based on artificial neural networks. Appl Sci 2019;9. https://doi.org/10.3390/app9142773.

[55] Parmley KA, Higgins RH, Ganapathysubramanian B, Sarkar S, Singh AK. Machine Learning Approach for Prescriptive Plant Breeding. Sci Rep 2019;9:1–12. https://doi.org/10.1038/s41598-019-53451-4.

[56] Abdipour M, Younessi-Hmazekhanlu M, Ramazani SHR. Artificial neural networks and multiple linear regression as potential methods for modeling seed yield of safflower (Carthamus tinctorius L.). Ind Crops Prod 2019;127:185–94.

[57] Pandith V, Kour H, Singh S, Manhas J, Sharma V. Performance Evaluation of Machine

Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis. J Sci Res 2020;64:394–8. https://doi.org/10.37398/jsr.2020.640254.

[58] Abbas F, Afzaal H, Farooque AA, Tang S. Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy 2020;10. https://doi.org/10.3390/AGRONOMY10071046.

[59] Crane-Droesch A. Ac ce pte us Machine learning methods for crop yield prediction and 2018.

[60] Folberth C, Baklanov A, Balkovič J, Skalský R, Khabarov N, Obersteiner M. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. Agric For Meteorol 2019;264:1–15. https://doi.org/10.1016/j.agrformet.2018.09.021.

[61] Amaratunga V, Wickramasinghe L, Perera A, Jayasinghe J, Rathnayake U, Zhou JG. Artificial Neural Network to Estimate the Paddy Yield Prediction Using Climatic Data. Math Probl Eng 2020;2020. https://doi.org/10.1155/2020/8627824.

[62] Matsumura K, Gaitan CF, Sugimoto K, Cannon AJ, Hsieh WW. Maize yield forecasting by linear regression and artificial neural networks in Jilin, China. J Agric Sci 2015;153:399–410. https://doi.org/10.1017/S0021859614000392.

[63] Maya Gopal PS, Bhargavi R. A novel approach for efficient crop yield prediction. Comput Electron Agric 2019;165:104968. https://doi.org/10.1016/j.compag.2019.104968.

[64] Rajkovi D, Pezo L, Lonˇ B, Zanetti F, Monti A, Kondi A. Yield and Quality Prediction of Winter Rapeseed — Artificial Neural Network and Random Forest Models 2022.

[65] Niedbała G. Application of artificial neural networks for multi-criteria yield prediction ofwinter rapeseed. Sustain 2019;11. https://doi.org/10.3390/su11020533.

[66] Kross A, Znoj E, Callegari D, Kaur G, Sunohara M, Lapen DR, et al. Using artificial neural networks and remotely sensed data to evaluate the relative importance of variables for prediction of within-field corn and soybean yields. Remote Sens 2020;12. https://doi.org/10.3390/rs12142230.

[67] Alvarez R. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. Eur J Agron 2009;30:70–7.

[68] Haghverdi A, Washington-Allen RA, Leib BG. Prediction of cotton lint yield from phenology of crop indices using artificial neural networks. Comput Electron Agric 2018;152:186–97. https://doi.org/10.1016/j.compag.2018.07.021.

[69] Yu X, Lu H, Liu Q. Deep-learning-based regression model and hyperspectral imaging for rapid detection of nitrogen concentration in oilseed rape (Brassica napus L.) leaf. Chemom Intell Lab Syst 2018;172:188–93. https://doi.org/10.1016/j.chemolab.2017.12.010.

[70] Klem K, Křen J, Kováč D, Holub P, Míša P, Svobodová I, et al. Improving nitrogen status

estimation in malting barley based on hyperspectral reflectance and artificial neural networks 2021.

[71] Berger K, Verrelst J, Féret J-B, Hank T, Wocher M, Mauser W, et al. Retrieval of aboveground crop nitrogen content with a hybrid machine learning method. Int J Appl Earth Obs Geoinf 2020;92:102174. https://doi.org/10.1016/j.jag.2020.102174.

[72] Abdipour M, Ramazani SHR, Younessi-Hmazekhanlu M, Niazian M. Modeling oil content of sesame (Sesamum indicum L.) using artificial neural network and multiple linear regression approaches. J Am Oil Chem Soc 2018;95:283–97.

[73] Zhu H, Chu B, Zhang C, Liu F, Jiang L, He Y. Hyperspectral Imaging for Presymptomatic Detection of Tobacco Disease with Successive Projections Algorithm and Machine-learning Classifiers. Sci Rep 2017;7:1–12. https://doi.org/10.1038/s41598-017-04501-2.

[74] Thakur A, Thakur R. Machine Learning Algorithms for Oilseed Disease Diagnosis. SSRN Electron J 2019. https://doi.org/10.2139/ssrn.3372216.

[75] Çetin N, Karaman K, Beyzi E, Sağlam C, Demirel B. Comparative Evaluation of Some Quality Characteristics of Sunflower Oilseeds (Helianthus annuus L.) Through Machine Learning Classifiers. Food Anal Methods 2021;14:1666–81. https://doi.org/10.1007/s12161-021-02002-7.

[76] Saad P, Ismail N. Artificial neural network modelling of rice yield prediction in precision farming. Artif Intell Softw Eng Res Lab, Sch Comput Commun Eng North Univ Coll Eng (KUKUM), Jejawi, Perlis 2009.

[77] Mokarram M, Bijanzadeh E. Prediction of biological and grain yield of barley using multiple regression and artificial neural network models. Aust J Crop Sci 2016;10:895–903.

[78] Chen C, McNairn H. A neural network integrated approach for rice crop monitoring. Int J Remote Sens 2006;27:1367–93.

[79] Rezazadeh Joudi A, Sattari M. Estimation of scour depth of piers in hydraulic structures using gaussian process Regression. Irrig Drain Struct Eng Res 2016;16:19–36.

[80] Singh A, Ganapathysubramanian B, Singh AK, Sarkar S. Machine Learning for High-Throughput Stress Phenotyping in Plants. Trends Plant Sci 2016;21:110–24. https://doi.org/10.1016/j.tplants.2015.10.015.

[81] Bashier IH, Mosa M, Babikir SF. Sesame Seed Disease Detection Using Image Classification. Proc 2020 Int Conf Comput Control Electr Electron Eng ICCCEEE 2020 2021:1–5. https://doi.org/10.1109/ICCCEEE49695.2021.9429640.

[82] Sahni V, Srivastava S, Khan R. Modelling Techniques to Improve the Quality of Food Using Artificial Intelligence. J Food Qual 2021;2021. https://doi.org/10.1155/2021/2140010.

[83] Khairunniza Bejo S, Mustaffha S, Khairunniza-Bejo S, Ishak W, Ismail W. Application of

Artificial Neural Network in Predicting Crop Yield: A Review Spectroscopy techniques View project Application of Artificial Neural Network in Predicting Crop Yield: A Review. J Food Sci Eng 2014;4:1–9.

[84] Sarvestani NS, Rohani A, Farzad A, Aghkhani MH. Modeling of specific fuel consumption and emission parameters of compression ignition engine using nanofluid combustion experimental data. Fuel Process Technol 2016;154:37–43.

[85] Sarbu C, Pop H. Principal component analysis versus fuzzy principal component analysis. Talanta -Oxford Then Amsterdam- 2005;65:1215–20.

[86] Ilin A, Raiko T. Practical approaches to principal component analysis in the presence of missing values. J Mach Learn Res 2010;11:1957–2000.

[87] Hoang N-D, Pham A-D, Nguyen Q-L, Pham Q-N. Estimating compressive strength of high performance concrete with Gaussian process regression model. Adv Civ Eng 2016;2016.

[88] Yang Y-K, Sun T-Y, Huo C-L, Yu Y-H, Liu C-C, Tsai C-H. A novel self-constructing Radial Basis Function Neural-Fuzzy System. Appl Soft Comput 2013;13:2390–404. https://doi.org/https://doi.org/10.1016/j.asoc.2013.01.023.

[89] Tatar A, Shokrollahi A, Mesbah M, Rashid S, Arabloo M, Bahadori A. Implementing Radial Basis Function Networks for modeling CO2-reservoir oil minimum miscibility pressure. J Nat Gas Sci Eng 2013;15:82–92. https://doi.org/https://doi.org/10.1016/j.jngse.2013.09.008.

[90] Ashtiani S-HM, Rohani A, Aghkhani MH. Soft computing-based method for estimation of almond kernel mass from its shell features. Sci Hortic (Amsterdam) 2020;262:109071.

[91] Zareei J, Rohani A. Optimization and study of performance parameters in an engine fueled with hydrogen. Int J Hydrogen Energy 2020;45:322–36. https://doi.org/10.1016/j.ijhydene.2019.10.250.

[92] Jiang P, Chen J. Displacement prediction of landslide based on generalized regression neural networks with K-fold cross-validation. Neurocomputing 2016;198:40–7.

[93] Rahimi M, Pourramezan M-R, Rohani A. Modeling and classifying the in-operando effects of wear and metal contaminations of lubricating oil on diesel engine: A machine learning approach. Expert Syst Appl 2022:117494. https://doi.org/https://doi.org/10.1016/j.eswa.2022.117494.

[94] Shim KB, Shin SH, Shon JY, Kang SG, Yang WH, Heu SG. Classification of a collection of sesame germplasm using multivariate analysis. J Crop Sci Biotechnol 2016;19:151–5.

[95] Fiseha B, Yemane T, Fetien A. Assessing inter-relationship of sesame genotypes and their traits using cluster analysis and principal component analysis methods. Int J Plant Breed Genet 2015;9:228–37.

[96] Ismaila A, Usman A. Genetic Variability for Yield and Yield Components in Sesame (

 doi:10.20944/preprints202208.0522.v1

Sesamum indicum L .) 2014;3:2012–5.

[97] Yol E, Karaman E, Furat S, Uzun B. Assessment of selection criteria in sesame by using correlation coefficients, path and factor analyses. Aust J Crop Sci 2010;4:598–602.

[98] Sudret B. Global sensitivity analysis using polynomial chaos expansions. Reliab Eng Syst Saf 2008;93:964–79.

[99] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global sensitivity analysis: the primer. John Wiley & Sons; 2008.

[100] Wan H-P, Ren W-X, Todd MD. Arbitrary polynomial chaos expansion method for uncertainty quantification and global sensitivity analysis in structural dynamics. Mech Syst Signal Process 2020;142:106732.