Vol. 21, No. 2, pp. 445-456 (2021)
Journal of Agricultural Physics
ISSN 0973-032X
http://www.agrophysics.in

Research Article

Mustard Yield Prediction using Machine Learning Approach

AVINASH GOYAL AND ANANTA VASHISTH*

Division of Agricultural Physics, ICAR-Indian Agricultural Research Institute, New Delhi-110012

ABSTRACT

Mustard yield prediction was done by machine learning approach using long term weather and mustard yield data. Model was developed by variable selection using stepwise multiple linear regression (SMLR) and artificial neural network (ANN), support vector machine (SVM), random forest (RF) techniques, variable extraction using principal component analysis (PCA) and ANN, SVM, RF techniques, optimal combination of model developed by different techniques. Results showed that model developed either by variable selection by SMLR or variable extraction by PCA and using ANN, SVM, RF techniques, SVM model performed best for mustard yield prediction. Among all six models, PCA-SVM performed best for the study area. Performance of the model developed by optimum combination performed better than the individual. By comparing the performance of the model developed by different techniques and can be used for district level mustard yield prediction.

Key words: Weather variables, stepwise multiple linear regression, artificial neural network, support vector machine, random forest, yield prediction

Introduction

Accurate and timely forecast of crop yields is necessary for crop management and planning decisions regarding import, export etc. Crop yield is affected by extreme weather events and climatic variability. Considering the challenge of food security at domestic and international level, it is desirable to develop an accurate and dynamic crop yield prediction model. To overcome the problems of forecasting non-linearity and non-stationary time series dataset several machine learning modern techniques has been used such as Artificial Neural Network (ANN), Support Vector Machine (SVM) and Random Forest (RF). R statistical software version 3.1.3 was used for developing model for mustard yield prediction and making a comparison between the developed model to determine the best one among them. Vashisth *et al*. (2014) reported that percentage deviation of wheat yield prediction using weather based statistical model at 45 and 25 days before harvesting by observed yield was 10.7% and 7%, respectively. The most important technique for variable extraction is principal component analysis that removes homogeneity in variables and creates uncorrelated variables known as Principal component (PC). Azfar *et al*. (2015) developed the model using principal component analysis of weekly data on weather variables and found to be most appropriate for providing rapeseed and mustard yield forecast one and half months before the harvest for Faizabad district of UP. SVM method used to overcome the over fitting problem in input dataset. Gandhi *et al*. (2016) evaluated SVM model and predicted the rice crop yield with 78.76% accuracy for 27 districts of Maharashtra state, India. Balakrishnan and Muthukumarasamy (2016) reported that SVMs approach was a better option compared to Naive Bayes approach for prediction of different crops at Thanjavur district, Tamil Nadu. Su *et al*. (2017)

*Corresponding author,
Email: ananta.iari@gmail.com

developed SVM-based Open Crop Model to predict the growing stages of rice and seed yield at regional level. Karimi *et al*. (2008) found that SVM technique over stepwise regression technique was more accurate as compared to predict the biophysical parameters and yield of corn. Palanivel and Surianarayanan (2019) reviewed linear regression, Artificial Neural Network and Support Vector Machine and found that SVM based prediction models to be more suitable for crop yield prediction.

Random forest technique avoid the over fitting in training dataset. The key advantage of the random forest technique is that can investigate nonlinear and hierarchical relationships between the predictors and the response using an ensemble learning approach. Gromping (2009) reported that RF regression technique provide more promising result for highly correlated input variables such as weather parameters, crop management and soil properties. Fukuda *et al*. (2013) demonstrated Random Forest models to predict maximum and mean value of mango fruit yield under different irrigated and rainfed condition. Everingham *et al*. (2016) reported that accuracy of sugarcane yield prediction done by RF techniques using ten years weather data was between 86.4% to 95.5%. Jeong *et al*. (2016) reported that wheat yield prediction done by Random forest and multiple linear regressions at global scale had nRMSE value 16 and 33% respectively.

Song *et al*. (2011) recommended that the optimal combination of different empirical crop yield forecasting model is more reliable approach to overdrawn the limitation of individual approach at regional scales. Pandey *et al*. (1992) combined the ARIMA model and remote sensing based technique to improve the accuracy of wheat yield estimated at Hissar, Haryana. They reported that optimal combination reduced the RMSE and percent deviation compared to both individual approaches. The aim of this study is do develop mustard yield prediction model by variable selection (SR) using SMLR variable extraction (PC) using PCA and ANN, SVM, RF techniques, optimal combination of model developed by different techniques and evaluate the performance of these models for improving the accuracy of mustard yield prediction.

Materials and Methods

Long term mustard yield data as well as daily weather data such as maximum and minimum temperature, morning and evening relative humidity, rainfall, bright sunshine hour from 1984 to 2019 were collected from IARI, New Delhi during mustard crop growing period.

Mustard seed yield shows an increasing trend over a long time series data. The increasing trend is generally due to improvement in crop production technology over time, such as, introduction of high yielding/stress tolerant cultivars, higher applications of input resources, and better technology for intercultural operations (Aggarwal *et al*., 2000). Mustard yield data follows the increasing trend with time (Fig.1). To understand the behaviour of weather variables on mustard yield and overcome the technological effects a modification has been proposed here. "Scaled normalized yield" was calculated using the formula as below:

$$\text{Scaled normalized yield} = \frac{\text{Normalized yield}_i - \text{Normalized yield}_{min}}{\text{Normalized yield}_{max} - \text{Normalized yield}_{min}} \quad \ldots(1)$$

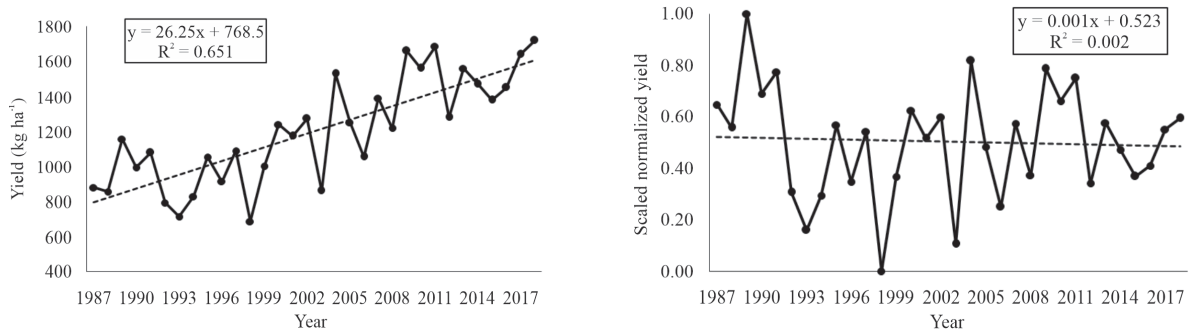

Fig. 1. Temporal plots of (a) Crop yield and (b) scaled normalized yield for a crop yield

Where, Normalized $yield_i$, Normalized $yield_{max}$, and Normalized $yield_{max}$ are the normalized yield deviations for current period, maximum normalized yield and minimum normalized yield among whole data set, respectively. The normalized yield is calculated as:

$$\text{Normalized yield} = \frac{\text{Yield}_i - \text{Yield}_{trend}}{\text{Yield}_{trend}} \quad \ldots(2)$$

Where, $Yield_i$ is crop yield of current period; $Yield_{trend}$ is the trend predicted yield for each year. $Yield_{trend}$ has been calculated as:

$$\text{Yield}_{trend} = a + b * \text{Time} \quad \ldots(3)$$

Where, a is the intercept and b is slope of linear regression between yield and time.

The resultant scaled normalized yield does not show any increasing trend and mean value of that is nearly 50 percent with zero value of coefficient of determination (Fig. 1). It remove the impact of developed technology on crop yield and gives a better representation of weather variables effect on crop yield. If there is no time trend is present in crop yield data than scaled normalized yield will be equal to observed yield.

Weather indices calculation

After calculating scaled normalized yield we developed the Z variables. There was six weather variables used to find out z variables such as maximum and minimum temperature (T_{max} and T_{min}, °C), morning and evening relative humidity (RH_{max} and RH_{min},%), Rainfall (mm) and bright sunshine hour (SSH, hr) during 40th to 13th standard meteorological week. There are two types of Z variables such as simple Z variable and weighted Z variables. Simple Z variables were developed by summing the each weather variable or their interactions between 40th to 13th standard meteorological week for each years. Weighted Z variables are the sum product of each weather variable or their interactions and it is in correlation with crop yield. The simple and weighted Z variables were computed by following equations.

$$Z_{ij} = \Sigma_{w=1}^{m} X_{iw} \quad \ldots(4)$$

$$Z_{ij} = \Sigma_{w=1}^{m} X_{iw} X_{ii'w} \quad \ldots(5)$$

Where, Z_{ij} is the simple Z variable; X_{iw} and $X_{ii'w}$ is the value of i^{th} weather variable and their interaction with i′th variable for w standard meteorological week; m is the standard meteorological weeks used for model development.

$$Z_{ij'} = \Sigma_{w=1}^{m} r^{j}_{iw} X_{iw} \quad \ldots(6)$$

$$Z_{ij'} = \Sigma_{w=1}^{m} r^{j}_{ii'w} X_{iw} X_{ii'w} \quad \ldots(7)$$

Where, $Z_{ij'}$ is the weighted Z variable; r^{j}_{iw} and $r^{j}_{ii'w}$ is the correlation coefficient of yield with i^{th} and their interaction with i′th variable for w standard meteorological week. The details of simple and weighted Z variables are shown in Table 1.

Selection and extraction of variables

These Z variables are very closely correlated to each other. Sometimes the irrelevant variables developed a good agreement and increase complexity in model. So it is important to reduce the correlation to avoid the over fitting problem in a model development. Therefore variable selection and extraction was done to reduce the dimensionality of the data in crop yield forecast. Stepwise multiple linear regression (SMLR) model run in SPSS version

Table 1. Simple and weighted weather indices used for developing mustard prediction model by different techniques

	Simple weather indices						Weighted weather indices					
	T_{max}	T_{min}	Rain	RH_{max}	RHmin	SSH	T_{max}	T_{min}	Rain	RH_{max}	RH_{min}	SSH
T_{max}	Z10						Z11					
T_{min}	Z120	Z20					Z121	Z21				
Rain	Z130	Z230	Z30				Z131	Z231	Z31			
RH_{max}	Z140	Z240	Z340	Z40			Z141	Z241	Z341	Z41		
RH_{min}	Z150	Z250	Z350	Z450	Z50		Z151	Z251	Z351	Z451	Z51	
SSH	Z160	Z260	Z360	Z460	Z560	Z60	Z161	Z261	Z361	Z461	Z561	Z61

13.0 for selection of highest important variables. Principal component analysis version 13.0 run in SPSS to extract the variables by combining them into a new reduced set of variable. The principal components (PCs) was selected on the basis of eigen values (>1) were able to describe more than 90 percent variability of input data set.

Multivariate techniques

Three multivariate techniques (artificial neural network, support vector machine and random forest) was used to develop mustard yield prediction model using statistical software two third data for training purpose and one third data for testing purpose was used.

Artificial neural network (ANN)

Artificial neural network consists of many artificial neurons that are connected together to network architecture specifically. Neural network has various architectures to approximate any linear function such as: feed forward network, feedback network, lateral network etc. ANN composed of three layers namely, input layer, hidden layer and output layer. Multilayer perception (MLP) technique is one of the popular neural network types than other different neural network types. The neurons are arranged in a successive pattern, through which information will flow unidirectional from the input layer will pass to the output layer through the hidden layer. This network interpreted as a form input-output model, with weights and threshold (biases) as free parameters of the model. Artificial neural network work through the optimized weighted value of variables, the method by which the optimized values are attained is called learning. In the learning process it tries to teach to produce the output based on the corresponding input provided. Learning will complete when the trained neural network can able to update the optimal weights and produce the output within the desired accuracy corresponding to the input pattern. The main objective of the neural network is to produce its own output having reduced discrepancies with target output value, which will help to transform the input into meaningful output.

We use "caret" package for cross validation, "ggplot2" package for data visualization and "nnet" package in R statistical software version 3.1.3. There are 10 fold cross validation has been used for prediction by ANN method using R version 3.1.3 (Kuhn, 2008). Size and decay is the regularization parameter to avoid over-fitting in "nnet" package that represent the number of units in hidden layer and parameters of weight decay in nodes, respectively. A schematically representation of the ANN model for forecasting has been shown in Fig. 2.

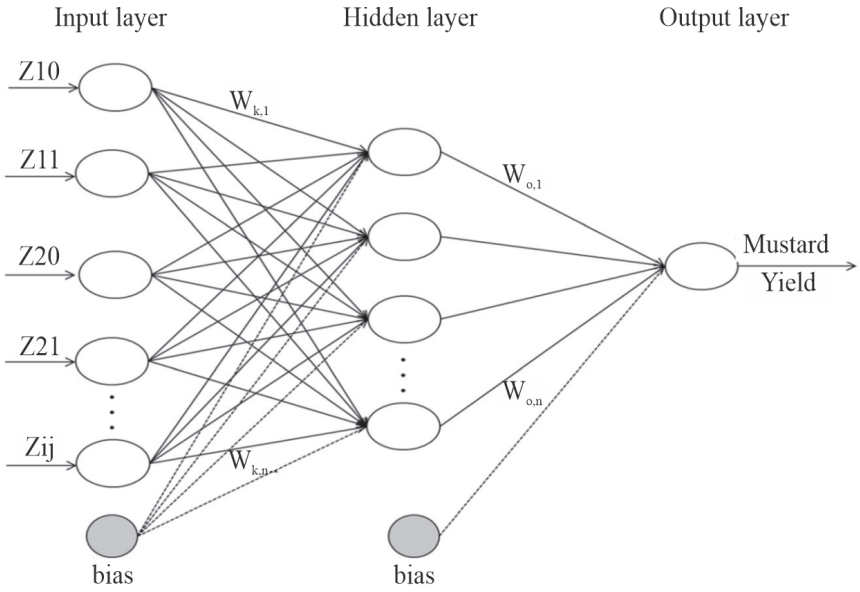

Fig. 2. Schematically representation of the basic ANN model

Support Vector Machine (SVM)

Support Vector Machines is a kernel-based, nonparametric, supervised machine learning technique used for prediction and classification of samples in two disjoint clusters (Pal, 2009). Nonparametric methods estimates the form of relationship between variable of interest and information seed yield directly based on the regression problem, as relationship between two is unknown. SVM is the useful tool with high accuracy for prediction and classification due to their capability to handle small training data sets. The schematically representation of SVM for crop yield forecasting has been shown in Fig. 3.

In this study, we used "e1071" package for SVM analysis, "caret" for cross validation and "ggplot2" for data visualization in R statistical software version 3.1.3. Gamma and cost are the parameters which is used for a cross validation. Gamma defines the distance of single data point from the hyperplane whereas value of cost decided the smoothness of hyperplane (large C smooth boundary). A low value gamma and large value of cost represents more accurate model for prediction.

Random forest (RF)

Random forest is an ensemble machine learning technique. It creates forest to enhance the performance of single decision tree by bootstrapping. At last it combines all the trees for better prediction. Each tree is different from another one due to presence of node and number of trees. The schematically representation of the random forest model has been shown in Fig. 4. Crop yield prediction is done "Random Forest" package. We used "caret" for cross validation and "ggplot2" for data visualization in R statistical software version 3.1.3 (Breiman, 2001). It is most important to decide the value of ntree and mtry for RF regression. The default value of ntree is 5. The value of mtry decides the number of variable randomly sampled at each split. The default value of mtry for classification is square root of variables and for regression is number of variable divided by three. More mtry showed a good agreement between trees.

Optimal combination of different models

The need of optimal combination to obtain diversified results because many forecasted model had similar accuracy so it is difficult to identify the best prediction model among them. The covariance of two forecasted results used in combination method to get the better property of forecast with least root mean square error. Optimal combination based on weights with the aim to minimize the expected loss of the combined forecast. The larger weight is responsible for better predictions and reduction in error.

We used variance of validated dataset to obtain optimal yield. The data sets were processed for analysis of variance and to develop an optimal combination model for all possible combination of ANN, SVM and RF by using MS-Excel office 2013. Following equation was used for optimal combination based on variance.

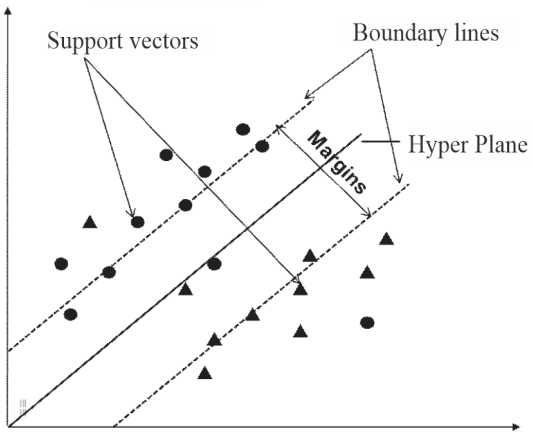

Fig. 3. Schematically representation of the basic SVM model

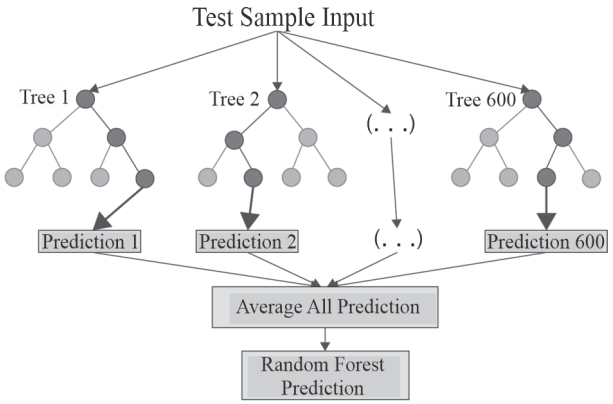

Fig. 4. Schematically representation of the basic Random forest model

$$X = \left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \frac{1}{\sigma_3^2}\right]^{-1} \left[\frac{X_1}{\sigma_1^2} + \frac{X_2}{\sigma_2^2} + \frac{X_3}{\sigma_3^2}\right] \quad \ldots(8)$$

Where, X_1, X_2 and X_3 are independent measurements (Testing data set) and σ_1^2, σ_2^2 and σ_3^2 are the variance of independent measurements (Testing data set) by ANN, SVM and RF, respectively.

Statistical test

Performance of model was done by calculating RMSE, nRMSE, RPD, MAE and percentage deviation.

Root mean square error (RMSE)

This is often used to measure the difference between predicted values from the model and actual observed values from the experiment that is being modeled. By this test, model performance during the calibration as well as validation period can be determined. It is also helpful in comparing individual model performance with that of other predictive models.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Pi - Oi)^2} \quad \ldots(9)$$

Where, RMSE is absolute root mean square error, Pi is the predicted value, O_i is the observed value and N is the number of observations

Normalized root mean square error (nRMSE)

If Pi, O_i, N and M are notated as predicted value, observed value, number of observations and mean of observed value, nRMSE can be written as the formula given below.

$$nRMSE = \frac{100}{M} * \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Pi - Oi)^2} \quad \ldots(10)$$

Normalized mean square error expressed in percentage, values close to zero indicates better model performance. nRMSE is a measure (%) of the relative difference of estimated versus observed data. The prediction is considered excellent with the nRMSE <10%, good if 10–20%, fair if 20–30%, poor if >30.

Ratio of performance to deviation (RPD)

RPD was also used to evaluate the prediction accuracy of the developed models. RPD<1.0 indicates very poor model and is not recommended for use; RPD between 1.0 and 1.4 indicates poor model; RPD between 1.4 and 1.8 indicates fair model which may be used for prediction; RPD values between 1.8 and 2.0 indicate good model which can be used for quantitative predictions; RPD between 2.0 and 2.5 indicates very good quantitative model for prediction, and RPD>2.5 indicates excellent model for prediction.

$$RPD = Sd/SEP \quad \ldots(11)$$

Where SEP = standard error of prediction, which is calculated as root mean squared error

$$SEP = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Oi - Pi)^2} \quad \ldots(12)$$

Sd = Standard deviation of the sample

$$Sd = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Oi - M)^2} \quad \ldots(13)$$

Percent Deviation

It is the difference between predicted and observed yield with reference to observed yield. The positive value of percent deviation shows overestimation and negative value shows under estimation of a model. Percent deviations calculated using following formula:

$$\text{Percent deviation} = \frac{Pi - Oi}{Oi} * 100 \quad \ldots(14)$$

Where, Pi is the predicted value and Oi is the observed value.

Mean Absolute Error (MAE)

Mean absolute error (MAE) of an estimator measures the average magnitude in predicted data set that is, the average difference between the estimated values and what is estimated. MAE calculated using following formula:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Pi - Oi| \quad \ldots(15)$$

Results and Discussion

Model was developed by artificial neural network, support vector machine and random forest techniques. The Z variable developed by weather parameters were taken as input parameter for developing crop yield prediction model using ANN, SVM and Random forest techniques. The z variable were selected by variable selection using stepwise regression (SR) technique and variable extraction using principal component analysis (PCA) technique.

Mustard yield prediction by variable selection using stepwise regression and ANN, SVM, RF techniques

Performance of the model developed for mustard yield prediction by variable selection using stepwise regression and ANN, SVM, RF techniques in R statistical software during calibration and validation are shown in Table 2. Results showed that model developed by ANN, SVM and Random Forest had RMSE value during calibration 151.2, 150.1 and 223.8 kg ha^{-1} respectively. During validation RMSE value was highest for model developed using SR-RF techniques (250.8 kg ha^{-1}) followed by SR-ANN (246.7 kg ha^{-1}) and SR-SVM (236.2 kg ha^{-1}) respectively. Mean absolute error (MAE) during calibration was 108.3, 91.3 and 183.7 kg ha^{-1} and during validation 179.9, 196.2 and 229.1 kg ha^{-1} respectively for model was developed by ANN, SVM and Radom forest. The value of nMAE calibration was 5.64, 4.75 and 9.55% and during validation was 9.02, 9.84 and 11.48% for ANN, SVM and RF respectively. nRMSE values calculated during calibration was 7.87, 7.81 and 11.64% and during validation was 12.37, 11.84 and 12.57% for ANN, SVM and RF respectively. The model predictions for study area were good with nRMSE value < 15% for all three developed model. Ratio of performance to deviation (RPD) values for the model developed by SR-ANN, SR-SVM and SR-RF techniques was 2.79, 2.81, 1.88 during calibration and 1.70, 1.78, 1.68 during validation, respectively. Results illustrate that among all three model developed, The SR-SVM model perform better followed by SR-ANN and SR-RF. Vashisth *et al*. (2018) reported that percentage deviation of estimated yield by actual yield of maize crop done at flowering stage and at grain filling stage was 10.3 and 7.1% by weather based statistical model. Ahmad (2017) noticed that ANN performed marginally better than RF to forecast the energy consumption. Palanivel and Surianarayanan (2019) reviewed several types of machine learning big data techniques such as linear regression, Artificial Neural Network and Support Vector Machine and found that SVM based prediction models are found to be more suitable for crop yield prediction.

Mustard yield prediction by variable extraction using principal component analysis and ANN, SVM, RF techniques

Performance of the model developed for mustard yield prediction by variable extraction using principal component analysis and ANN, SVM, RF techniques in R statistical software during calibration and validation are shown in Table 3. Result showed that the RMSE value during calibration was lowest for PC-SVM followed 39.8 kg ha^{-1} for PC-ANN (101.2 kg ha$^{-1)}$ and PC-RF (217.4 kg ha^{-1}) respectively. During validation PC-RF showed the highest RMSE values 271.5 kg ha^{-1} followed by 214.2 kg ha^{-1} for PC-ANN and 187.4 kg ha^{-1} for PC-SVM. Mean absolute error (MAE) during calibration was lowest

Table 2. Mustard yield prediction model developed by variable selection (SR) using SMLR and ANN, SVM and RF techniques

Accuracy parameters	SR-ANN		SR-SVM		SR-RF	
	Calibration	Validation	Calibration	Validation	Calibration	Validation
MAE (kg ha^{-1})	108.36	179.86	91.28	196.21	183.70	229.11
nMAE(%)	5.64	9.02	4.75	9.84	9.55	11.48
RMSE(kg ha^{-1})	151.25	246.71	150.14	236.21	223.84	250.83
nRMSE(%)	7.87	12.37	7.81	11.84	11.64	12.57
RPD	2.79	1.70	2.81	1.78	1.88	1.68

Table 3. Mustard yield prediction model developed using variable extraction (PC) by PCA and ANN, SVM and RF techniques

Accuracy parameters	PC-ANN		PC-SVM		PC-RF	
	Calibration	Validation	Calibration	Validation	Calibration	Validation
MAE(kg ha^{-1})	65.24	139.74	38.31	155.96	186.05	244.70
nMAE(%)	3.44	6.80	2.02	7.59	9.82	11.91
RMSE(kg ha^{-1})	101.18	214.24	39.83	187.43	217.41	271.52
nRMSE(%)	5.34	10.43	2.10	9.12	11.48	13.22
RPD	4.15	1.90	10.55	2.17	1.93	1.50

for PC-SVM (38.3 kg ha^{-1}) followed by PC-ANN (65.2 kg ha^{-1}) and PC-RF (186.0 kg ha^{-1}). The value of MAE for validation was lowest for PC-ANN (139.7 kg ha^{-1}) followed by PC-SVM (155.5 kg ha^{-1}) and PC-RF (244.7 kg ha^{-1}) for study area.

During calibration nMAE was 3.44, 2.02 and 9.82% and during validation was 6.80, 7.59 and 11.91% for ANN, SVM and RF respectively. During calibration values for nRMSE was 5.34, 2.10 and 11.48% and during validation was 10.43, 9.12 and 13.22% for ANN, SVM and RF respectively. The model predictions were excellent having value < 10% for model developed using PC-SVM and good with nRMSE value 10.43 and 13.22% for model developed using PC-ANN and PC-RF techniques. Ratio of performance to deviation (RPD) values from the model developed by PC-ANN, PC-SVM and PC-RF were 4.15, 10.55, 1.93 during calibration and 1.90, 2.17 and 1.50 during validation respectively. Among all the three models developed using principal component analysis extraction techniques by PC-SVM performed best followed by PC-ANN and PC-RF respectively.

By comparing model developed either using variable section by stepwise regression model or variable extraction by principal component analysis, SVM is performing best followed by ANN and RF. The feature extraction is useful for improving the performance of regression models, improving the stability against noise, avoiding over-fitting, reducing the training and testing time, and reducing the measurement and storage requirements. There were several researchers who used variable extraction technique for forecasting (Azfar *et al.,* 2015; Annu *et al*., 2017, Suzuki *et al*., 2020). Singh *et al.* (2014) reported that statistical models based on weather indices can successfully simulate multi-stage yield forecast of wheat at mid-season and at pre-harvest for Amritsar, Bhatinda and Ludhiana districts. This model is simple, does not require any sophisticated statistical tools, and can be used satisfactorily for district, agro-climatic zone and state level forecasting. Vashisth and Aravind (2020) reported that Elastic Net, LASSO and SMLR model based on weather parameters can be used for multistage mustard yield estimation and Elastic Net performed best among all the three models followed by LASSO and SMLR model.

Optimal combination of mustard yield prediction model

Optimal combination of models was used for predicting mustard yield. It chooses weights to minimize the expected errors of the combined forecast. There were four combinations ANN+SVM+RF, ANN+SVM, ANN+RF and SVM+RF used to combine the predicted results for the study area. Accuracy of prediction was improved by combining the different developed models.

The results obtained from optimal combination techniques used for model developed by variable extraction by PCA and ANN, SVM and RF techniques are presented in Fig. 5. Value of RMSE for mustard yield prediction by different optimal combination model was lowest for ANN+ SVM (179.9 kg ha^{-1}) followed by ANN+SVM+RF (203.1 kg ha^{-1}) and SVM+RF (216.6 kg ha^{-1}) and ANN+RF (230.72 kg ha^{-1}). nRMSE had similar pattern with lowest value for ANN+SVM (8.75%) followed by ANN+SVM+RF (9.88%), SVM+RF(10.55%) and ANN+RF (11.23%). Optimum combination techniques used for model developed by ANN+SVM

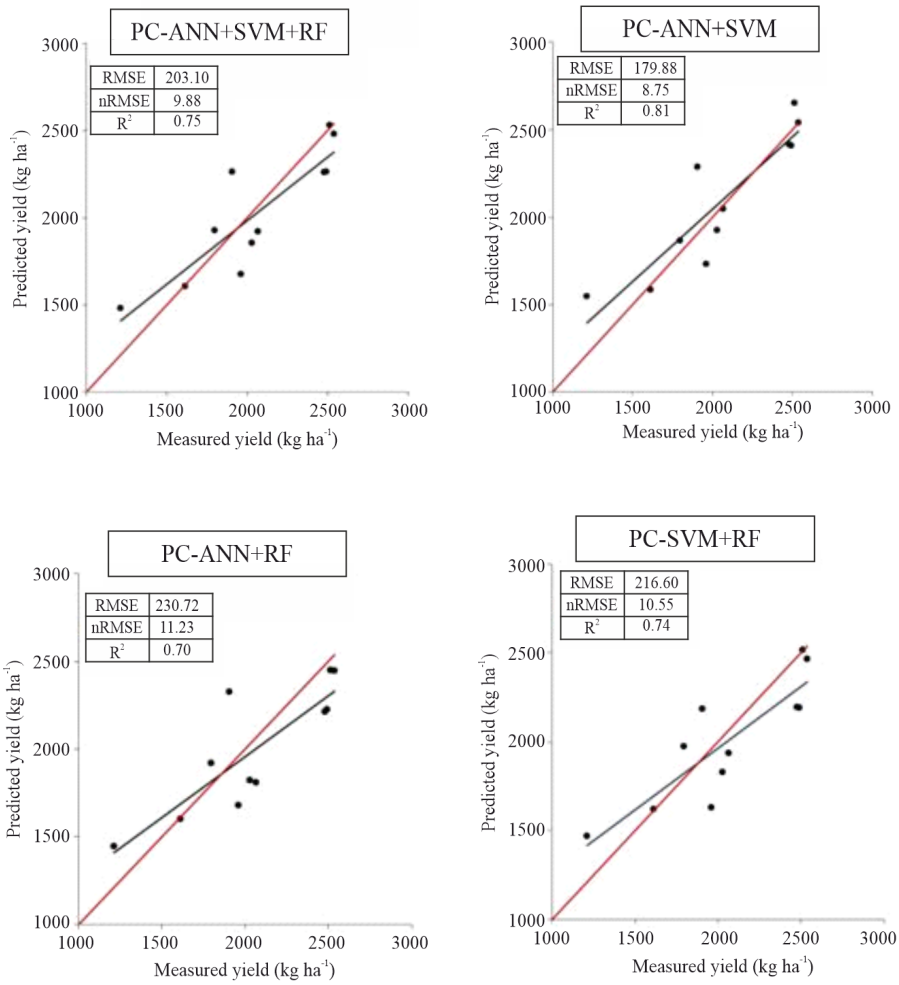

Fig. 5. Optimal Combination of model developed by variable extraction (PC) using PCA and ANN, SVM, RF techniques

and ANN+SVM+RF performed excellent with nRMSE value less than 10% followed by model developed by SVM+RF (10.55%). Model developed by ANN+RF performed good with nRMSE value less than 12%. The coefficient of determination was highest for ANN+SVM combination with R^2 value 0.81 followed by ANN+SVM+RF with R^2 value 0.75 and SVM+RF with R^2 value 0.74 and ANN+RF with R^2 value 0.70.

Optimal combination techniques used for model developed by variable selection by SMLR and ANN, SVM and RF techniques are presented in Fig. 6. Optimal combination techniques was used to predict the mustard yield by combination of SR-ANN, SR-SVM and SR-RF. Mustard yield prediction done by all four combination had MAE value lowest for SVM+RF (151.9 kg ha^{-1}) followed by 154.0 kg ha^{-1} for ANN+RF, 154.8 kg ha^{-1} for ANN+SVM and 159.86 kg ha^{-1} for ANN+SVM+RF. Value of RMSE and nRMSE value lowest for ANN+SVM+RF (204.4 kg ha^{-1} and 10.2%) followed by ANN+RF (207.7 kg ha^{-1} and 10.41%), ANN+SVM (219.0 kg ha^{-1} and 10.94%) and SVM+RF (219.4 kg ha^{-1} and 11.0%), respectively. Model developed by all four combination performed good with nRMSE value less than 11%. Coefficient of determinant was highest for ANN+RF with R^2 value 0.78 followed by ANN+SVM+RF with R^2 value 0.77, ANN+SVM with R^2 value 0.75 and SVM+RF with R^2 value 0.73.

Among all the optimum combination used for mustard yield prediction ANN+ SVM (PCA based) combination performed best followed by ANN +

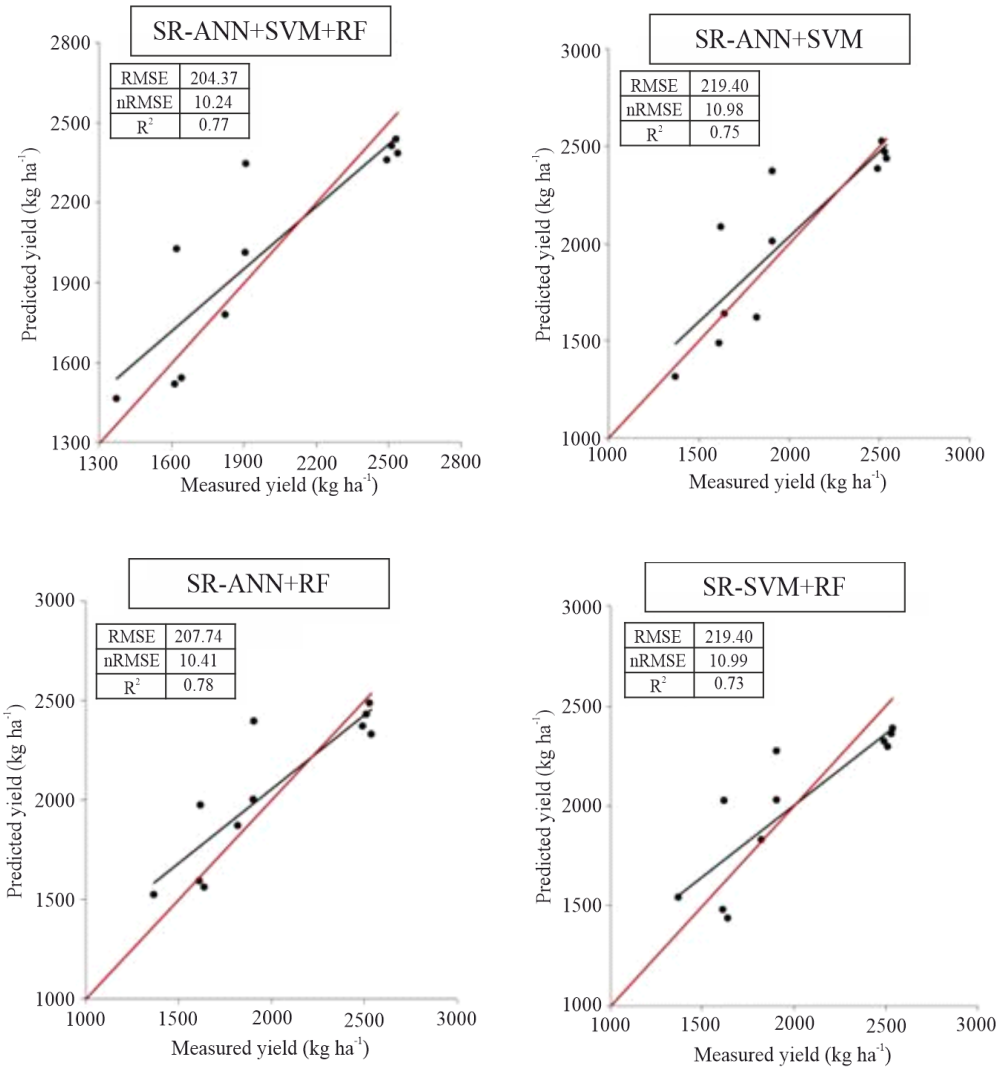

Fig. 6. Optimal Combination of model developed by variable selection (SR) using SMLR and ANN, SVM, RF techniques

SVM+RF (PCA based), ANN+SVM+RF (SMLR based), ANN + RF (SMLR based), SVM+ RF (PCA based), ANN+ SVM (SMLR based), SVM + RF (SMLR based), ANN+ RF (PCA based). The PCA based combination of ANN+SVM technique showed a good agreement between observed and predicted values for study area. Song *et al*. (2011) recommended that the optimal combination of different empirical crop yield forecasting model is more reliable approach to overdrawn the limitation of individual approach at regional scales. Hsiao and Wan (2014) concluded that combination of different forecast model is able to develop more reliable forecasting model to overdrawn the limitation of each model.

Conclusion

Model developed either by variable selection by SMLR or variable extraction by PCA and using ANN, SVM, RF techniques, SVM model performed best for mustard yield prediction done for the study area. Among all six models developed for mustard yield prediction, PCA-SVM performed best for the study area. Performance of the model developed by optimum combination performed better than the individual. By comparing the performance of the model developed by different techniques variable extraction by PCA performed better than variable selection by SMLR. Optimum combination of PC-(ANN+SVM) performed best followed by PC-SVM

and can be used for mustard yield forecast at district level.

References

Aggarwal, P.K., Bandyopadhyay, S.K., Pathak, H., Kalra, N., Chandra, S. and Kumar, S. 2000. Analysis of yield trend of the rice-wheat system in north-western India . *Outlook in Agriculture* **29**(4): 259-268.

Ahmad, M.W., Mourshed, M. and Rezgui, Y. 2017. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings* **147**: 77-89.

Annu, Sisodia, B.V.S. and Rai, V.N. 2017. An application of principal component analysis for pre- harvest forecast model for wheat crop based on biometrical characters. *International Research Journal of Agricultural Economics and Statistics* **8**: 83–87.

Azfar, M., Sisodia, B.V.S., Rai, V.N. and Devi, M. 2015. Pre-harvest forecast models for rapeseed & mustard yield using principal component. *Mausam*, **4**: 761–766.

Balakrishnan, N. and Muthukumarasamy, G. 2016. Crop production-ensemble machine learning model for prediction. *International Journal of Computer Science and Software Engineering* **5**(7): 148-153.

Breiman, L. 2001. Random forests. *Kluwer Academic Publishers. Manufactured in The Netherlands,* **45**: 5-32.

Everingham, Y., Sexton, J., Skocaj, D. and Bamber, G.I. 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development: 27-36.

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V. and Muller, J. 2013. Random Forests modelling for the estimation of mango (*Mangifera indica* L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agricultural and Water Management* **116**: 142-150.

Gandhi, N., Armstrong, L.J., Petkar, R.O. and Kumar A. 2016. Rice crop yield prediction in India using support vector machines. *International Joint Conference on Computer Science and Software Engineering*, *Khon Kaen, Thailand:* 1-5.

Gromping, U. 2009. Variable importance assessment in regression: Linear regression versus random forest. *American Statistician* **63**(4): 308-319.

Hsiao, C. and Wan, S.K. 2014. Is there an optimal forecast combination? *Journal of Econometrics* **178**: 294–309.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R. and Kim, S.H. 2016. Random forests for global and regional crop yield predictions. *Plos One* | DOI:10.1371/ journal.pone.0156571 2016: 1-15.

Karimi, Y., Prasher, S.O., Madani, A. and Kim, S. 2008. Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations. *Canadian Biosystems Engineering* **50**: 7.14-7.20.

Kuhn, M. 2008. Building predictive models in R using caret package. *Journal of Statistical Software* **28**: 1–6.

Pal, M. 2009. Kernel methods in remote sensing: A review. *ISH Journal of Hydraulic Engineering* **15**: 194–215.

Palanivel, K. and Surianarayanan, C. 2019. An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology* **10(**3): 110-118.

Pandey, P.C., Dadhwal, V.K., Sahai, B. and Kale, P. P. 1992. An optimal estimation technique for increasing the accuracy of crop forecasts by combining remotely sensed and conventional forecast results. *International Journal of Remote Sensing* **13**(14): 2735-2741.

Singh, A.K., Vashisth, Ananta, Sehgal,V.K, Goyal, A., Pathak, H. and Parihar, S.S. 2014. Development of Multi Stage District Level Wheat Yield Forecast Models. *Journal of Agricultural Physics* **14**(2): 189-193.

Song, H., Zhang, R., Zhang, Y., Xia, F. and Miao, Q. 2011. Energy consumption combination forecast of Hebei province based on the IOWA operator. *Energy Procedia* **5**: 2224-2229.

Su, Y.X., Xu, H. and Yan, L.J. 2017. Support vector machine-based open crop model (SBOCM): Case of rice production in China. *Saudi Journal of Biological Science* **24**(3): 537–547.

Suzuki, M., Shibahara, T. and Muragaki, Y. 2020. A method to extract feature variables contributed in nonlinear machine learning prediction. *Methods of Information in Medicine* **59**(1): 1-8.

Vashisth, Ananta and Aravind, K.S. 2020. Multistage mustard yield estimation based on weather variables using multiple linear, LASSO and Elastic Net Models for semi arid region of India. *Journal of Agricultural Physics* **20**(2): 213-223.

Vashisth, Ananta, Singh R. and Choudary, Manu. 2014. Crop yield forecast at different growth stage of wheat crop using statistical model under semi arid region. *Journal of Agro ecology and Natural Resource Management*: 1-3.

Vashisth, A., Goyal, A. and Roy, D. 2018. Pre harvest maize crop yield forecast at different growth stage using different model under semi arid region of India. *International Journal of Tropical Agriculture* **36**(4): 915-920.

Received: July 10, 2021; Accepted: October 19, 2021