https://doi.org/10.17221/15/2021-RAE

Models for feature selection and efficient crop yield prediction in the groundnut production

Kuruguntu Mohan Krithika, Nachimuthu Maheswari, Manickam Sivagami*

School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India

**Corresponding author: km.krithika2016@vitstudent.ac.in*

Citation: Krithika K.M., Maheswari N., Sivagami M. (2022): Models for feature selection and efficient crop yield prediction in the groundnut production. Res. Agr. Eng., 68: 131–141.

Abstract: Tamil Nadu ranks high in groundnut production in India. The yield prediction of the crop over Tamil Nadu will be highly useful in improving the efficiency of the production. This article aims to identify an efficient machine learning model to predict the groundnut crop yield and analyse the performance of the tested models. The study used the irrigation, rainfall, area and production data as factors for the groundnut crop yield across the districts of Tamil Nadu. This article identified the best set of features for training the models and studied various prediction models to evaluate the performance on the collected data. The trained and tested data were evaluated using various performance measures. The results of the study show that LASSO and ElasticNet provide the optimal results with the lowest RMSE and RRMSE values of 491.603 and 490.931 kg·ha^{-1}, 20.68 and 20.66%, respectively. The models showed the lowest MAE and RMAE values as well (333.154 and 331.827 kg·ha^{-1} and 14.53%, 14.51%, respectively) when compared to other models. The identification of the right time to sow and area to irrigate through feature selection and the prediction of the yield will improve the yield of the groundnut crops. This helps farmers to make practical decisions and reap the benefits.

Keywords: experimental models; groundnut yield; performance evaluation; prediction accuracy; subset selection

India stands third in oilseed production and first in groundnut production around the globe. Groundnut covers 40% of the area under oilseed production in India. In Tamil Nadu, the groundnut is cultivated in an area of around 0.619 million ha with a yearly production of 1.098 million tonnes. Among the production, 70% is from simply rain fed cultivation and the other 30% is from cultivated under irrigation according to Maya Gopal and Bhargavi (2019a). Accurate and efficient methods to predict the crop yield helps economists and officials in the planning process of agricultural practices according to Maya Gopal and Bhargavi (2019b). Kouadio et al. (2018) analysed that the availability of varied data has urged researchers to use data driven models to understand and produce accurate results. The prediction of the yield of any crop is not only dependent on environmental factors, such as the area, irrigation, rainfall, etc., but also the prediction algorithm, in order to expect precise results according to Sirsat et al. (2019).

However, when such models are to be applied on a large scale regions (like districts and states), the availability and collection of data for creating models is highly problematic according to Kouadio et al. (2018). Basso et al. (2013) analysed that the precision of results in the prediction of the yield for any crop can be achieved by providing appropriate inputs and selecting proper models without changing the traditional agricultural practices and their systems.

Various researchers have performed yield prediction analyses of different crops with data across India, such as paddy crops in Tamil Nadu using multiple linear regression (MLR) – artificial neural network (ANN) by Sirsat et al. (2019), pearl millet in Jaipur using an artificial neural network by Meena and Singh (2013), rice in Maharashtra using a support vector machine (SVM) model by Gandhi et al. (2016a) and other different algorithms. The rice crop yield prediction has been studied by using various factors like soil properties by Casanova

https://doi.org/10.17221/15/2021-RAE

et al. (1999), rainfall and fertilisers by Ramesh and Vardhan (2015), temperature and irrigation by Sirsat et al. (2019) and various other factors. ANN is one of the commonly available machine learning approaches used for yield prediction by many researchers like Gonzalez-Sanchez et al. (2014) and Haghverdi et al. (2018). The ANN model requires more time to develop than other linear regression methods because it is problem driven and there is no predefined method to develop the network according to Kaul et al. (2005). It has been exhaustively used by many researchers in the field. For example, Kaul et al. (2005) analysed the ability of ANN to predict the yields of corn and soybean in Maryland under typical conditions. Gandhi et al. (2016b) used ANN to predict the crop yield of rice in India using data from the districts of Maharashtra, while Emamgholizadeh et al. (2015) performed a predictive analysis of sesame seed yield predictions using ANN and multiple linear regression (MLR) algorithms. Maya Gopal and Bhargavi (2019a) introduced a hybrid of MLR and ANN to predict the crop yield using Tamil Nadu rice data which shows better accuracy than the individual models and this study implements the proposed algorithm to know if similar results can be obtained. The regression models least absolute shrinkage and selection operator (LASSO), ElasticNet (ENET) and support vector regression (SVR) have been used in prediction models, but not as quite often as the MLR and ANN models. Research by Jaikla et al. (2008), Sharif et al. (2017) and Das et al. (2018) are examples of the same, where the authors used LASSO regression, support vector machine, gradient boosting and random forest (RF) to predict Alzheimer's disease. RF proved to have the best accuracy. Mupangwa et al. (2020) used linear discriminant analysis (LDA) and logistic regression (LR) to predict the maize yield compared with non-linear methods. The performance of the LDA algorithm was the best among the models. Pallavi et al. (2021) discussed that data mining assists in the analysis of future patterns and characteristics that helps companies to make better decisions. For a particular region, they used the random forest approach to forecast the agricultural yield.

The comparison of different models and their performance on similar data has been common research area in the field of machine learning since they are data driven. Sharif et al. (2017) compared the regression models (stepwise, LASSO, ridge and ElasticNet) with varying α values and their performance was analysed to predict the winter oilseed rape yield. Wallach and Goffinet (1989) used the mean squared error as the major criterion to assess the quality of a model during the prediction of values. Safa and Samarasinghe (2011) used the metrics root mean squared error (RMSE) and the correlation coefficient to evaluate the performance of models. Maya Gopal and Bhargavi (2019b) evaluated models using RMSE, the mean absolute error (MAE), and the R value and compared them to evaluate their performances. Sirsat et al. (2019) used RMSE and the relative root mean squared error (RRMSE) to compare the models in predicting the grape vine yield for different phenological stages. The authors used the weather and soil data of Andhra Pradesh in the RF classifier model only. The authors Kumar and Sreenivasulu (2017) used remote sensing images of the Chitoor district, Andhra Pradesh, in the regression model to estimate the groundnut yield. The authors Shah and Shah (2018) included the soil, rainfall and weather parameters of Gujarat state in various models and achieved the best results using the K nearest neighbour model.

To the authors' knowledge, there is no study on groundnut crop yield prediction using data from the Tamil Nadu districts using a comparative analysis of the model performances. Here, models, such as LASSO, ElasticNet, SVR, SVR-ANN, MLR-ANN, along with the commonly used MLR and ANN are used to perform the prediction. The correlation matrix, the variance inflation factor, and backward elimination methods are used and the best sets of features are selected for training the models. Also, the results of these outputs, using the performance metrics RMSE, MAE, RRMSE, and R score are collectively compared for the performance evaluation of those models.

This study aims at understanding the performance of different models in the yield prediction on the collected groundnut crop dataset. The study uses the irrigation, rainfall, area and production data as factors for the groundnut crop yield across the districts of Tamil Nadu. This article identifies the best set of features to train the models and studies various prediction models. The prediction of the crop yield ahead of time helps the agricultural policy makers and involved farmers to take the required measures to store and acquire a market for the crop. It also helps the associated industries to plan their business and logistics as per Johnson et al. (2016), especially in Tamil Nadu where the groundnut is one of the major agricultural crops.

https://doi.org/10.17221/15/2021-RAE

MATERIAL AND METHODS

The data have been collected from online sources and official state departments. They have been pre-processed by removing missing values, NaN (not a number) values and filling them by the mean and median imputation methods in order to be trained by the models to predict the groundnut crop yield. The NaN values are insignificant and the imputation values did not vary the results of the prediction models. The models trained are compared for better performance and accuracy in this work.

Dataset and pre-processing

The study used variables for the groundnut crop yield across the districts of Tamil Nadu. This geographical area and its groundnut production have not been explored yet. The data were collected from the Tamil Nadu Statistical Department, Agricultural Department and web sources over the period of 2007–2017. A part of the data was manually collected from the Agricultural Department of Tamil Nadu to ensure authenticity.

The collected data have been aggregated year-wise and district-wise. It includes data from 30 districts of Tamil Nadu that belong to seven different agro-climatic regions as defined by the Agricultural Department of Tamil Nadu. The aggregated dataset contains the area of irrigation based on sources, the rainfall in millimetres for four different seasons, the area of the groundnut crop sown in hectares, the groundnut production in tonnes, and the yield of the crop in kg per hectare. As a result, the dataset documents 15 features of 244 instances. Figure 1 shows the study area classified based on the agro-climatic regions in Tamil Nadu.

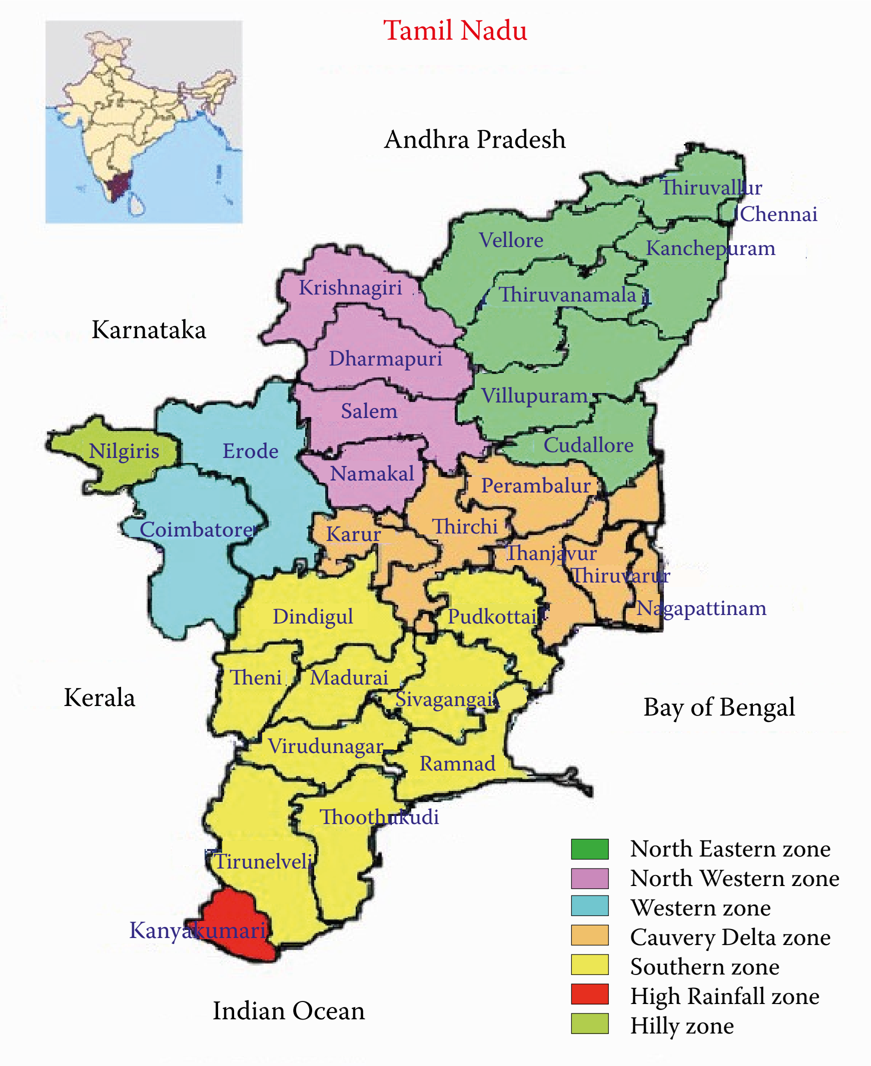

Figure 1. Area of study for the data classified based on the agro-climatic regions (Government of Tamil Nadu 2017)

https://doi.org/10.17221/15/2021-RAE

Table 1. Variance inflation factor of the variables in the dataset

Features	Variance inflation factor
Source canals (in ha)	1.5605
Source tanks (in ha)	6.4813
Source tube wells and other wells (in ha)	4.0970
Source open wells (in ha)	2.5640
Source other sources (in ha)	1.3154
Total area irrigated (in ha)	8.5974
Actual south-west monsoon (in mm)	1.0316
Actual north-east monsoon (in mm)	2.9454
Actual winter season (in mm)	9.8333
Actual hot weather season (in mm)	2.0220
Average rainfall (in mm)	9.2589
Area sown (in ha)	2.7176

Feature selection

Kouadio et al. (2018) analysed that the accuracy of a prediction model highly depends on the feature subsets used for training the model and, thus, the selection of relevant features plays an important role. The unique subsets of features are identified using the feature selection algorithm used for the crop yield prediction for better accuracy. Three different subset selection methods namely the correlation matrix, variance inflation factor and backward elimination method are used to identify the feature subsets that provide better accuracy. The set of features selected through the above methods are used to implement the MLR and RF models without cross-validation. Figure 2 explains the same.

Correlation matrix (CORR). The correlation matrix shows the Pearson's correlation coefficients between the groups of variables. Each variable chosen at random from the table is correlated with each of the other variables in the table. The correlation between the features is compared by using a heat map and the *P*-value generated in the correlation matrix.

For the given data, Pearson's coefficient values of the variables corresponding to the yield variable are compared. The variables whose corresponding values are greater than 0.1 and are less than 0.9 are selected as the features for further processing.

Thus, the features selected using the correlation matrix are the area irrigated by the canals, tube wells, rainfall in the north-east monsoon, average rainfall in the year, area of the sown crop, and groundnut production.

Variance inflation factor (VIF). The VIF method is a quick method to identify features as it eliminates independent features which are correlated using a one-time search over the predictor variables. It is an effective method to test the strength of each predictor that can be selected to create the model. The variance inflation factor (VIF) is computed as Equation (1):

$$VIF = \frac{1}{1-R^2} \tag{1}$$

where: R^2 – the coefficient of determination of the regression line.

A variance inflation factor based algorithm for variable selection chooses the independent features with lower collinearity by identifying the collinearity between all the independent features of the dataset. Table 1 shows the VIF factors of the variables in the dataset. According to the thumb rule, the features with a VIF factor >3 are eliminated as mentioned by Glen (2015). Thus, the selected variables based on the VIF factor are the areas irrigated by the canals, open wells, other sources, rainfall in the south-west monsoon, rainfall in the north-east monsoon, rainfall in the hot weather season and the area of the sown crop.

Backward elimination (BE). The backward elimination method of the feature selection begins with including all the variables in the dataset. The model fit criterion is identified and the variable which has been removed gives the most insignificant decrease in the model fit which is then deleted from the features. This process is repeated until no further variables can be removed without loss of the model fit.

The features selected from the data set using the backward elimination are the area irrigated by the canals, the area irrigated by the tanks, the open wells, the total area irrigated, the south-west monsoon, the north-east monsoon rainfall, the winter season rainfall and the average rainfall.

Experimental models

This study aims to develop a model using the subset of features that have been selected after proper analysis. A four-fold cross validation is used for this model development (similar to the one by Kouadio et al. (2018) which used (k –1) folds of data in training and building the model and the k^{th} fold to vali-

https://doi.org/10.17221/15/2021-RAE

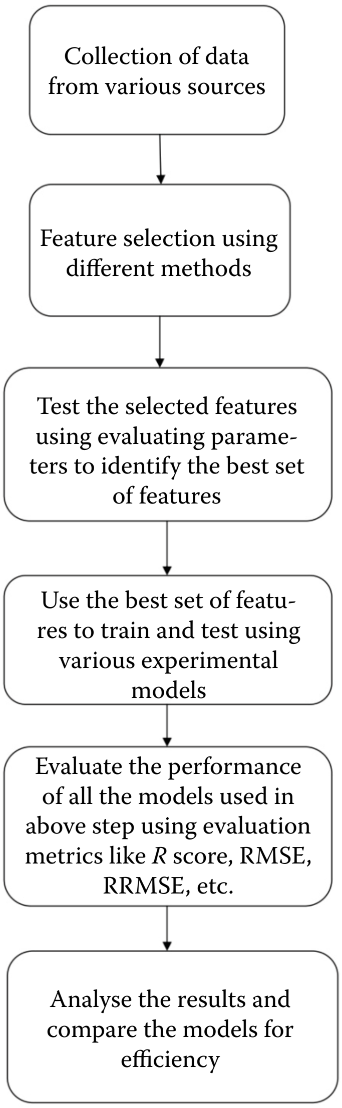
Collection of data from various sources
Feature selection using different methods
Test the selected features using evaluating parameters to identify the best set of features
Use the best set of features to train and test using various experimental models
Evaluate the performance of all the models used in above step using evaluation metrics like *R* score, RMSE, RRMSE, etc.
Analyse the results and compare the models for efficiency

Figure 2. Flowchart of the methodology used for this study

RMSE – root mean squared error; RRMSE – relative root mean squared error

date the model where every round of testing uses a different set of data chosen from those folds. The models are developed using python programming in a Jupyter notebook similar to the experiment by Basso et al. (2013), as mentioned in Figure 2.

Random forest (RF). Random forest is a supervised learning model in general. It is an ensemble learning model applicable to both classification and regression problems. The algorithm works by creating a number of decision trees during the training of data in order to give an output by simply calculating the mean of the prediction of those trees individually. In this algorithm, random decision trees are developed based on input data samples. An "out-of-bag" prediction is made where every decision tree is created from a bootstrap sample and one-third of the samples are left for validation and estimates are made from them. The random forest algorithm helps to rectify the problem of over-fitting in decision trees. In this work, a random forest algorithm with 1 000 trees has been implemented for the best results.

Multiple linear regression (MLR). The statistical model MLR is the most frequently used algorithm for forecasting crop yields by experts. Many researchers have used this model to predict the yield in various fields. MLR is a regression model where the dependent variable Y (production) is linearly related to multiple independent variables $x_1, x_2,... x_n$ where $x_1, x_2,... x_n$ are the selected input features. The MLR equation can be written as Equation (2):

$$Y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n \tag{2}$$

where: $b_0, b_1, b_2, ..., b_n$ – the parameters to be calculated; b_0 – the bias value and $b_1, b_2,...., b_n$ – the coefficients of the independent variables in the equation; these coefficients are calculated from the training dataset and using the regression equation.

In the current study, the area irrigated by the canals, the tube wells, the rainfall in the north-east monsoon, the average rainfall in the year, the area of the sown crop and the groundnut production are the independent predictor variables used in the MLR and the yield of the crop is the dependent variable to be calculated.

LASSO. The least absolute shrinkage and selection operator (LASSO) is a regression method used for analysis. It performs feature selection and regularisation to improve the prediction accuracy. It performs L1 regularisation where the absolute value of the coefficient's magnitude is added as a penalty. In this experiment, the mixing parameter is assigned as α = 0 for the LASSO penalty to perform the L1 regularisation. It is implemented using LASSOCV from the sklearn package in python.

ElasticNet (ENET). In machine learning and statistics, ElasticNet is viewed as a regularised regression model that linearly combines the penalties L1 of LASSO regression and L2 of ridge regression. It aims at minimising the loss function. In this work, the mixing parameters are set at α = 0 for the ridge model penalty and α = 1 for the LASSO model penalty.

Artificial neural network (ANN). The artificial neural network's operation is similar to the operation of a human brain. The neurons in the ANN model that are connected at each level are seen as the neurons in the brain that continuously transmit information. The model learns the pattern in a dataset by continuously training over the data for a number of times. It shows the ability of the model to identify the non-linearity between the predictor and the predicted variables.

A conventional ANN has an input layer, hidden layer and output layer. The hidden layer is placed between the input and output layers to perform non-linear transformations to the input values entered in the network. The number of neurons in the input layer is the same as the number of independent input features and one dependent variable is the single neuron in the output layer. The learning rates and epochs are identified by the trial-and-error method for the best results.

In this study, the input variables are the six independent features selected using the correlation matrix-based feature selection algorithm from the dataset and the output is the yield of the crop. Weights w_i are used to connect the nodes in the network between the layers and, additionally, the bias (b_i) is used for the hidden and output layers. This work includes only one hidden layer since the data is comparatively smaller. All three layers are fully connected. The best results are achieved by the 6-1-1 ANN structure with only three layers.

Hybrid multiple linear regression-ANN. In a traditional neural network model, the bias value and initial weights provided for the input layers are chosen at random. In a hybrid regression-neural network model, these initial weights and bias values fed to the input layers of the ANN are calculated from the MLR equation, see Equation (2) provided above. The coefficients of the independent variables in the MLR equation, Equation (2), are used to feed the input layer in the ANN model. The coefficients b_1, b_2,... b_n which are calculated using the training dataset are used as the initial weights w_1, w_2...w_n and b_0 is used as the bias value.

Support vector regression (SVR). Support vector regression is not anything like the other regression models. It predicts a continuous variable using a support vector machine (SVM) algorithm which is generally used to classify the data samples. All the other linear regression models are focused on minimising the error between the actual and predicted values, whereas support vector regression aims to identify the best fit line that lies within the threshold error value. All the prediction lines are classified either as ones that pass through the error boundary or that do not pass through the boundary where the boundary is the space defined by two parallel lines. The latter lines are not considered in the algorithm since the difference of the actual and predicted values is not within the limits of the threshold value, ε (epsilon). The lines that pass through the boundary are the ones considered as strong support vectors to make the required predictions. Figure 3 depicts this regression algorithm where the parallel lines are the boundaries considered at an ε deviation.

A kernel function is necessary in the SVR since it performs linear regression with high dimensional data and a kernel is the function that helps in mapping a lower dimensional data sample into a higher dimensional data. There are various kernel functions such as polynomial, Gaussian, sigmoid, etc. In this work, the SVR model uses a linear kernel function. It gave better accuracy when compared with the other kernels by the trial-and-error method.

Hybrid SVR-ANN. In a linear SVM, the hyper plane that separates the classes of data as accurately as possible is represented by the weights. The coordinates of a vector which are orthogonal to the hyper plane are provided by those weights. The hybrid SVR-ANN uses these weights as the initial parameters (weights and bias) to feed the first input layer of the ANN. This is similar to the hybrid MLR-ANN algorithm but uses the coordinate of the vectors as weights.

This algorithm was experimented with in this study from the inspiration of MLR-ANN which showed better performance compared to the individual MLR and ANN methods.

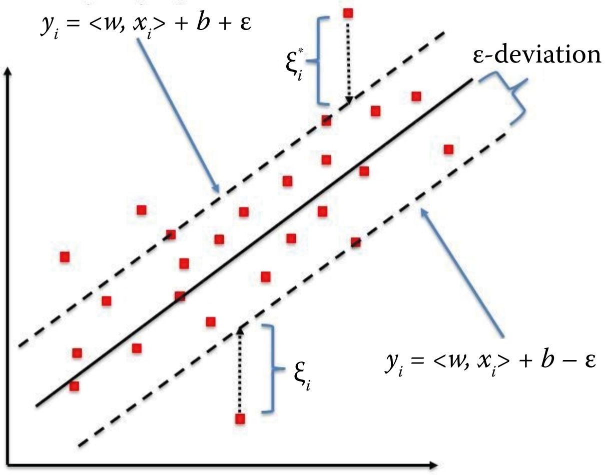
$y_i = <w, x_i> + b + \varepsilon$
ξ_i^*
ε-deviation
ξ_i
$y_i = <w, x_i> + b - \varepsilon$

Figure 3. Support vector regression

https://doi.org/10.17221/15/2021-RAE

Performance evaluation metrics

To compute the accuracy of the LASSO, ElasticNet, RF, MLR, ANN, SVR, SVR-ANN and MLR-ANN models, the actual observed yield data from the dataset and predicted crop yield output of the tested data are compared. The correlation coefficient (R), root mean squared error (RMSE), relative root mean squared error (RRMSE), mean absolute error (MAE) and relative mean absolute error (RMAE) whose formulae, Equations (3–7), respectively, are used:

$$R = \frac{\sum_{i=1}^{i=N}\left[\left(Y_i^a - Y_{mean}^a\right)\left(Y_i^p - Y_{mean}^p\right)\right]}{\sqrt{\sum_{i=1}^{i=N}\left(Y_i^a - Y_{mean}^a\right)^2}\sqrt{\sum_{i=1}^{i=N}\left(Y_i^p - Y_{mean}^p\right)}} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\left(Y_i^a - Y_i^p\right)^2} \tag{4}$$

$$\text{RRMSE} = 100 \times \frac{\text{RMSE}}{Y_{mean}^a} \tag{5}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{i=N}\left|Y_i^a - Y_i^p\right| \tag{6}$$

$$\text{RMAE} = 100 \times \frac{\text{MAE}}{Y_{mean}^a} \tag{7}$$

where: N – the number of data points in the testing phase; Y_i^a – the i^{th} actual data value; Y_{mean}^a – the mean actual data value; Y_i^p – refers to the i^{th} predicted value; and Y_{mean}^p – refers to the mean of predicted value.

Every performance metric individually focuses on the particular characteristic aspects of the error. In order to make conclusions about the performance of any model, such metrics need to be analysed in combination. Statistical metrics, such as RMSE, MAE and R score, are important in giving a logical and multidimensional comparison between various models.

RESULTS AND DISCUSSION

Feature selection. The features have been extracted using the correlation matrix, variance inflation factor and backward elimination methods. The accuracy of the models and respective subset of features are compared using the evaluation metrics root mean squared error (RMSE), mean absolute error (MAE) as mentioned in Figure 2. The results of RMSE and MAE are shown in Table 2 and Table 3, respectively.

Table 2. RMSE values of the models (kg·ha^{-1})

CORR	604.74	467.01
VIF	820.00	690.29
BE	724.70	860.46

RMSE – root mean squared error; CORR – correlation matrix; VIF – variance inflation factor; BE – backward elimination

The calculated RMSE and MAE values indicated that the features selected by the correlation method are more accurate for the yield prediction using both the MLR and random forest models (604.74 kg·ha^{-1} and 457.78 kg·ha^{-1} and 467.01 kg·ha^{-1} and 360.96 kg·ha^{-1}, respectively).

Thus, the features selected for the further analysis are extracted by the correlation matrix method. The features are the areas irrigated by the canals, tube wells, rainfall in the north-east monsoon, average rainfall in the year, groundnut production.

Experimental models. The current study aims to analyse the prediction behaviour of the models for the selected variables using python programming language. The results of the experiment performed for this study are provided in this section. Among the 14 crop features to predict the yield, the best six features are selected. Various feature selection methods like the correlation matrix, variance inflation factor and backward elimination are used. They are trained and tested on the MLR and RF models to identify the best set of features. A four-fold cross validation has been performed where three quarters of the data is used for training the model and one quarter of the data is tested by all the models that are compared in this work.

Table 3. MAE values of the models in kg·ha^{-1}

Performance measure	MLR	RF
CORR	457.78	360.96
VIF	645.94	523.64
BE	565.80	637.78

MAE – mean absolute error; MLR – multiple linear regression; RF – random forest;CORR – correlation matrix; VIF – variance inflation factor; BE – backward elimination

Table 4. *R* score, RMSE, MAE, RRMSE, and RMAE values for the prediction model

Performance measure	*R* score	RMSE (kg·ha^{-1})	MAE (kg·ha^{-1})	RRMSE (%)	RMAE (%)
RF	0.449	933.321	569.596	38.58	23.13
ELASTICNETCV	0.550	490.931	331.827	20.66	14.00
LASSOCV	0.549	491.603	333.154	20.68	14.02
MLR	0.452	923.449	554.071	38.14	22.44
ANN	0.521	645.852	587.740	26.42	24.04
MLR-ANN	0.523	647.907	480.089	26.51	19.64
SVR	0.234	1 079.209	684.767	44.18	27.66
SVR-ANN	0.499	777.771	660.720	31.82	27.03

RF – random forest; ELASTICNETCV – elements of statistical learning, a regression model; LASSOCV – least absolute shrinkage and selection operator; MLR – multiple linear regressions; ANN – artificial neural network; MLR-ANN – multiple linear regressions and artificial neural network; SVR – support vector regression; SVR-ANN – support vector regression and artificial neural network; RMSE – root mean score error; MAE – mean absolute error; RRMSE – relative root mean square error; RMAE – relative mean absolute error

Discussions. As a part of this work, the MLR, ANN and the hybrid MLR-ANN methods were compared. Similarly, the SVR, ANN and the hybrid SVR-ANN methods were compared to understand the performance of the models with other machine learning models.

Table 4 clearly shows that the MLR model has the highest RMSE (923.449 kg·ha^{-1}) whereas the ANN and hybrid MLR-ANN models show similar *R* score, RMSE and RRMSE values. On the other hand, the absolute error is comparatively less in the hybrid model (480.089 kg·ha^{-1}) whereas the MLR method has 554.071 kg per hectare. The ANN method has an absolute error of 587.740 kg·ha^{-1} being the highest of them all. It is clearly understood that the MLR method shows the lowest performance among them and the hybrid proves to be better than both since the absolute error and relative absolute error are the lowest for this model. Thus, defining the input weights of the ANN model using the coefficients of the MLR model and combining them gives better performance than the models when trained individually.

Clearly, the SVR model shows the highest RMSE and MAE values among all of the experimented models (1 079.209 and 684.767 kg·ha^{-1}, respectively). The statistical *R* score also indicates 23.4% accuracy only which is the least accurate in the set of chosen algorithms. The SVR-ANN method does not perform better than the conventional ANN method. The ANN method has the lowest values when compared to the SVR and the hybrid SVR-ANN models as shown in Table 4. This result is in contrast to the previous comparison of the hybrid MLR-ANN method comparison stating not all combinations of algorithms provide the expected results. The performance of every model is data driven and differs with the various datasets.

Figures 4–8 show the comparison of all the models. The authors Kumar and Sreenivasulu (2017) used the remote sensing images of the Chitoor district,

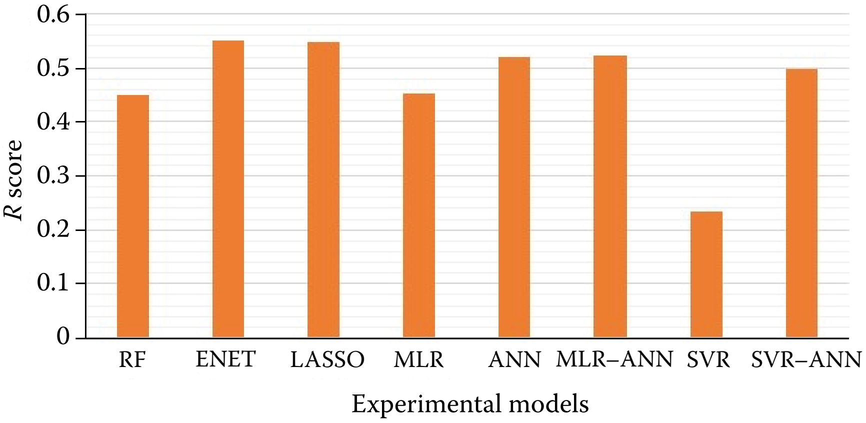

Figure 4. *R* score for the ML models

RF – random forest; ENET – ElasticNet; LASSO – least absolute shrinkage and selection operator; MLR – multiple linear regressions; ANN – artificial neural network; MLR-ANN – multiple linear regressions and artificial neural network; SVR – support vector regression; SVR-ANN – support vector regression and artificial neural network

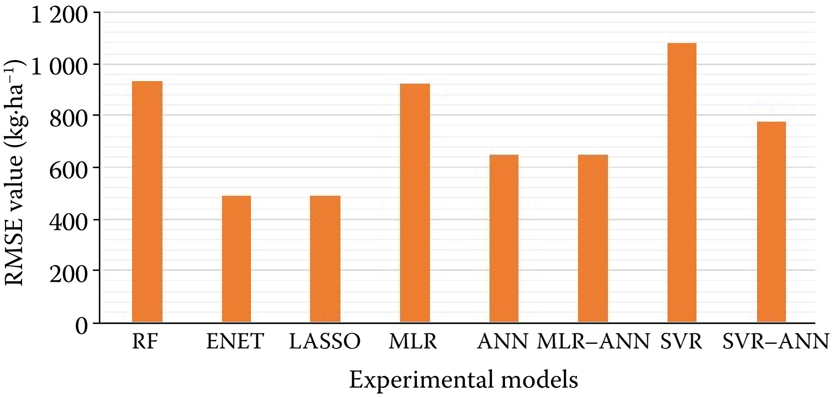

Figure 5. RMSE (kg·ha^{-1}) for the ML models

RMSE – root mean squared error; RF – random forest; ENET – ElasticNet; LASSO – least absolute shrinkage and selection operator; MLR – multiple linear regressions; ANN – artificial neural network; MLR-ANN – multiple linear regressions and artificial neural network; SVR – support vector regression; SVR-ANN – support vector regression and artificial neural network

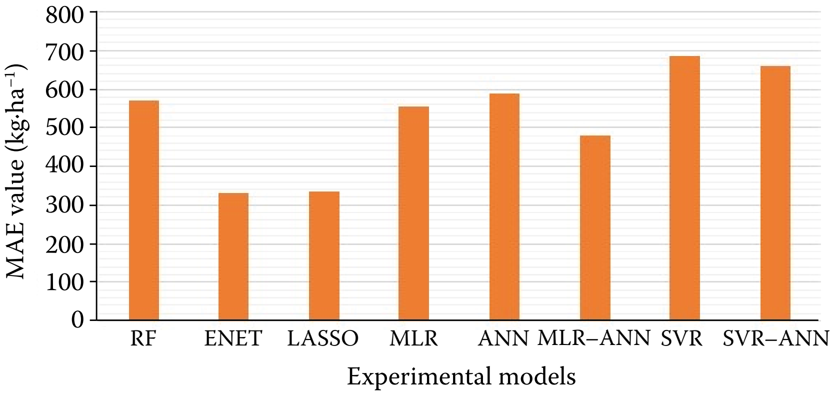

Figure 6. MAE (kg·ha^{-1}) for the ML models

MAE – mean absolute error; RF – random forest; ENET – Elastic Net; LASSO – least absolute shrinkage and selection operator; MLR – multiple linear regressions; ANN – artificial neural network; MLR-ANN – multiple linear regressions and artificial neural network; SVR – support vector regression; SVR-ANN – support vector regression and artificial neural network

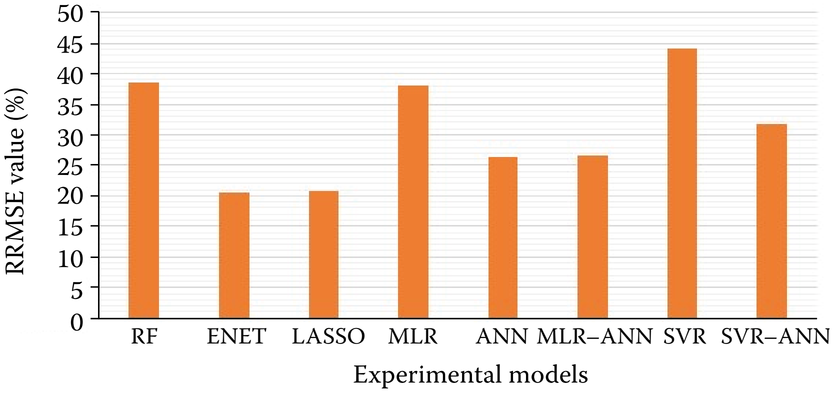

Figure 7. RRMSE (%) for the ML models

RRMSE – relative root mean squared error; RF – random forest; ENET – ElasticNet; LASSO – least absolute shrinkage and selection operator; MLR – multiple linear regressions; ANN – artificial neural network; MLR-ANN – multiple linear regressions and artificial neural network; SVR – support vector regression; SVR-ANN – support vector regression and artificial neural network

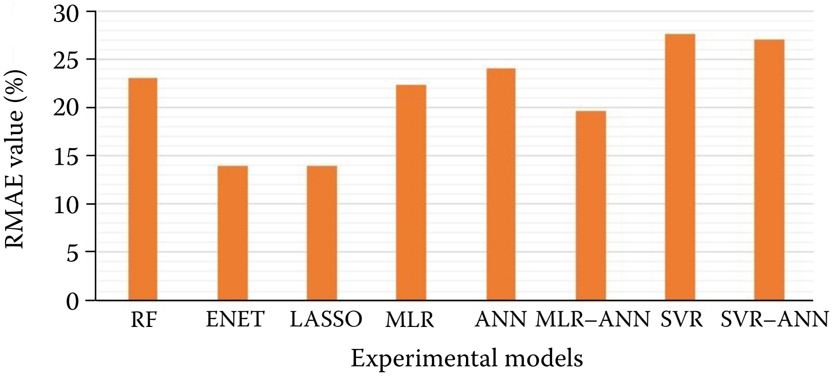

Figure 8. RMAE (%) for the ML models

RMAE – relative mean absolute errror; RF – random forest; ENET – ElasticNet; LASSO – least absolute shrinkage and selection operator; MLR – multiple linear regressions; ANN – artificial neural network; MLR-ANN – multiple linear regressions and artificial neural network; SVR – support vector regression; SVR-ANN – support vector regression and artificial neural network

Andhra Pradesh, in a regression model to estimate the groundnut yield. The authors used the weather and soil data of Andhra Pradesh in the random forest classifier model only. The authors Shah and Shah (2018) included the soil, rainfall and weather parameters of Gujarat state in various models and achieved the best result using the *K* nearest neighbour model. There is no study on the groundnut crop yield prediction using the data from the Tamil Nadu districts and a comparative analysis of the model performances. In our work, the RMSE values are the lowest for the LASSO and ElasticNet models which were 491.603 and 490.931 kg·ha^{-1}, respectively. Also, the RRMSE values are the lowest for those models (20.68 and 20.66%, respectively). The MAE and RMAE models show the lowest values for the LASSO and ElasticNet models as compared to the other models. Both these models gave similar accuracy rates which explains that the L2 regularisation has the least impact on the regression line, but the L1 regularisation ensures higher performance. The precision is achieved because of the in-built cross validation in these models. It is expected that the in-built feature selection method of these models helps in further identifying the best features to predict the output and, hence, giving the maximum prediction accuracy.

CONCLUSION

Standard performance metrics are used to calculate the accuracy of a prediction. The results of the conventional models MLR, RF, SVR, LASSO, ElasticNet, ANN, MLR-ANN and the hybrid SVR-ANN are compared. The results of the LASSO and ElasticNet models show better prediction accuracy and provided optimal results with the lowest RMSE and RRMSE values of 491.603 and 490.931 kg·ha^{-1}, 20.68 and 20.66%, respectively. The models showed the lowest MAE and RMAE values as well (333.154 and 331.827 kg·ha^{-1}, i.e. 14.53 and 14.51%, respectively) when compared to the other models using the groundnut crop dataset to predict the yield. Tamil Nadu being one of the major producers of groundnuts, the feature selection identified the factors influencing the crop yield which can be prioritised in decision making. The scope of this paper is data specific and applies to Tamil Nadu and its districts only because the seasons, irrigation conditions and other factors differ in other parts of the country which can alter the data set and its patterns. The future scope would be to include various data points from various parts of the country and world and identify patterns in the groundnut production and the most efficient models for the yield prediction. Furthermore, this efficient yield prediction model will help in the understanding and planning of farming practices.

Acknowledgement: We thank the Department of Agriculture of state of Tamil Nadu and the online source, open government data portal of Tamil Nadu (data.tn.gov) for providing with agricultural data for research purposes.

REFERENCES

Basso B., Cammarano D., Carfagna E. (2013): Review of crop yield forecasting methods and early warning systems. In: Proceedings of the First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics (FAO Headquarters), July 18–19, 2013, Rome, Italy: 18–19.

Casanova D., Goudriaan J., Bouma J., Epema G.F. (1999): Yield gap analysis in relation to soil properties in direct-seeded flooded rice. Geoderma, 91: 191–216.

Das B., Nair B., Reddy V.K., Venkatesh P. (2018): Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. International Journal of Biometeorology, 62: 1809–1822.

Emamgholizadeh S., Parsaeian M., Baradaran M. (2015): Seed yield prediction of sesame using artificial neural network. European Journal of Agronomy, 68: 89–96.

Gandhi N., Armstrong L.J., Petkar O., Tripathy A.K. (2016a): Rice crop yield prediction in India using support vector machines. In: 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), July 13–15, 2016, Khon Kaen, Thailand: 1–5.

Gandhi N., Petkar O., Armstrong L.J. (2016b): Rice crop yield prediction using artificial neural networks. In: IEEE International Conference on Technological Innovations in ICT for Agriculture and Rural Development (TIAR), July 15–16, 2016, Chennai, India: 105–110.

Glen S. (2015): Variance inflation factor. Available at https://www.statisticshowto.com/variance-inflation-factor/ (accessed Feb 1, 2021).

Gonzalez-Sanchez A., Frausto-Solis J., Ojeda-Bustamante W. (2014): Attribute selection impact on linear and nonlinear regression models for crop yield prediction. The Scientific World Journal, 2014: 1–10.

Government of Tamil Nadu (2017): Report No.1 of 2017 – Economic Sector Government of Tamil Nadu [Dataset]. Available

https://doi.org/10.17221/15/2021-RAE

at https://cag.gov.in/webroot/uploads/download_audit_report/2017/Report_No.1_of_2017_-_Economic_Sector_Government_of_Tamil_Nadu.pdf (accessed Feb 4, 2021).

Haghverdi A., Washington-Allen R.A., Leib B.G. (2018): Prediction of cotton lint yield from phenology of crop indices using artificial neural networks. Computers and Electronics in Agriculture, 152: 186–197.

Jaikla R., Auephanwiriyakul S., Jintrawet A. (2008): A rice yield prediction using a support vector regression method. In: 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, May 14–17, 2008, Krabi, Thailand: 29–32.

Johnson M.D., Hsieh W.W., Cannon A.J., Davidson A., Bédard F. (2016): Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agricultural and Forest Meteorology, 218: 74–84.

Kaul M., Hill R.L., Walthall C. (2005): Artificial neural networks for corn and soybean yield prediction. Agricultural Systems, 85: 1–18.

Kouadio L., Deo R.C., Byrareddy V., Adamowski J.F., Mushtaq S. (2018): Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. Computers and Electronics in Agriculture, 155: 324–338.

Kumar K.S., Sreenivasulu G. (2017): Locally received NOAA based crop yield estimation using vegetation index and atmospheric parameters for Chittoor district. International Journal of Applied Engineering Research, 12: 9688–9696.

Maya Gopal P.S., Bhargavi R. (2019a): A novel approach for efficient crop yield prediction. Computers and Electronics in Agriculture, 165: 1–9.

Maya Gopal P.S., Bhargavi R. (2019b): Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms. Applied Artificial Intelligence, 33: 621–642.

Meena M., Singh P.K. (2013): Crop yield forecasting using neural networks. In: International Conference on Swarm, Evolutionary, and Memetic Computing, Dec 19–21, 2013, Chennai, India: 319–331.

Mupangwa W., Chipindu L., Nyagumbo I., Mkuhlani S., Sisito G. (2020): Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in Eastern and Southern Africa. SN Applied Science, 2: 1–14.

Pallavi K., Pallavi P., Shrilatha S., Sushma, Sowmya S. (2021): Crop yield forecasting using data mining. Global Transitions Proceedings of International Conference on Computing Systems and Applications, 2: 402–407.

Ramesh D., Vardhan B.V. (2015): Analysis of crop yield prediction using data mining techniques. International Journal of Research in Engineering and Technology, 4: 470–473.

Safa M., Samarasinghe S. (2011): Determination and modelling of energy consumption in wheat production using neural networks: A case study in Canterbury province, New Zealand. Energy, 36: 5140–5147.

Shah V., Shah P. (2018): Groundnut crop yield prediction using machine learning techniques. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3: 1093–1097.

Sharif B., Makowski D., Plauborg F., Olesen J.E. (2017): Comparison of regression techniques to predict response of oilseed rape yield to variation in climatic conditions in Denmark. European Journal of Agronomy, 82: 11–20.

Sirsat M.S., Mendes-Moreira J., Ferreira C., Cunha M. (2019): Machine learning predictive model of grapevine yield based on agro climatic patterns. Engineering in Agriculture, Environment and Food, 12: 443–450.

Wallach D., Goffinet B. (1989): Mean squared error of prediction as a criterion for evaluating and comparing system models. Ecological Modelling, 44: 299–306.

Received: February 5, 2021
Accepted: January 30, 2022
Published online: September 14, 2022