Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2022.019947
Article

Tech Science Press

Enrichment of Crop Yield Prophecy Using Machine Learning Algorithms

R. Kingsy Grace*, K. Induja and M. Lincy

Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore, 641653, India
*Corresponding Author: R. Kingsy Grace. Email: kingsygrace.r@srec.ac.in
Received: 01 May 2021; Accepted: 02 June 2021

Abstract: Strong associations exist between the crop productivity and the seasonal, biological, economical causes in natural ecosystems. The linkages like climatic conditions, health of a soil, growth of crop, irrigation, fertilizers, temperature, rainwater, pesticides desired to be preserved in comprehensively managed crop lands which impacts the crop potency. Crop yield prognosis plays a vibrant part in agricultural planning, administration and environs sustainability. Advancements in the field of Machine Learning have perceived novel expectations to improve the prediction performance in Agriculture. Highly gratifying prediction of crop yield helps the majority of agronomists for their rapid decision-making in the choice of crop to be cultivated. This paper makes an attempt to suggest which crop can be sown at a particular district in Tamil Nadu, depending on the factors required for the growth of the crop based on the research outcomes. This is achieved by applying clustering on the attributes in the dataset using E-DBSCAN and is compared with three different algorithms such as DBSCAN, CLARA and K-means. The best suitable factor for the growth of a crop for a location is predicted using clustering techniques. The accuracy of crop yield prediction is calculated for three crops, namely, rice, wheat and maize. The proposed method outperforms in terms of Bias, F1 Score, MAPE, MDAE, MSE, RAE, RMSE and MAE with existing algorithms in the literature.

Keywords: Machine learning; DBSCAN; K-means; Multiple linear regression

1 Introduction

Precision farming is an innovative idea to improve crop yield which also reduces the contamination and crop yielding expenses. The concept of precision farming is based on the factors such as ecological and soil parameters which is either temporal or spatial. Precision farming will augment the crop yield and is financially viable to the farmers [1,2].

Agronomics crop production rely on various aspects. It is dependent on factors such as health of the soil, climatic conditions, nature of cultivation, irrigation type, needed fertilizers, atmospheric conditions, rainfall, harvesting, pesticides and other factors [3]. There is no proper yield because of unpredictable climatic changes and reduction in water resource. The majority of farmers are not receiving the estimated crop yield in India. Earlier days yield prediction was performed by considering farmer's previous experience

on a particular crop. Crop yield forecast is a significant farming problem. But in the recent times, the traditional way of predicting proves to be unfavorable, because of varying climatic and biological constraints. Agricultural planning is essential to estimate the crop yield. The factors involved to make decisions for farming are, namely, the price of commodity, type of soil and environmental conditions. The large amounts of data that are nowadays virtually harvested along with the crops have to be analyzed so that they are manipulated to give the maximum production. The concept of data analytics is adopted in agriculture because the agricultural data is huge (crops, season, soil content, temperature range, soil Ph, minerals etc.). During data analytics, huge amount of data is efficiently gathered, store and analyze for decision making. Data analytics is generally grouped as predictive and descriptive type. Predictive type is essentially used in farming [4].

In the most characterizing time, processing moved from huge centralized computers to cloud and supercomputers. The algorithms which are used to unravel computer are improved consequently through experience and by the utilization of data. It is viewed as a piece of artificial intelligence and is called as Machine learning (ML) [5]. ML is a technique for information examination that computerizes insightful model structure. The leeway of ML system is to learn from precedent knowledge and formulate pronouncement with negligible human intercession. The three classes of ML algorithms are specifically, supervised learning, unsupervised learning and semi-supervised learning. The technique for recognizing comparable congregation of information in a dataset is called clustering. It is the most typically utilized strategy of unsupervised learning. Clustering is employed to discover data clusters with the end goal that each group has the most firmly harmonized data. Individual data points in each cluster are relatively more like rudiments of that cluster than those of different clusters. Given a bunch of data points, the clustering technique is focused to arrange every data point into a particular cluster.

Machine learning based data analytics is an automated prediction and analytical process which is involved in the transformation of data into useful information by applying two separate processes of knowledge discovery and prediction. Here knowledge is discovered from the results obtained from the clustering process. The prediction of crop yield is made by applying predictive modeling techniques on clustering [6–8]. The application of clustering techniques to crop research data enables the customizing of information on sowing crops depending on various attributes like districts, crop type, and season. The proposed method focuses on analyzing agricultural data and finding the factors which maximize the crop production using unsupervised machine learning algorithms, specifically, clustering algorithms. Also, the best factors for the growth of a crop yield and helping the farmers to choose the high yielding crop with respect to the location to get the maximum yield. The factors contained in the dataset are clustered using different clustering algorithms. The clustered results are then fed to the Multiple Linear Regression algorithm for prediction. The error functions such as Mean absolute error (MAE), Relative absolute error (RAE), Mean square error (MSE) and Root mean square error (RMSE) are calculated. Finally, the accuracy of the system is verified using quality metrics such as purity, F-measure, recall, precision and rand index. The remaining part of the paper is structured as follows: Section 2 describes related work of various models used for crop yield forecasting. Section 3 presents the proposed methodology. Section 4 discusses the results and finally, concluding remarks and future work is discussed in Section 5.

2 Literature Survey

The crop productivity is one of the major problems in agronomics and is deliberated in several researches. This section confers the crop productivity techniques discussed in the research. Majumdar et al. have proposed optimal parameters which are used to increase the crop yield [4]. The input dataset [9–13] consists of six years data with parameters such as crop (cotton, groundnut, jowar, rice and wheat), season (kharif, rabi, summer), area, production, average temperature in centigrade, average rainfall (mm),

pH value, soil type, nitrogen (kg/Ha), phosphorus, potassium. The proposed modified Density based spatial clustering of applications with noise (DBSCAN) algorithm is used for clustering the agricultural data of different districts having similar type of temperature, rain and soil [4]. DBSCAN is compared with Partition around medoids (PAM) and Clustering for large applications (CLARA) and these algorithms are applied to the agricultural data of the districts producing highest crop yield. Result shows that DBSCAN provides the better clustering quality than PAM and CLARA. Good clustering is exhibited in CLARA than PAM.

Ahamed et al. [14] have discussed the prediction of annual yield of crops in different districts of Bangladesh. The dataset is collected from Bangladesh agricultural research institute (BARI). Environmental, biotic and area central variables are considered as the input variables. K-means classification and Linear regression are used for prediction. From the result, three best crops for major agricultural districts of Bangladesh have been suggested.

Ramesh & Vishnu have presented user friendly interface for farmers by giving analysis on rice production [8]. The East Godavari district from 1955 to 2009 data is used for prediction. Multiple linear regression and Density based clustering techniques are used for prediction of rice yield. The prediction results for Multiple linear regression vary from −14% to +13% and Density based clustering technique from −13% to +8%. Gholap et al. have proposed a methodology for classification of soil and to predict untested attributes using regression technique [15]. The dataset is collected from the district of Pune, India. Classification is done by Naive Bayes, J48, JRip. The prediction algorithms Linear regression, Least medium square regression are implemented using Weka tool. The results show that the Least medium square regression is more accurate but Linear regression consumes less time comparatively.

Sujatha & Isakki have introduced crop yield estimation approaches based on environmental season, biological and economic reasons [16]. Data mining techniques are used to find the excellent crop which gains the farming area. Hemageetha has presented a survey made on forecasting of crop yield using classification techniques [17]. It describes six suitable classification algorithms for analyzing soil to use for agriculture. The data mining algorithms used are association rule mining, classification techniques include Naïve Bayes, J48 and K-means for clustering. The classification algorithms are used based on their fertility.

Aishwarya has discussed the prediction of rice productivity in Bangladesh [18]. Input variables are environmental, biotic and area of production. The dataset was collected from Warangal for 20 years. Each year data contains 7 attributes, i.e., rainfall, maximum temperature, minimum temperature, hours of sunshine, wind speed, humidity and cloud coverage. Each year's worth of rice yield contains 3 rice varieties namely Boro, Aman and Aus. Clustering techniques are applied to divide regions; and then suitable classification techniques is applied to obtain crop yield predictions. ANN provides better prediction for some of the crops such as Aus having more missing values. Linear regression provides better prediction performance for Boro and Amon.

Diepeveen & Armstrong have described data mining techniques for crop performance variability. Data is collected from the Department of agriculture and food western Australia [19]. The key attributes in the data includes nutrition and soil type, grain yield and quality, sowing and harvest dates and tolerance to environmental stresses. Multivariate mixed model using R and asreml-r are used to produce the predictions for all the dimensions of the data. This predicted new dataset is analyzed with principal components analyses. The predictions from these mixed models were then put into a data-cube implemented in Postgresql. The data-cube was then used for reporting and querying variety of predictions. As a result of this research, it would be suggested that the growers could use data mining techniques to identify high performing varieties for their specific locations and farming practices through the adoption of predicted varieties.

Ramesh & Ramar have discussed to classify soil using data mining techniques [20]. The proposed methodology was developed by commercial and research center and the dataset collected from Kanchipuram district. The soil of Kanchipuram district is organized into 8 classes. Data was collected from 7 common soil types to classify the soil. Soil classification deals with systematic categorization of soil based on their characterization. K-nearest neighbor and support vector machines are used to classify the data. Weka is used to interpret the soil. Result states that small number of traits determines their effectiveness with standard statistical techniques. They also suggest that data mining technique can be used in the field of soil.

Van Evert et al. [21] have analyzed big data techniques for weed data management. Traditionally statistical methods have been used in agricultural applications. Some of the traditional methods are Principal component analysis (PCA), regression and Analysis of variance (ANOVA). These methods are not capable of handling large number of variables in Big Data applications. To overcome this, machine learning models have to be used to process agricultural data. In addition to that appropriate training methods have to be chosen for real time implementation of large volume of agricultural data analysis [21]. Bose et al. [22] have proposed an image processing technique named Spiking neural networks (SNNs) to perform remote sensing spatiotemporal analysis. This process is based on image time series and is used to estimate crop yield from agricultural images. When compared to traditional methods, SNN performs better in terms of crop yield estimation.

Maya Gopal et al. [23] have predicted crop yield with respect to environmental, weather conditions and biological features. MLR is used to predict the accuracy of feature selection algorithm. The metrics such as RMSE, MAE, R, and RRMSE are calculated for the proposed algorithm. 85% accuracy is achieved in the proposed feature selection algorithm. Maya Gopal et al. [24] have proposed important feature selection for prediction of crop yield. After feature selection, the MLR is used for prediction. SFFS algorithm is used for selecting 5 features and the MLR prediction provides 85% of accuracy. Mohammed et al. [25] have proposed different irrigation methods for the production of crop.

Multi-model ensemble (MME) method is used by Iizumi et al. [26] for crop yield prediction. The crops which are studied are (*i*) maize, (*ii*) rice, (*iv*) wheat and (*v*) soybean. Niedbała [27] have introduced a Multilayer perceptron (MLP) model to predict the rapeseed yield. RAE, RMS, MAE, MAPE are the metrics used to find the errors in the proposed system. The minimum of the error obtained in the proposed system is 9.43%. Ricciardi et al. [28] have presented a survey on crop production with the help of open-access dataset. This dataset consists of 154 crops and 11 farm size classes. The dataset is developed as csv file and contains the overall crop production of 51.1% in globally.

MARS-crop prediction system is proposed for European Union by Velde et al. [29]. The proposed MCYFS method uses MAPE as an accuracy indicator. Lecerf et al. [30] have proposed a crop model using meteorological indicators for their forecasting. Using the proposed model maize grain prediction is found before 80 days of harvest and the soft wheat prediction done before one month of harvest. Meshesha et al. [31] have proposed a crop prediction for Ethiopia. Four crops such as wheat, rice, maize and teff were used for the prediction. The EVI and NDVI indicators are used in the proposed system to measure the prediction. Geoffrey et al. have proposed ensemble based for crop yield measures which is related with dynamical climate. The probabilistic prediction is done with the help of RPSS and ROCSS. Usually, the prediction happened before sowing and is approximately before 2-months [32]. Lambert et al. [33] have proposed a crop prediction using Sentinel-2 time-series. A review on machine learning based crop yield prediction is presented in [5]. Tab. 1 provides the comparison of crop prediction methods in the literature.

Table 1: Comparison on different crop yield prediction methods

S. no.	Reference	Proposed algorithm	Data set	Crops	Performance
1.	Data Mining Techniques [4]	DBSCAN	6-year data of Karnataka	cotton, groundnut, jowar, rice and wheat	Better than PAM and CLARA
2.	Data Mining Techniques [14]	K-means and classification and regression using k-NN, Neural Network and Linear Regression	Dataset collected by Bangladesh Agricultural Research Institute (BARI)	Rice	Three best crops for major agricultural districts have suggested
3.	Data Mining Techniques [8]	Multiple Linear Regression and Density Based Clustering Technique	East Godavari district, Andhra Pradesh from 1955 to 2009	region specific crop	Multiple Linear Regression varies from −14% to +13% and Density Based Clustering Technique from −13% to +8%
4.	Classification Techniques and Prediction [15]	Naïve Bayes, JRip, J48 & regression technique	Soil samples taken from 3 regions of Pune district such as Khed, Bhor, and Velhe (1988 soil samples)	Not Mentioned	Accuracy of Naïve Bayes =38.40%, JRip= 90.24%, J48= 91.90%
5.	Rice Prediction [18]	ANN & Linear regression	Dataset from Warangal for 20 years	3 rice varieties namely Boro, Aman and Aus	ANN provides better prediction for some of the crops such as Aus having more missing values. Linear regression provide better prediction performance for Boro and Amon.
6.	Data mining techniques for crop performance variability [19]	principal component analyses	Department of Agriculture and Food Western Australia	Most of the crops	Identify high performing varieties for their specific locations and farming practices through the adoption of predicted varieties
7.	Image Time Series [22]	Spiking Neural Networks (SNNs)	Data from: China (Hebei, Henan, Shandong, Anhui, and Jiangsu)	maize	Average accuracy = 95.64%

(Continued)

Table 1 (continued)

S. no.	Reference	Proposed algorithm	Data set	Crops	Performance
8.	Optimizing crop yield prediction [23]	Multiple Linear Regression (MLR)	secondary sources of state Agriculture Department, Government of Tamil Nadu, India, for over 30 years	Most of the crops	Accuracy = 85 %
9.	Global crop yield forecasting [26]	multi-model ensemble	JRA-25 reanalysis (1984–2010)	maize, rice, wheat and soybean	predictions are reliable in 23–32%
10.	Crop prediction [33]	supervised Random Forest	2014–2017 STARS project	Cotton, maize, millet, while peanut and sorghum	80% overall accuracy

3 Methodology

Cluster algorithms are used to cluster data objects based on their characteristics, and aggregation of data objects based on their similarities [16]. Clustering algorithms are classified as unsupervised learning. Clustering methods can handle problems with noise and outliers efficiently and requires only minimum amount of domain knowledge to determine input parameters. The steps to perform clustering techniques and their performance analysis for crop yield prediction are discussed in this section. Fig. 1 shows the system architecture for crop yield prediction. Different parts of the system architecture are: (i) Dataset, (ii) Pre-Processing, (iii) Clustering, (iv) Cluster Validation, (v) Prediction and (vi) Visualization.

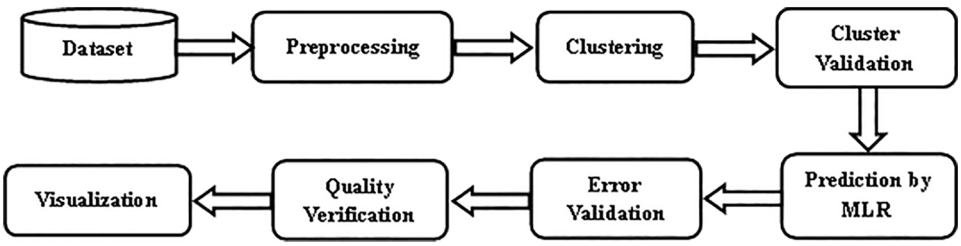

Figure 1: System architecture for crop yield prediction

The dataset used in this work is taken from Open Government Data (OGD) [34,35]. The dataset consists of 1,00,000 records containing 22 attributes such as districts of Tamil Nadu, pH level, temperature, sunlight, minerals like phosphorous, potassium, boron, carbon, nitrogen, sulphur, calcium, magnesium, manganese, zinc, iodine, copper. The crops considered are rice, wheat, maize for which the prediction process is performed. Preprocessing helps to correct the incomplete data, remove noisy data and all duplicate records. It also helps to achieve accuracy, completeness, consistency, timeliness, and interpretability.

Various algorithms are available to perform clustering. In this proposed work, K-means, CLARA, DBSCAN and E-DBSCAN are used to analyze the crop yield prediction. Depending on the working of each algorithm, the attributes are clustered based on similarities. Elements belonging to one cluster are similar and the elements from various clusters differ in their features. The data points can be scattered into any shapes, hence the clustering algorithms clusters them in both arbitrary shapes and specific shapes like square, circle, oval etc. The clustered results are predicted followed by error validation and quality verification. Cluster validation process takes the factors such as, cluster size, number of clusters formed, minimum cluster size, noise, average distance of the cluster points, and median distance of the clusters to check the quality of the clusters formed.

Prediction technique in data mining discovers the relationship between variables which is either dependent or independent. In this work, prediction is used to find the productivity of a crop in a particular location and the important parameters required. The predicted results are visualized using different charts. The steps of how the clustering process is performed are shown in Fig. 2. The features are selected from the collected data set. Then clustering algorithm is implemented. Finally, the results are validated.

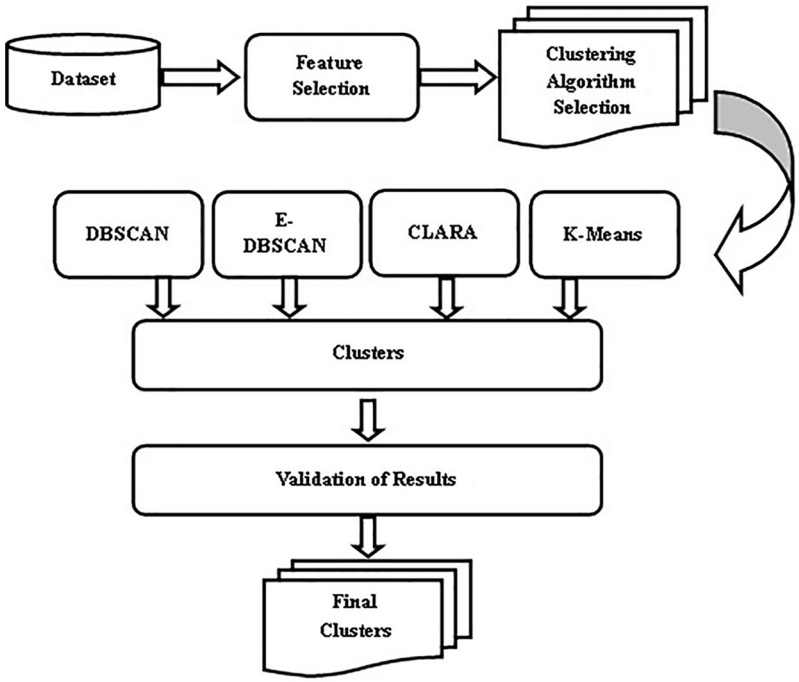

Figure 2: Steps in clustering process

3.1 Density-Based Spatial Clustering and Application with Noise (DBSCAN)

Density based clustering algorithms create clusters according to the high density of members of a data set, in a determined location. It aggregates some distance notion to a density standard level to group members in clusters [5]. DBSCAN is density-based clustering algorithm. It identifies clusters which are in different shape and also works well in the noisy and outlier's dataset. The number of clusters generated using DBSCAN is automatic. The steps followed in DBSCAN algorithm is shown in Algorithm 1.

1. For each data point, compute the distance between the data point and the other points. Finds all neighbor points within distance *eps* (*eps* is reachability maximum distance) of the starting point. Each point, with a neighbor count greater than or equal to *MinPts*, (*MinPts*: reachability minimum number of points) is marked as core point or visited.
2. For each core point, if it's not already assigned to a cluster, create a new cluster. Find recursively all its density connected points and assign them to the same cluster as the core point.
3. Iterate through the remaining unvisited points in the dataset .
4. The K -Nearest Neighbour method (KNN) supports DBSCAN algorithm in finding the optimal *eps* value.
5. The value of k will be specified by the user and corresponds to *MinPts*.
6. Next, these k -distances are plotted in an ascending order. The aim is to determine the "*knee*", which corresponds to the optimal *eps* parameter.
7. A *knee* corresponds to a threshold where a sharp change occurs along the k -distance curve.

Algorithm 1: DBSCAN Algorithm

3.2 Clustering Large Applications (CLARA)

CLARA is a partitioning based clustering algorithm which is used to handle large datasets [36]. Each partition, the objects are closer and in different partition the objects are far away. K-Medoids is the predecessor of CLARA. Algorithm 2 shows the steps of CLARA algorithm.

1. Split randomly the data sets in multiple subsets with fixed size
2. Compute the algorithm on each subset and choose the corresponding k representative objects (medoids). Assign each observation of the entire data set to the closest medoid.
3. Calculate the mean (or the sum) of the dissimilarities of the observations to their closest medoid. This is used as a measure of the goodness of the clustering.
4. Retain the sub -dataset for which the mean (or sum) is minimal. A further analysis is carried out on the final partition.
5. Note that, each sub -data set is forced to contain the medoids obtained from the best sub -data set until then. Randomly drawn observations are added to this set until the subsets with fixed size has been reached.

Algorithm 2: CLARA Algorithm

3.3 K-Means

K-means clustering clusters the dataset based on predefined number of clusters. K centroids are calculated and the data objects are clustered in each centroid [37]. The objective function is shown in Eq. (1).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \tag{1}$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster centres. The algorithm steps for K-means clustering are shown in Algorithm 3.

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Algorithm 3: K-Means clustering

3.4 E – DBSCAN

Finding neighborhood in DBSCAN is time expensive because it uses spatial methods. So, this method is not suitable for high dimension data [38]. The improved version of the DBSCAN is E – DBSCAN (Enrichment of DBSCAN) used in the proposed system to get higher prediction results in crop yield prediction. The steps for E – DBSCAN are shown in Algorithm 4.

1. For each data point, compute the distance between the data point and the other points. Find all neighbor points within distance *eps* of the starting point. Each point, with a neighbor count greater than or equal to *MinPts,* is marked as core point or visited.
2. Based on eps one lists is created for neighborhood calculation, datapoint D is ordered non-decreasingly with respect to distance. Two thresholds have calculated (Threshold_1 = p.dist – Eps & Threshold_2 = Eps + p.dist)
3. q.dist < Threshold_1 and Distance(q, p) < = eps, add the datapoint to the cluster.
4. q.dist > Threshold_2 and Distance(q, p) < = eps, add the datapoint to the cluster.

Algorithm 4: E – DBSCAN clustering

3.5 Multiple Linear Regression

Here, the algorithm used for prediction is Multiple Linear Regression. Multiple linear regression is an extension of the simple linear regression where multiple independent variables exist [39]. It is used to analyze the effect of more than one independent variable $x_1, x_2, \ldots . x_k$ on the dependent variable y. For a given dataset, $(y, x_1, x_2, \ldots . x_k)$ the multiple linear regression fits the dataset to the model in Eq. (2).

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i \tag{2}$$

where β_0 is the y-intercept and the parameters $\beta_0, \beta_1, \ldots, \beta_k$ are called partial coefficients.

4 Experimental Results

The proposed work is executed using the R studio and the results are deliberated in this segment. The datasets utilized in this research were sourced from the 24 districts of Tamil Nadu out of which five districts were preferred arbitrarily. The Fig. 3 is sketched amidst the districts and the production values. The crops deliberated are rice, wheat, maize for the selected districts, namely, Ariyalur, Coimbatore, Cuddalore, Dharmapuri and Dindugal. The district of Ariyalur shows the maximal production among all other districts, in which the top-most crop yield is for maize followed by rice and wheat. The remaining districts Dharmapuri, Dindugal, Coimbatore, Cuddalore are rated on the basis of crop production. Based on the yield in every district, Ariyalur has the highest yield for rice and wheat. Ariyalur and Dharmapuri shows the maximal productivity for maize.

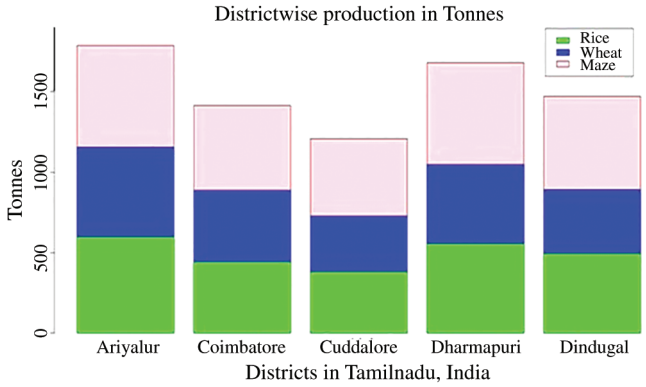

Figure 3: Crop yield prediction based on location (Districts in Tamilnadu, India)

The correlation between crops and attributes is depicted in the Tab. 2. Correlation determines the accordance between one dependent variable with other independent variables. The correlation is calculated for three crops, namely, rice, wheat and maize with their attributes. The parameters investigated and the correlation among the designated crops (Wheat, Rice and Maize) are Ph (pouvoirhydrogene), I (iodine), K (potassium), C (carbon), N (nitrogen), Zn (zinc), Temperature, Ca (calcium), Cu (copper), Mg (magnesium), Mn (Manganese). When the correlation value of an attribute earshot to 100% indicates that the attribute is holding better correlation with the crop. For all the crops, pH correlation is 90%. Likewise, the correlation value for iodine, potassium, carbon, nitrogen is above 80%.

Table 2: Correlation between crops and attributes

Wheat	Rice	Maize
Ph = 0.8957	Temp = 0.9039	Ph = 0.9014
I = 0.8867	Ph = 0.9019	Temp = 0.8987
K = 0.8832	Mn = 0.8995	P = 0.8983
C = 0.8825	N = 0.8967	Mg = 0.8962
N = 0.8823	K = 0.8912	K = 0.8909
Zn = 0.8753	Mg = 0.8898	Ca = 0.8912
Mg = 0.8752	C = 0.8873	Cu = 0.8903

Figs. 4–6 depicts each crop with its best factors, to be considered based on the correlation among the crops and the factors. The factors deliberated are Ph level, temperature, sunlight, minerals like phosphorous, potassium, boron, carbon, nitrogen, sulphur, calcium, magnesium, manganese, zinc, iodine, copper. For wheat, the best three factors are ranked as Ph, zinc, boron. Similarly, for rice, the best three factors are ranked as temperature, Ph, and magnesium. The best three factors for maize are ranked as the best parameter is Ph, temperature, boron.

Tab. 3 shows the evaluation on clustering algorithms based on clustering parameters and their values. The clustering algorithms accomplished are E-DBSCAN, DBSCAN, CLARA and K-Means. The cluster parameters considered are cluster size, minimum cluster size, noise, diameter of each cluster, average distance of each cluster, median distance between the clusters formed, separation value between the clusters, average value between the cluster formed, average distance within the data points of a cluster,

dunn (Dunn index value gives minimum separation/maximum diameter value) and s-index value where C1, C2, C3 denotes clusters which are formed.

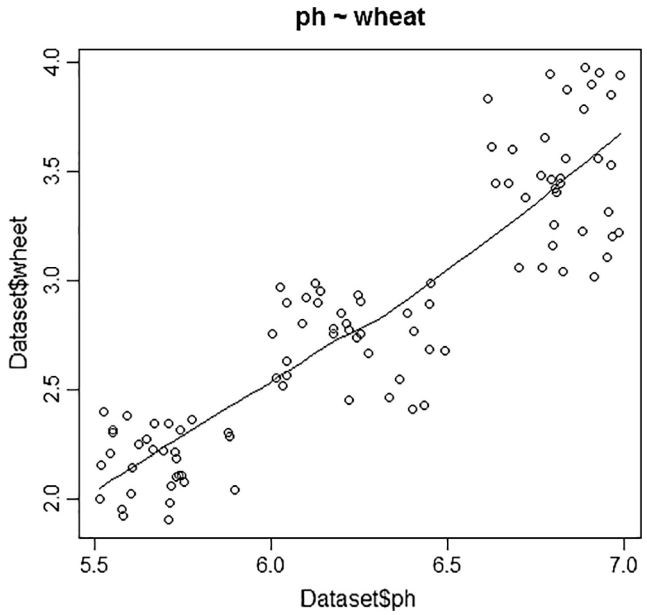

Figure 4: Plot for wheat vs pH

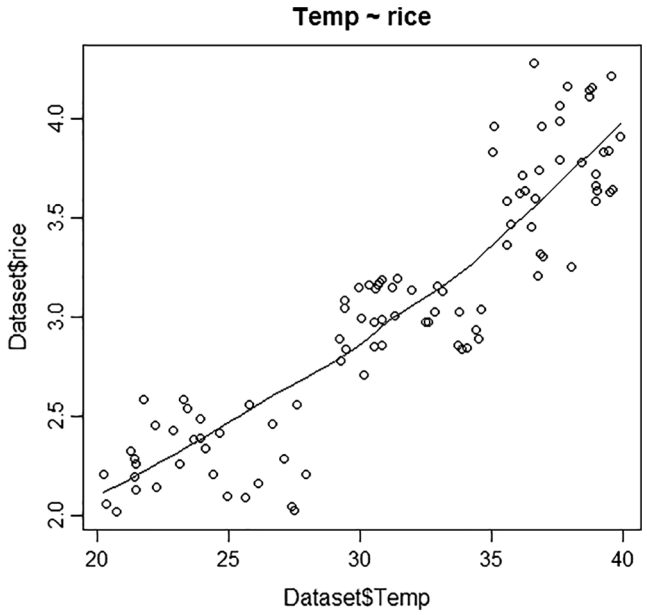

Figure 5: Plot for rice vs temperature

Fig. 7 shows the detailed values of Quality validation [40,41]. The cluster quality is diagnosed using the quality metrics like Purity, F-measure, Recall, Precision, and Rand Index. Purity is high to get good cluster performance. The other quality metrics value is high in E-DBSCAN for getting better cluster. The F-measure, Recall and Precision value for E-DBSCAN, DBSCAN, CLARA and K-Means is 40%. Similarly, purity is

78% for E-DBSCAN, purity is 70% for DBSCAN, 60% for CLARA, 50% for k-means. The Rand index value is 85% for E-DBSCAN, 80% for DBSCAN, is 70% for CLARA and 60% for k-means. The clustering quality of E-DBSCAN is better than DBSCAN, CLARA and K-means. Tab. 4 is having the quality metrics formulas for clustering algorithms.

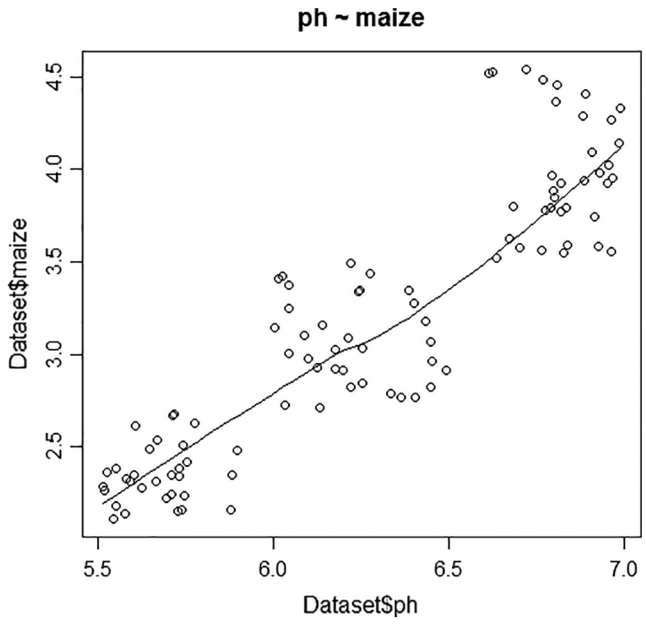

Figure 6: Plot for Maize vs pH

Table 3: Comparison on clustering algorithms

Validation factors	DBSCAN			CLARA			K-Means			E- DBSCAN		
Number of Cluster	3			3			3			3		
Min cluster size	299			318			317			270		
Noise	0			0			0			0		
Cluster size	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
	312	389	299	318	322	360	323	360	317	302	394	305
Diameter	190.1	230.4	143.9	181.7	202.3	243.9	209.1	243.5	155.7	170.1	240.4	113.9
Average distance	68.2	120.8	77.3	73.4	89.4	99.7	89.6	99.7	73.1	58.2	130.8	66.3
Median distance	64.7	123.9	79.5	72.6	88.1	88.9	88.3	88.9	72.4	60.7	143.9	69.5
Separation	11.2	11.3	19.8	19.1	19.1	55.6	53.7	55.6	53.7	9.2	9.3	15.8
Average between	342.67			314.05			314.06			350.4		
Average within	62.3			88.6			88.3			60.1		
Dunn	0.043			0.078			0.22			0.035		
S index	56.16			65.54			76.72			50.13		

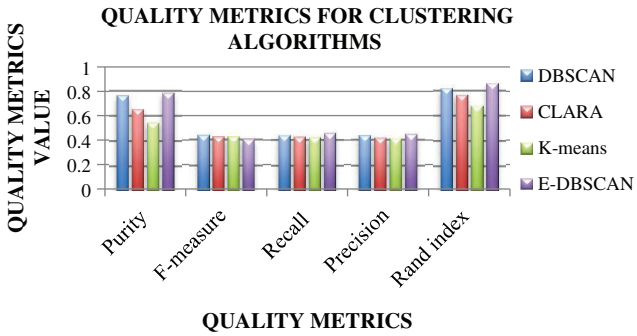

Figure 7: Quality metrics for clustering algorithms

Table 4: Quality metrics for clustering algorithms

Quality metrics formula for clustering algorithms	Formula
Purity	$Purity = \frac{1}{N}\sum_{q=1}^{k} \max_{1<j<l} n_q^j$ N – *Total no. of samples* q – *Cluster* n_q^j - *No. of samples in cluster*
F-Measure	$F - Measure = 2 * precision * recall / (precision + recall)$
Recall	$Recall = TruePositives / (TruePositives + FalseNegatives)$
Precision	$Precision = TruePositives / (TruePositives + FalsePositives)$
Rand Index	$Rand\ Index = \frac{Index - Expected\ Index}{Max\ Index - Expected\ Index}$

Machine learning is used to find the rapport amongst input and output using learning [21]. Training data is called as input data. If the training data is (Xi,Yi) where i = 1,. . .,n and (X_i,Y_i) exemplifies the prior season yield. Machine learning is used to learn the function f which is Y = f (X) and it fits the training data. The learning is said to be good if the average mean square error should be minimum when compared to other types of errors [20]. The attributes considered for crop yield prediction are pH, temperature, I, K, Cu, Ca, Mg and Zn. The proposed method uses multiple linear regression (MLR) to forecast the crop yield for 15 years (2000 to 2014). The predicted results for Wheat, Maize and Rice are shown in Figs. 8–10. Tab. 5 presents the details of actual crop yield and predicted crop yield with error. The accuracy of the system lies within the range of 90 to 95 percent. The error percentage is calculated to find the accuracy of the predicted crops. The error is from −0.3 to 0.65.

Fig. 11 visualizes the performance measures. Performance measures are important part to ensure the efficiency of the system. The performance measures parameters considered are Accuracy, Bias, F1score, MAPE (Mean Absolute Percentage Error), MDAE (Median Absolute Error), MSE (Mean Squared Error), RAE (Relative Absolute Error), RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) (Maria Halkidi et al. 2001). The value for Accuracy and F1 score is nearer to 100%. For the system to be less erratic, RMSE and MAE should be ranged from 0 % to 50 %. Similarly, Bias, MAPE, MAE, MDE, RAE, values should be nearer to 0%. From All the formulas for performance measure parameters are specified in Tab. 6. From Fig. 11, it is observed that the predicted results are highly accurate.

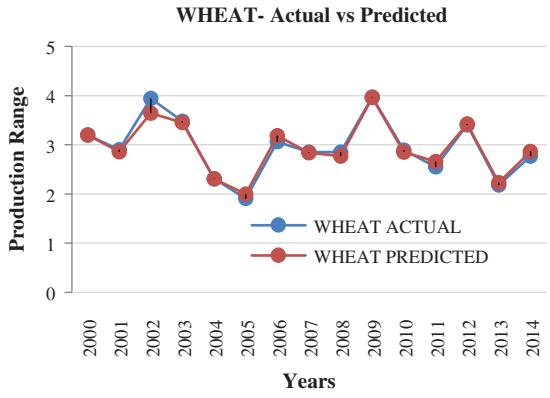

Figure 8: Prediction of wheat

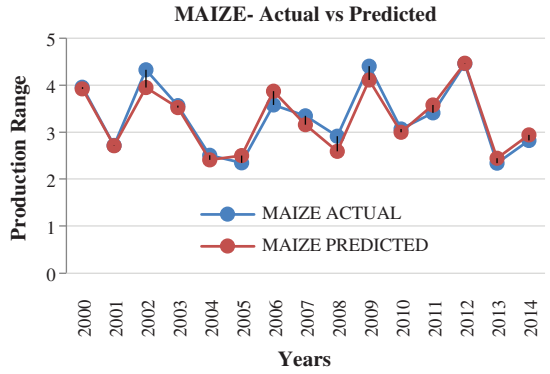

Figure 9: Prediction of maize

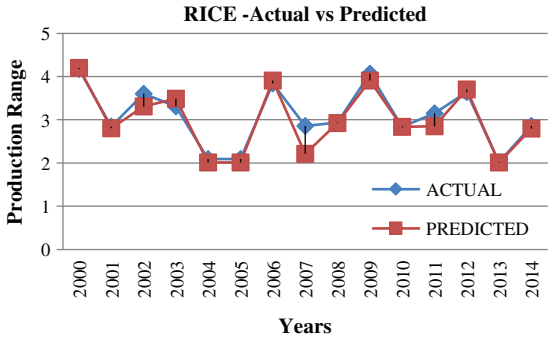

Figure 10: Prediction of rice

Table 5: Actual crop yield and predicted crop yield

Year	Rice				Wheat				Maize			
	Actual	Predicted	Error	% Error	Actual	Predicted	Error	% Error	Actual	Predicted	Error	% Error
2000	4.168	4.201	0.033	0.79	3.199	3.199	0	3.955	3.9205	−0.0345	−0.87	3.955
2001	2.845	2.805	−0.04	−1.41	2.898	2.856	−0.042	2.714	2.712	−0.002	−0.07	2.714
2002	3.602	3.31	−0.292	−8.11	3.941	3.643	−0.298	4.329	3.951	−0.378	−8.73	4.329

Table 5 (continued)

Year	Rice				Wheat				Maize			
	Actual	Predicted	Error	% Error	Actual	Predicted	Error	% Error	Actual	Predicted	Error	% Error
2003	3.31	3.49	0.18	5.44	3.479	3.449	−0.03	3.562	3.521	−0.041	−1.15	3.562
2004	2.089	2.015	−0.074	−3.54	2.315	2.305	−0.01	2.508	2.412	−0.096	−3.83	2.508
2005	2.094	2.012	−0.082	−3.92	1.907	2	0.093	2.348	2.501	0.153	6.52	2.348
2006	3.836	3.908	0.072	1.88	3.059	3.183	0.124	3.575	3.873	0.298	8.34	3.575
2007	2.853	2.211	−0.642	−22.5	2.849	2.838	−0.011	3.345	3.158	−0.187	−5.59	3.345
2008	2.938	2.921	−0.017	−0.58	2.849	2.772	−0.077	2.916	2.59	−0.326	−11.18	2.916
2009	4.067	3.905	−0.162	−3.98	3.973	3.961	−0.012	4.404	4.112	−0.292	−6.63	4.404
2010	2.838	2.835	−0.003	−0.11	2.89	2.851	−0.039	3.067	2.996	−0.071	−2.31	3.067
2011	3.147	2.85	−0.297	−9.44	2.553	2.661	0.108	3.411	3.578	0.167	4.9	3.411
2012	3.637	3.698	0.061	1.68	3.405	3.423	0.018	4.456	4.466	0.01	0.22	4.456
2013	2.021	2.011	−0.01	−0.49	2.185	2.233	0.048	2.342	2.442	0.1	4.27	2.342
2014	2.855	2.798	−0.057	−2	2.772	2.862	0.09	2.823	2.942	0.119	4.22	2.823

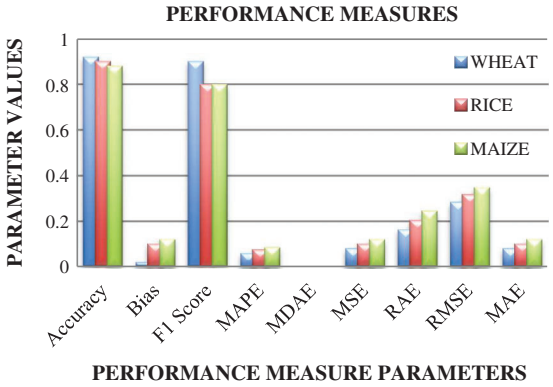

Figure 11: Performance measures on prediction

Table 6: Formula of performance measure parameters

Performance measure parameters	Formula
Accuracy	$\boldsymbol{Accuracy} = \dfrac{\boldsymbol{P} - \boldsymbol{A}}{\boldsymbol{P}}$
Bias	$Bias = Average\ Predictions - Average\ of\ labels\ in\ dataset$
F1 Score	$F1 = 2X\dfrac{Precision * Recall}{Precision + Recall}$
MAPE	$MAPE = \dfrac{1}{N}\sum_{k=1}^{N}\left\lvert\dfrac{A_k - P_k}{A_k}\right\rvert$

(Continued)

Table 6 (continued)

Performance measure parameters	Formula
MDAE	$MDAE = \frac{1}{n}\sum_{i-1}^{n} \lvert A_k - P_k \rvert$
MSE	$MSE = \frac{1}{N}\sum_{k=1}^{N} (A_k - P_k)^2$
RAE	$RAE = \frac{\sum_{i=1}^{N} \lvert A_k - P_k \rvert}{\sum_{i=1}^{N} \lvert A - A_k \rvert}$
RMSE	$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N} (\lvert A_k - P_k \rvert)^2}$
MAE	$MAE = \frac{1}{N}\sum_{k=1}^{N} \lvert A_k - P_k \rvert$
Percentage of Prediction Error	$\% \, Error = (A_k \;\; P_k) \; / \; A_k \; * \; 100$
A-Actual value, P-Predicted value	

5 Conclusion

In this research, machine learning based crop yield prediction is presented which validates the performance using the clustering algorithms like E-DBSCAN, DBSCAN, CLARA, K-Means and Multiple Linear Regression. The feature selection method efficaciously found important features, and discovered that ecological aspects had a greater consequence on the crop yield. Accuracy of this model was found to be comparatively subtle to the prediction techniques. The knowledge data discovery of machine learning algorithms is efficient and accurate. The clustering techniques are applied to crops (rice, wheat, and maize) dataset. It is inferred that the clustering results of E-DBSCAN is better than DBSCAN, CLARA and K-means. The outcomes illustrate that, DBSCAN plays unprecedented performance in clustering. Also, the prediction using multiple linear regression is less error prone. The accuracy of wheat is maximum and among the considered districts of Tamil Nadu and Ariyalur illustrates the highest production. The predicted results are authentic in terms of the performance measures like Accuracy, Bias, F1 Score, MAPE, MDAE, MSE, RAE, RMSE and MAE. Thus, the proposed system, E-DBSCAN aids the farmers to hand-pick the precise crop for harvest to upsurge their productivity reckon on location and finest influences for the growth of crops.

Acknowledgement: The authors acknowledge the support and encouragement of the management and Principal of Sri Ramakrishna Engineering College. We would also like to thank the referees and the editors for their helpful comments and suggestions.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Gopala Pillai and L. Tian, "In-field variability detection and spatial yield modelling for corn using digital aerial imaging," *Transactions of the ASAE*, vol. 42, no. 6, pp. 1911–1920, 1999.

[2] Y. Lan, Y. Huang, D. E. Martin and W. C. Hoffmann, "Development of an airborne remote sensing system for crop pest management: system integration and verification," *Applied Engineering in Agriculture*, vol. 25, no. 4, pp. 607–615, 2009.

[3] United Nations University, Japan, Causes of shortage," 2017. [Online]. Available: http://archive.unu.edu/unupress/unupbooks/uu22we/uu22we0a.htm.

[4] J. Majumdar, S. Naraseeyappa and S. Ankalaki, "Analysis of agriculture data using data mining techniques: Application of big data," *Journal of Big Data*, vol. 4, pp. 1–15, 2017.

[5] T. Klompenburg, A. Kassahun and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, pp. 1–18, 2020.

[6] J. Han and M. Kamber, *Data mining concepts and techniques*, 2nd edition. Elsevier, 2017.

[7] R. Sujatha and P. Isakki, "A study on crop yield forecasting using classification techniques," in *Int. Conf. on Computing Technologies and Intelligent Data Engineering*, Kovilpatti, India, pp. 1–4, 2016.

[8] D. Ramesh and B. Vishnu Vardhan, "Analysis of crop yield prediction using data mining techniques," *International Journal of Research in Engineering and Technology*, vol. 4, no. 1, pp. 470–473, 2015.

[9] Agriculture - Data.gov, Dataset in agricultural sector," 2017. [Online]. Available: https://data.gov.in/.

[10] CWWG data, Crop wise agriculture data," 2017. [Online]. Available: https://agricoop.nic.in/all-india-crop-situation.

[11] Karnataka State Department of Agriculture (KSDA), Agriculture data of different districts," 2017. [Online]. Available: http://14.139.94.101/fertimeter/Distkar.aspx.

[12] Karnataka State Department of Agriculture (KSDA), Karnataka Agricultural Statistics. 2017. [Online]. Available: http://raitamitra.kar.nic.in/ENG/statistics.asp.

[13] J. Majumdar, Agriculture data based on weather, temperature, and relative humidity," 2017. [Online]. Available: http://dmc.kar.nic.in/trg.pdf.

[14] T. M. S. Ahamed, N. T. Mahmood, N. Hossain, M. T. Kabir, K. Das *et al.,* "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh," in *16th Int. Conf. on Software Engineering*, Bangladesh, pp. 1–6, 2015.

[15] J. Gholap, A. Ingole, J. Gohil, S. Gargade and V. Attar, "Soil data analysis using classification techniques and soil attribute prediction," *International Journal of Computer Science*, vol. 3, pp. 415–418, 2012.

[16] R. Sujatha and P. Isakki, "A study on crop yield forecasting using classification techniques," in *IEEE Conf. on Computing Technologies and Intelligent Data Engineering*, Kovilpatti, India, pp. 1–4, 2016.

[17] N. Hemageetha, "A survey on application of data mining techniques to analyse the soil for agricultural purpose," in *3rd Int. Conf. on Computing for Sustainable Global Development*, New Delhi, pp. 3112–3117, 2016.

[18] B. R. Aishwarya, "A literature study on application of data mining tools for rice yield prediction," *International Journal of Innovative Technology and Research*, vol. 4, no. 1, pp. 2757–2759, 2016.

[19] D. Diepeveen and L. Armstrong, "Identifying key crop performance traits using data mining," in *World Conf. on Agricultural Information and IT*, IAALD AFITA WCCA2008, Tokyo, Japan, pp. 1–21, 2008.

[20] V. Ramesh and K. Ramar, "Classification of agricultural land soils: A data mining approach," *Agricultural Journal*, vol. 6, no. 3, pp. 82–86, 2011.

[21] F. K. Van Evert, S. Fountas, D. Jakovetic, V. Crjevic, I. Travlos *et al.,* "Big data for weed control and crop protection," *Weed Research*, vol. 57, pp. 218–233, 2017.

[22] P. Bose, N. K. Kasabov, L. Bruzzone and N. Harto, "Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 11, pp. 6563–6573, 2016.

[23] P. S. Maya Gopal and R. Bhargavi, "Optimum feature subset for optimizing crop yield prediction using filter and wrapper approaches," *ASABE—Applied Engineering in Agriculture*, vol. 35, no. 1, pp. 9–14, 2019.

[24] P. S. Maya Gopal and R. Bhargavi, "Selection of important features for optimizing crop yield prediction," *International Journal of Agricultural and Environmental Information Systems*, vol. 10, no. 3, pp. 54–71, 2019.

[25] A. T. Mohammed, S. Irmak, W. L. Kranz, S. Van Donk and C. Dean Yonts, "Grain yield, crop and basal evapotranspiration, production functions and water productivity response of drought-tolerant and non-drought-tolerant maize hybrids under different irrigation levels and population densities," *Applied Engineering in Agriculture*, vol. 35, no. 1, pp. 61–81, 2019.

[26] T. Iizumi, Y. Shin, W. Kim, M. Kim and J. Choi, "Global crop yield forecasting using seasonal climate information from a multi-model ensemble," *Climate Services*, vol. 11, pp. 13–23, 2018.

[27] G. Niedbała, "Simple model based on artificial neural network for early prediction and simulation winter rapeseed yield," *Journal of Integrative Agriculture*, vol. 18, no. 1, pp. 54–61, 2019.

[28] V. Ricciardi, N. Ramankutty, Z. Mehrabi, L. Jarvis and B. Chookolingo, "An open-access dataset of crop production by farm size from agricultural censuses and surveys," *Data in Brief*, vol. 19, pp. 1970–1988, 2018.

[29] M. Van der Velde and L. Nisini, "Performance of the MARS-crop yield forecasting system for the European Union: Assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015," *Agricultural Systems*, vol. 168, pp. 203–212, 2019.

[30] R. Lecerf, A. Ceglar, R. López-Lozano, M. Van Der Velde and B. Baruth, "Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe," *Agricultural Systems*, vol. 168, pp. 191–202, 2018.

[31] D. T. Meshesha and M. Abeje, "Developing crop yield forecasting models for four major Ethiopian agricultural commodities," *Remote Sensing Applications: Society and Environment*, vol. 11, pp. 83–93, 2018.

[32] G. E. O. Ogutu, W. P. H. Franssen, I. Supit, P. Omondi and R. W. A. Hutjes, "Probabilistic maize yield prediction over east Africa using dynamic ensemble seasonal climate forecasts," *Agricultural and Forest Meteorology*, vol. 250–251, pp. 243–261, 2018.

[33] M. J. Lambert, P. C. S. Traoré, X. Blaes, P. Baret and P. Defourny, "Estimating smallholder crops production at village level from sentinel-2 time series in mali's cotton belt," *Remote Sensing of Environment*, vol. 216, pp. 647–657, 2018.

[34] District-wise, "Season-wise crop production statistics," 2017. [Online]. Available: https://data.gov.in/.

[35] Computer Science and Engineering, "IIIT-Delhi Institutional Repository," 2017. [Online]. Available: https://repository.iiitd.edu.in.

[36] A. Kassambara, CLARA," 2017. [Online]. Available: http://www.sthda.com/english/articles/27-partitioning-clustering-essentials/89-clara-clustering-large-applications.

[37] A. Trevino, K-means Clustering," 2017. [Online]. Available: https://www.datascience.com/blog/k-means-clustering.

[38] M. Kryszkiewicz and P. Lasek, *TI-DBSCAN: Clustering with dbscan by means of the triangle inequality*. Berlin Heidelberg: Springer-Verlag, pp. 60–69, 2010.

[39] A. M. Keith, Advanced Statistics, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1197/j.aem.2003.09.006.

[40] R. Jain and C. Dubes, Algorithms for Clustering Data, Cluster Validation. 2017. [Online]. Available: www.cs.kent.edu/~jin/DM08/ClusterValidation.pdf.

[41] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2/3, pp. 107–145, 2001.