Environment, Development and Sustainability
https://doi.org/10.1007/s10668-019-00445-x

Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression

Ekaansh Khosla[1] · **Ramesh Dharavath**[1] · **Rashmi Priya**[1]

Received: 17 October 2018 / Accepted: 8 August 2019

Abstract
At the present time, one of the most important sources of survival as well as the most crucial factor in the growth of Indian economy is agriculture. More than 70% of the Indian population is involved in agricultural activities. The crop yield prediction is one of the most desirable yet challenging tasks for every nation. Nowadays, due to the unpredictable climatic changes, farmers are struggling to obtain a good amount of yield from the crops. To feed the increasing population of India, there is a need to incorporate the latest technology and tools in the agricultural sector. This study focuses on the prediction of major kharif crops in Andhra Pradesh's one of the largest costal districts: Visakhapatnam. As rainfall is the main factor in determining amount of kharif crop production, in this study, first we predict the amount of monsoon rainfall by using modular artificial neural networks (MANNs), and then, we predict the amount of major kharif crops that can be yielded by using the rainfall data and area given to that particular crop by using support vector regression (SVR). By using the methodology of MANNs-SVR, proper agricultural strategies can be made in order to increase the yield of the crops. Comparison with other machine learning algorithms has been done which shows that the proposed methodology outperforms in predicting the instances for kharif crop production.

Keywords Agriculture · Crop modeling · Machine learning · Support vector regression · Yield prediction

✉ Ramesh Dharavath
drramesh@iitism.ac.in

Ekaansh Khosla
ekaanshkhosla007@gmail.com

Rashmi Priya
rashmi.17DR000561@cse.ism.ac.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, Jharkhand 826004, India

Published online: 21 August 2019

1 Introduction

Agriculture is one of the real wellsprings of the Indian economy (Kumar and Bhattachrya 2015; Sujatha and Isakki 2016). Over 70% of the Indian population is associated with agrarian exercises (Borlaug 2002). The forecast for agricultural yield is one of the most attractive yet difficult errands for each country. These days, because of the unusual climatic changes, farmers are attempting to acquire a decent measure of yield from the products. According to the data survey 2012, it was found that India lacks behind in the rice yield (kg per hectare) from many countries like China, Bangladesh, Vietnam, Indonesia, Japan, Brazil and Philippines (Kannan and Sundaram 2011). Another major problem is diminishing cultivable land due to the rising population. Other influencing factors are depending on traditional agricultural methods and level of infrastructure, etc. In Andhra Pradesh (AP), 60% of the population is involved in agricultural and related activities (Ramachandran et al. 2010). Rice is the prime food crop of AP. In AP, the other season crops such as bajra, ragi, maize, sugarcane, cotton are also cultivated. Four major rivers of India: Krishna, Penna, Godavari and Tungabhadra run across the state providing irrigation. In addition to that, many multistate irrigation projects are under development which include Nagarjuna Sagar Dam and Godavari River Basin Irrigation Projects. In the 1970s, Andhra Pradesh went for the Green Revolution in rice cultivation. At constant prices in 2012–2013, the average per capita income of agriculture was 850 million US Dollars in the state.

Visakhapatnam is the Andhra Pradesh's largest city as well as financial capital. Its geographical location is amidst Eastern Ghats mountain range and the coast of Bay of Bengal. Figure 1 depicts the position of Visakhapatnam in India. It is the 14th largest city and the 9th most populous metropolitan city of India with a population of over 2 million (Maiti and Agrawal 2005). As of now, Visakhapatnam is the 9th largest contributor to India's overall GDP with an output of $43.5 billion. In Visakhapatnam, 70% of the households depends

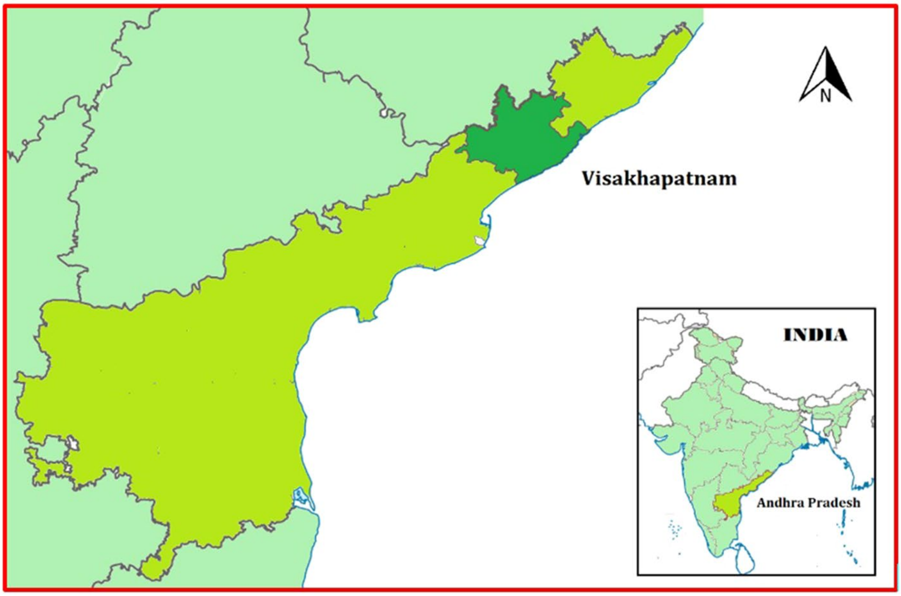


Fig. 1 Position of Visakhapatnam in India

on the agriculture. In the district, rice is the staple crop, and hence, paddy is the principal crop followed by ragi, bajra and jowar. Because of the limitation of the major irrigation system in Visakhapatnam, about 36% of the area which is cropped is irrigated with minor irrigation tanks or medium irrigation system; the rest of the area which is to be cultivated is covered with dry crops (Gaddeyya and Kumar 2014).

In order to feed the increasing population of India, there is a need to incorporate the latest technological methodologies in the agricultural sector. In addition to this, farmers require timely advice to predict the crop productivity so that they can make proper strategies to increase the yield of their crops. Precision agriculture is an approach which uses technology in order to make sure that soil and crops get what they need for an optimum productivity and health. In precision agriculture, the real-time farm and weather data are collected using the sensors and are used to make predictions to help farmers in taking correct farm-related decision. Small sensors are deployed in the farm, where it collects and sends the data to the relevant data storage node (McBratney et al. 2005). The data collected are huge in volume and hence can be processed using big data analytics. Big data provide facilities like data storage, data processing and data analysis with accuracy. So its use in the field of agriculture can benefit farmers and nation's economic growth (Howe et al. 2008). With the help of big data analytics and related machine learning algorithms, crop productivity can be increased by many folds. To accommodate the prediction strategies for kharif, in this study, we present a methodology named MANNs-SVR. By using MANNs-SVR, first, we predict amount of occurrence of rainfall in the region of Visakhapatnam. Then, by using attributes which give the most information about production of kharif crops in Visakhapatnam, we predict amount of common kharif crops, i.e., rice, ragi, maize and bajra, that can be yielded in the upcoming years. Most of the past studies were mainly focused on image processing and prediction using statistical models. Since the proposed work uses machine learning approaches, it is fast in computation and more efficient compared to statistical methods proposed in the past, and also this is simple in terms of application implementation.

The road map of this paper is organized as follows: Sect. 2 describes the related works in the field of rainfall prediction and precision agriculture. Section 3 discusses about the proposed methodology for predicting the rainfall as well as yield of kharif crops in Visakhapatnam. In Sect. 4, we evaluate the proposed methodology and discuss the experimental results. In Sect. 5, we conclude the methodology for the yield of kharif crops.

2 Related works

In the literature, many agriculture-related studies have been suggested to make the agricultural environment more suitable (Morshed et al. 2013). In the current scenario, researchers' focuses mainly on the analytical mechanisms, which produce the limited information of the crops. These data may not be sufficient to yield the crops. Hence, in order to make farm-specific data analytics, a good connection is required between these two parties. Agricultural production mainly depends on weather and climate changes. Proper weather conditions result in higher production of a suitable crop. Using high-quality seeds also leads to higher productivity. But, the problem with using high-quality seeds is that while making predictions we have to examine the genotype and the phenotype of the crop and their response to the environment (Parent and Tardieu 2014). Other factors which can affect the productivity of certain crop are water and nutrient content in the soil, weeds and their control (Rosenzweig et al. 2015).

In the past studies, a number of related crop prediction methodologies have been presented. Lobell and Field (2007) presented a work on effects of global warming and climate change on six most growing crops. Yinhong Kang et al. 2009 described a comprehensive study of climate, water and crop yield model to find the impacts of climate on the crop. Cantelaube and Terres (2005) proposed a weather forecast model, by using meteorological data for crop yield in Europe. A precision agriculture-based study on statistical model is presented by Bornn and Zidek (2012) for the Canadian Prairies. For Poland, a suitable crop growth condition and yield based on AVHRR are discussed (Dabrowska-Zielinska et al. 2002). Considering the remotely collected data of soil moisture and leaf area index, Ines et al. (2013) proposed a crop model using sequential data assimilation in order to find yield in maize crop. Because of the impact of extreme climatic conditions of winter and summer on Mediterranean crop, the work presented by Moriondo et al. (2011) suggests a suitable seasonal crop. The authors De Wit and Van Diepen (2007) presented a data assimilation model based on ensemble Kalman filter in order to overcome the error incurred in the crop yield model due to uncertainty in temporal rainfall distribution. Haboudane et al. (2002) presented a work to find the crop yield by using chlorophyll content in the leaf area index. Based on the study of four spectral vegetation indices, namely soil adjusted vegetarian index, perpendicular vegetation index, normalized difference vegetation index and green vegetation index, a neural network supported methodology is presented to predict the corn yield (Panda et al. 2010). A statistical model on agriculture is proposed to find the effects of climate change on crop yield using historical data (Lobell and Burke 2010). In order to find out water consumption by crop, crop yield and use of water efficiency, a prediction model based on SVAT crop growth was proposed (Mo et al. 2005). Oettli et al. (2011) proposed a comparative study of nine regional climate models in West Africa for agriculture. Ullah et al. (2018) presented the use of diurnal temperature range with the combination of two climate models to find relationship between DTR, daily mean temperature and crop yield. To build the gap between dynamic and statistical downscaling, Salehnia et al. (2019) made a comparative study between the two to evaluate the historical precipitation data.

The production of a certain crop mainly depends on the weather. On the other hand, a suitable weather prediction strategy is required to increase the productivity of the crop. Wu et al. (2010) proposed a model in which they compared modular artificial neural network with artificial neural network, *k*-nearest neighbors and linear regression. This study showed that optimal rainfall predicting model could be constructed from modular artificial neural networks (MANNs) coupled with preprocessing techniques, such as singular spectrum analysis, and without using preprocessing techniques, i.e., in normal mode. In the present study, the amount of rainfall that can occur in Visakhapatnam is predicted first by using modular artificial neural networks. Then, by using attributes, which give most information about production of kharif crops in Visakhapatnam, the amount of common kharif crops, i.e., rice, ragi, maize and bajra that can be yielded in the upcoming years in city, is predicted.

3 Proposed methodology

This section describes the working methodology of the proposed algorithms, dataset and experimental setup that are used in predicting the amount of yield of kharif crop in the region of Visakhapatnam. The proposed model for prediction of various kharif crops in Visakhapatnam in the upcoming years is shown in Fig. 2.

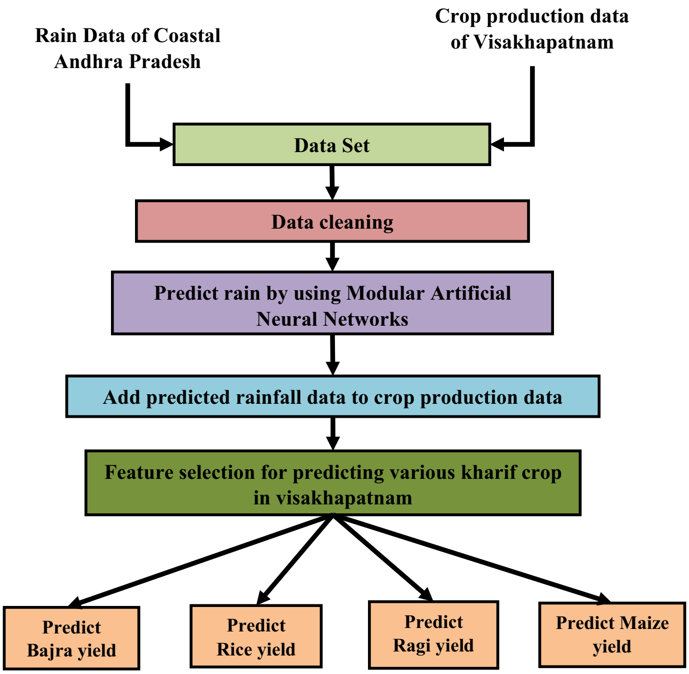


Fig. 2 Workflow for prediction of Kharif crops in Visakhapatnam

3.1 Methodology

In this section, we describe the proposed methodology with its related variants. To predict the monsoon rainfall, first, input/output pairs are reconstructed; then, modular artificial neural network is applied in order to predict the monsoon rainfall in Visakhapatnam.

3.1.1 Construction of input/output pairs

Let us consider $\{X_1, X_2, \ldots, X_N\}$ is a rainfall series. The series can be regenerated into a series of delay vectors as $X_t = \{X_t, X_{t+\tau}, X_{t+2\tau}, \ldots, X_{t+(m-1)\tau}\}$ where $X_t \in R^m$, τ is the delay time as multiple of the sampling period and m is the embedded dimension, i.e., the data space dimension. Let the rainfall $X_{t+T+(m-1)\tau}$ at T-step lead be associated with the vector X_t, then the past data can be encapsulated into set of pairs as $\{X_{t+T+(m-1)\tau} : t = 1, \ldots, n\}$, where n is the number of pairs with its functionality $n = N - (m-1)\tau$.

The functional relationship between the input vector X_t at time t and the predicted output $X^F_{t+T+(m-1)\tau}$ at time $t + T$ can be written as follows:

$$X^F_{t+T+(m-1)\tau} = f(X_t) + e_t \tag{1}$$

where e_t is the typical noise term, $X^F_{t+T+(m-1)\tau}$ is the prediction of $X_{t+T+(m-1)\tau}$ and $f(.)$ is the mapping function. In the current study, $f(.)$ is approximated by modular artificial neural networks.

3.2 Modular artificial neural networks (MANNs)

For constructing modular artificial neural networks, first, the training data are divided into several clusters, and then, on each cluster artificial neural network (ANN) is applied. In the present study, Fuzzy C-Means clustering is used to generate the clusters (Wang et al. 2006; Bezdek 2013).

Figure 3 shows the methodology for predicting rainfall by using modular artificial neural networks. According to the figure, first the model input by one of the six methods is decided, which are: linear correlation analysis (Sudheer et al. 2002), average mutual information (Fraser and Swinney 1986), partial mutual information (May et al. 2008), false nearest neighbors (Kennel et al. 1992), SLR and MOGA (Giustolisi and Savic 2006). After obtaining input–output pairs, data are divided into three clusters by Fuzzy C-Means clustering technique based on magnitudes of the rainfall that are low, medium and high. After this, the artificial neural network is applied to each of the clusters. The final output comes from one of the three local models.

3.3 Artificial neural networks (ANNs)

The multilayer perceptron is one of the most popular paradigms of artificial neural networks, which use error back propagation to train the network. In the architecture of artificial neural networks, the input layer, hidden layer and the output layer consist of the number of neurons. In the literature, artificial neural networks with one hidden layer are used for hydrologic modeling, because these networks provide enough complexity in accurately simulating the nonlinear properties of hydrologic process (Dawson and Wilby 2001; De Vos and Rientjes 2005). Based on Eq. (1), the artificial neural network forecasting model is formulated as:

$$X^F_{t+T+(m-1)\tau} = f(X_t, w, \theta, m, h) = \theta_0 + \sum_{j=1}^{h} w_j^{out} \varphi\left(\sum_{i=1}^{m} w_{ji} X_{t+(i-1)\tau} + \theta_j\right) \tag{2}$$

Model notations used in the above equation are described as follows: (1) $\varphi \rightarrow$ transfer functions, (2) $w_{ji} \rightarrow$ weights describing connection between ith node of the input layer and

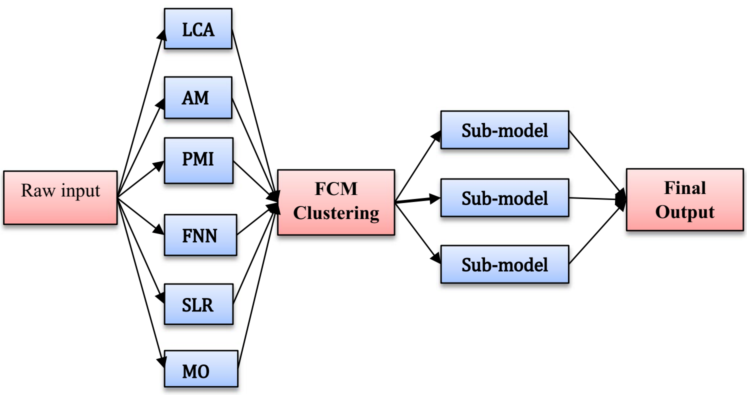


Fig. 3 Methodology for predicting rainfall

jth node of the hidden layer, (3) $\theta_j \rightarrow$ biases related to jth node of the hidden layer, (4) $w_j^{\text{out}} \rightarrow$ weights related to connection between jth node of the hidden layer and node of the output layer and (5) $\theta_0 \rightarrow$ bias at the output layer. In order to apply Eq. (2) to the rainfall predictions, a suitable training algorithm is required in order to optimize values of w and θ.

3.4 Applications of model

3.4.1 Determination of model inputs

In the present study, artificial neural networks provided with Levenberg–Marquardt (L–M) training algorithm and hyperbolic tangent sigmoid transfer functions are used as a point of reference to examine previously mentioned six model inputs in terms of root mean square error (RMSE). Results from six methods are presented in Table 1. The results are based on one-step lead prediction, and let m be the target value at one-step prediction horizon. Table 1 shows that RMSE values of these methods are close to one another. Hence, these methods are mutually replaceable. For better calculation and accuracy, the LCA method is chosen for this study.

3.4.2 Identification of model

For the identification of artificial neural network with known inputs and outputs, the number of hidden nodes h in the hidden layer must be optimized. For this, the number of hidden neurons is systematically increased from 1 to 10 until reasonable performance is obtained. Based on Levenberg–Marquardt training algorithm and hyperbolic tangent sigmoid transfer functions, the identified configurations of artificial neural network for proposed study are 12-5-1. The same method is used to identify the structure of modular artificial neural network, and the only difference is that the identification is repeated three times with each time being for a local artificial neural network. Consequently, modular artificial neural network is obtained as 12-3/2/5-1 for the proposed study in Visakhapatnam.

It is very important to standardize or normalize the training data in order to improve the performance of the model. In the past study, we can find a few models (Cannas et al. 2002; Rajurkar et al. 2002; Campolo et al. 2003). Standardization is basically a method to rescale the training data to [−1, 1], [0, 1] or even smaller interval depending on the transfer function used in artificial neural network. In normalization, the training data are rescaled to a Gaussian function with a mean of 0 and unit standard deviation which is done by subtracting the mean and dividing by the standard deviation. When the normalization approach is adopted, the artificial neural network uses the linear function instead of the hyperbolic

Table 1 Comparison of methods to determine mode inputs using ANN model

Methods	τ	m	Effective inputs	Identified ANN	RMSE
LCA	1	20	The last 12	(12-5-1)	63.74
AMI	1	12	The last 12	(12-5-1)	63.74
PMI	1	12	$X_{t-11,t-10,t-5,t}$	(4-5-1)	71.07
FNN	1	20	The last 5	(5-9-1)	79.85
SLR	1	12	Except for X_{t-4}	(11-9-1)	66.12
MOGA	1	12	$X_{t-11,t-9,t-7,t-5,t-4,t-1,t}$	(7-1-1)	76.24

tangent sigmoid transfer function in the output layer. Also, some studies have indicated that consideration of statistical principles may improve the performance of ANN model (Cheng and Titterington 1994). Fortin et al. (1997) recommended that training data should be normally distributed. Sudheer et al. (2002) exemplified that ANN cannot account for trends and heteroscedasticity in the data; hence, stationarity should be considered in the development of ANN. The results indicated that to improve the model performance, data transformation can be used which reduces the skewness of data. In order to perform the prediction, we adopt the methodology proposed by Oettli et al. (2011).

3.5 Methodology used for prediction of production of kharif crop

In the present study, a number of regression techniques such as k-nearest neighbor regression, random forest regression, linear regression and support vector regression were applied in order to predict the production of kharif crop in which SVR showed the best results. Hence, in the current study SVR is applied for predicting production of kharif crops in Visakhapatnam.

3.5.1 Support vector regression (SVR)

Support vector regression (SVR) and support vector machine (SVM) use the same concepts with a few dissimilarity. For solving regression problems, we can extend the methodology used by support vector classification. Constructing a support vector classification model depends only on a subset of training data because the cost function for constructing the support vector classification model does not consider training points beyond a certain boundary. Similarly, a subset of data is required for building a support vector regression model (Curtin et al. 2013). The SVR is put together as minimization of the following (Basak et al. 2007). Figure 4 shows different parameters used in these equations.

$$\frac{1}{2}||w||^2 + c\sum_{i=1}^{N} \xi_i + \xi_I^* \tag{3}$$

With constrains, when $y = wx + b$,

$$\begin{aligned} & y_i - wx - b \leq \varepsilon + \xi_i \\ & wx_i + b - y_i \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{aligned}$$

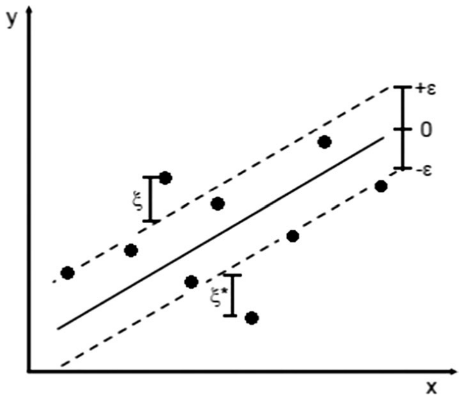
where ξ_i and ξ_i^* are the slack variables and $-\varepsilon$ and $+\varepsilon$ are the distance of hyper plane and the boundary line.

In the present study, due to limitation of high-speed computing system, linear support vector regression is used and formulated as (Juang and Hsieh 2012):

$$y = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*)(x_n \cdot x) + b \tag{4}$$

where α_n, α_n^* are nonnegative multipliers with for each observation x_n and $x_n.x$ is the dot product.

Fig. 4 Parameters used in SVR minimization equation



4 Experimental and result analysis

4.1 Experimental setup

In this section, we discuss the predicted results of production of kharif crop in Visakhapatnam in the upcoming years. For making yield predictions of all the kharif crops, in the present study, we divide the training data in the form of four consecutive years, and then, we use three years as training data in each of the divided parts and one year for testing the model. The input for the model is weather parameter, and the output is yield obtained. In order to evaluate the proposed methodology with modular artificial neural networks (MANNs) and support vector regression (SVR), the following experimental cluster setup has been considered.

Operating system	Ubuntu 16.04
Workstation Quad core x64-based processor	With 4 clusters (each with 8 GB RAM)
Backend	Tensorflow
Storage	1 TB with 280 GB extra SATA

4.2 Dataset

The dataset considered in the present study is collected from various sources and merged into one, such that a dataset sufficiently large enough for the study is created. The rainfall data collected from Indian government site (data.gov.in) are from 1901 till 2015. Again the rainfall data for the years 2016 and 2017 were collected from meteorological site of India. Similarly, the crop-related data of rice, maize, bajra and ragi were collected from data.gov.in, which contain the attributes like area under production, total yield and the cropping season. The rainfall data contain the rainfall amount of each month from 1901 till 2017. There is also a dedicated field for the amount of rainfall received in monsoon every year. A dedicated data attribute for monsoon was taken with the purpose to make the proposed model work efficiently.

The data given were for costal area in Andhra Pradesh. As Visakhapatnam comes in the coastal area of Andhra Pradesh, we use the same data for our experiment. Table 2 shows the basic structure of dataset used for predicting the monsoon rainfall in Visakhapatnam in the years 2018 and 2019. For training our model, we use the data from 1901 to 2000, and to test our model, we use the data from 2001 to 2017. With this, the final prediction has been done for the monsoon rainfall in Visakhapatnam in the years 2018 and 2019.

After predicting the monsoon rainfall, we use the predicted rainfall in order to find the yield of kharif crop in the upcoming years. For this, first, we added the rainfall information to the crop production data, and after that, we perform feature selection on the attributes. With this step, important features are selected in predicting the yield of the different kharif crops. The data for crop production in Visakhapatnam were taken from data.gov.in, an Indian government data site. As the data for yield prediction of kharif crop were from only 1997 to 2014, the rest of the data were taken from other government agencies. Table 3 shows the dataset used for predicting bajra, Table 4 shows the dataset used for predicting maize, Table 5 shows the dataset used for predicting rice, and Table 6 shows the dataset used for predicting ragi. As kharif crop only depends on rainfall in the months of June and September, in our training data, for predicting yield of the crops, we only use monsoon months as attributes. Also, as production of a certain crop depends on factors like area which can be given to that particular crop, consumption of that particular crop, etc., which can differ time to time, for this experiment we divided our training data in the form of four consecutive years, and then, we use 3 years as training data in each of the divided parts and 1 year for testing the model. For example, we consider training data for the years 1997, 1998, 1999, 2001, 2002, 2003, etc., for training our model and then tested our model in the years 2000, 2004 and so on.

4.3 Examining the dataset

In this section, we analyze how the data of rainfall change different Kharif crops over the years in Visakhapatnam.

4.3.1 Analyzing the monsoon rainfall in Visakhapatnam

Figure 5 shows the monsoon rainfall in Visakhapatnam from 1901 to 2017. In Visakhapatnam, min monsoon rainfall between 1901 and 2017 was 393.2 mm in the year 1920, whereas max monsoon rainfall was 954 mm in the year 2010. The monsoon rainfall in Visakhapatnam from 1901 to 2017 is 655.1 mm.

4.3.2 Analysis for different kharif crop productions in Visakhapatnam

Figure 6 shows the area given and production of related crop in Visakhapatnam from 1997 to 2017. Green line shows the area in hector, and red line shows its production in tones. From Fig. 6a, we can see that area given to bajra over the years in Visakhapatnam has been decreased, and because of that, its production has also decreased in these years. From Fig. 6b, we can infer that area given to maize crop has not changed over the time, but there has been changes in the production of maize in Visakhapatnam. From Fig. 6b, we can conclude that production of maize in Visakhapatnam does not depend highly on the area given to maize; it rather depends on many other factors also. From

Table 2 Structure of Dataset used for predicting the rainfall in Visakhapatnam

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Monsoon
1901	18.8	80.9	7.2	28.7	68.7	77.7	113	133.7	125.3	173.4	164.8	1.5	993.8	449.7
1902	2	0	2.8	23.9	37.6	72.6	144.5	236.1	204.5	262	50.4	27.1	1063.6	657.7
1903	0.8	13.3	0.2	6.2	73.4	154	248.6	258	216.5	159.1	173.9	12.1	1316.2	877.1
1904	1.3	0	5.4	3	136.3	107.8	120.2	117.7	116.8	240.9	0	10.7	860.2	462.6
1905	1.1	16.7	68	37	68.8	84.4	64.6	210.8	170.2	66	7.4	0	795.2	530.1

Table 3 Dataset used for predicting bajra

Year	Crop	Area	Jun	Jul	Aug	Sep	Monsoon	Production (t)
1997	Bajra	17,700	63.1	156	128.1	315.5	662.7	9700
1998	Bajra	25,000	136.2	214.1	252.2	255.5	858	28,700
1999	Bajra	18,351	126.4	154.2	136.6	137.5	554.7	16,956
2000	Bajra	17,746	200.5	154.1	311.5	102.9	768.9	19,663
2001	Bajra	15,686	117.2	123	170.7	187.2	598.1	13,019

Table 4 Dataset used for predicting maize

Year	Crop	Area	Jun	Jul	Aug	Sep	Monsoon	Production (t)
1997	Maize	6400	63.1	156	128.1	315.5	662.7	6500
1998	Maize	6400	136.2	214.1	252.2	255.5	858	8600
1999	Maize	6419	126.4	154.2	136.6	137.5	554.7	9134
2000	Maize	6704	200.5	154.1	311.5	102.9	768.9	6979
2001	Maize	6879	117.2	123	170.7	187.2	598.1	7567

Table 5 Dataset used for predicting rice in Visakhapatnam

Year	Crop	Area	June	July	Aug	Sep	Monsoon	Production (t)
1997	Rice	95,600	63.1	156	128.1	315.5	662.7	79,700
1998	Rice	114,900	136.2	214.1	252.2	255.5	858	141,300
1999	Rice	103,372	126.4	154.2	136.6	137.5	554.7	134,010
2000	Rice	115,239	200.5	154.1	311.5	102.9	768.9	183,345
2001	Rice	86,579	117.2	123	170.7	187.2	598.1	145,020

Table 6 Dataset used for predicting ragi in Visakhapatnam

Year	Crop	Area	Jun	Jul	Aug	Sep	Monsoon	Production (t)
1997	Ragi	35,800	63.1	156	128.1	315.5	662.7	14,600
1998	Ragi	33,300	136.2	214.1	252.2	255.5	858	31,900
1999	Ragi	33,856	126.4	154.2	136.6	137.5	554.7	35,041
2000	Ragi	34,119	200.5	154.1	311.5	102.9	768.9	32,038
2001	Ragi	32,390	117.2	123	170.7	187.2	598.1	37,151

Fig. 6c, we can say that area given to rice crop has not changed over the time, but we see from the figure that production of rice depends on area given to the crop. And finally from Fig. 6d, we can infer that area given to ragi over the years in Visakhapatnam has been decreased, and because of that, its production has also decreased in these years.

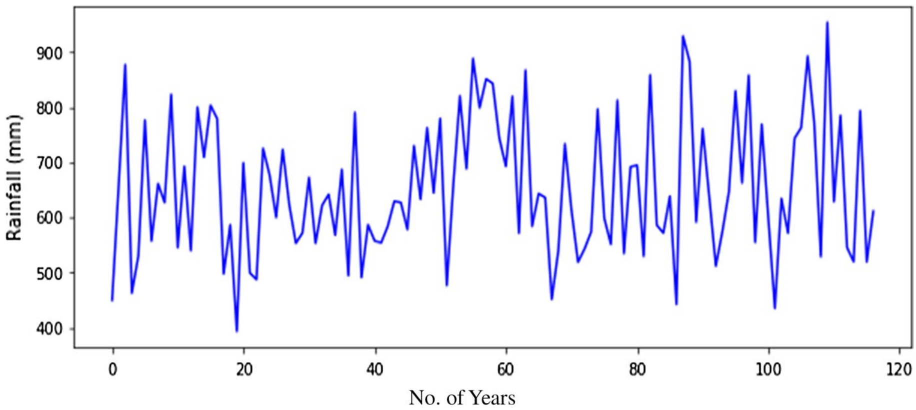
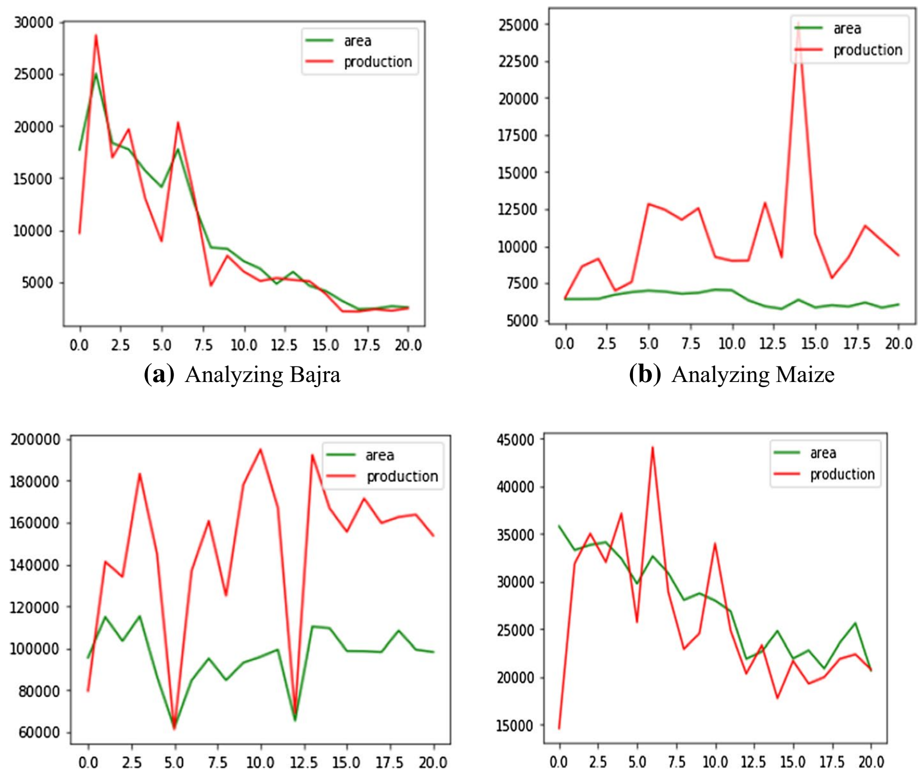


Fig. 5 Analyzing the monsoon rainfall in Visakhapatnam



(a) Analyzing Bajra

(b) Analyzing Maize

(c) Analyzing Rice

(d) Analyzing Ragi

Fig. 6 Analyzing different kharif crops in Visakhapatnam

From Fig. 6, we can conclude that area given to bajra and ragi has been decreased over the years, and due to that, their production has also decreased, whereas no changes are seen in the case of maize and rice. The decrease in area can be due to increasing population in the city and less area being given to these particular crops, or maybe the demand of bajra and ragi has been decreasing over the years in the city.

4.4 Prediction of rainfall in Visakhapatnam

For predicting the rainfall in Visakhapatnam, we use the data as depicted in Table 2, and because our study is limited to kharif crops only, we removed all the attributes accept the monsoon rainfall and then applied modular artificial neural networks to the data. We trained our model from 1901 to 1999 by using linear correlation analysis (LCA) method. After this, we test our model on monsoon rainfall data from 2000 to 2017 based on one-step lead prediction and finally predicted monsoon rainfall in the upcoming years: 2018 and 2019. For predicting monsoon rainfall in 2019, we use the predicted monsoon rainfall data of the year 2018. Figure 7 shows the amount of rainfall that can occur in the monsoon season in the upcoming years.

From Fig. 7, we can conclude that our model gives accurate results with 58.4-mm mean error. To calculate the predicted rainfall in the months of June, July, August and September, we used the data presented in Table 7.

The rainfall prediction in Visakhapatnam for upcoming years by our model is depicted in Table 8. We use these data for the future predictions.

The predicted data are added to the dataset shown in Tables 3, 4, 5 and 6 in order to predict yield of bajra, maize, rice and ragi, respectively.

4.5 Predicting the yield of bajra in Visakhapatnam

For the yield prediction of bajra in Visakhapatnam, first, we apply feature selection method so that only important features can be selected in order to give the predictions. Figure 8 shows the important features in determining the yield of bajra in Visakhapatnam. From Fig. 8, we can infer that area, rainfall in September and total monsoon rainfall are three most important features in determining the amount of yield in Visakhapatnam. We also note that the total area is one of the important attributes with importance of more than 80% in determining the yield of bajra in Visakhapatnam.

Figure 9 shows the performances of the various regression algorithms used for making predictions of yield of bajra. From Fig. 9a, we can infer that SVR gave the most

Fig. 7 Predicted rainfall in Visakhapatnam in the upcoming years

Table 7 Mean rainfall data of Visakhapatnam from 1901 to 2017

Time	Mean rainfall (mm)	% of monsoon rainfall
June	123.7	18.8
July	173.8	26.5
August	176	26.8
September	181.7	27.9
Monsoon	655.1	100

Table 8 Predicted rainfall in Visakhapatnam (mm)

Year	June	July	August	September	Monsoon
2018	158.5	223.4	225.9	235.25	843.2
2019	123.4	174	175.9	183.2	656.7

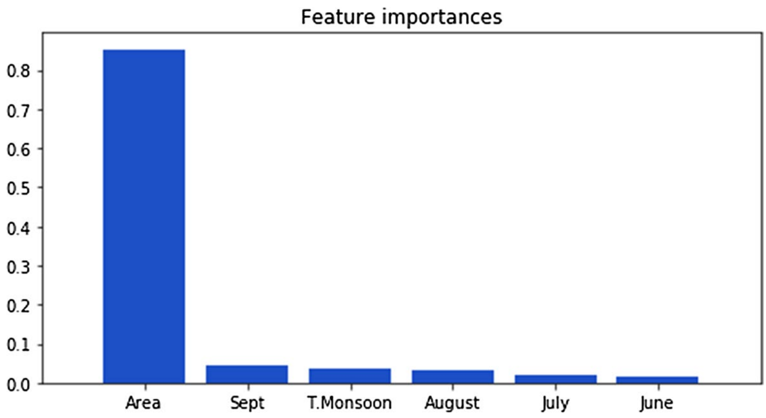
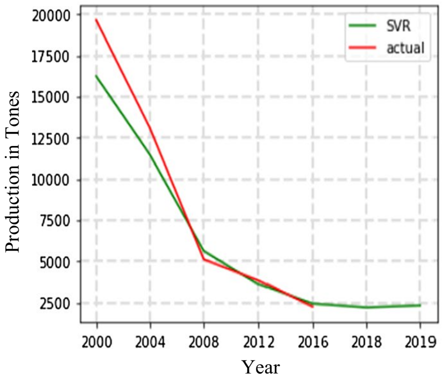


Fig. 8 Features that are most important in determining yield of bajra in Visakhapatnam

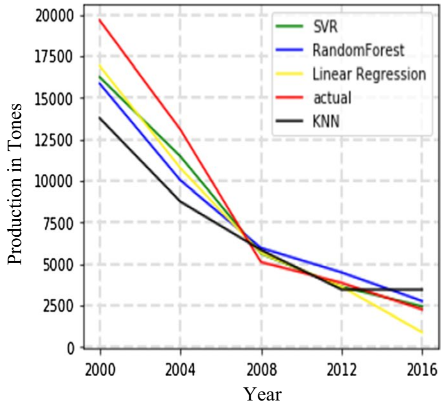
optimal performance with 1709 t root mean square error. Then, for predicting the yield of bajra in the years 2018 and 2019 we take mean of last 5 years' area which came out to be 2653 and we used that area for our prediction. Figure 9b shows the predicted production of bajra in Visakhapatnam by SVR in the years 2018 and 2019.

From Fig. 9, we also infer that there is a general decrease in production of bajra in Visakhapatnam. This is because of the increasing population in the city and less area being given to the bajra crop. The predictions of production of bajra made by our model is as follows:

- 2018 → 2199 t production → 828 kg/ha yield
- 2019 → 2320 t production → 874 kg/ha yield



(a) Performance of various algorithms



(b) Predicting yield of Bajra by SVR

Fig. 9 Predicting yield of bajra in Visakhapatnam

4.6 Predicting yield of maize in Visakhapatnam

To predict the yield of maize in Visakhapatnam, first, we apply feature selection method so that only important features can be selected in order to give the predictions. Figure 10 shows the important features in determining the yield of maize. From Fig. 10, we can infer that all the features are somewhat important in determining the yield in Visakhapatnam so we use all these features in our prediction for yield of maize.

Figure 11 shows the performances of the various regression algorithms used for making predictions for yield of maize in Visakhapatnam. From Fig. 11a, we can infer that SVR gave the most optimal performance with 1696 t root mean square error. Then, for predicting the yield of bajra in the years 2018 and 2019 we take mean of last 5 years' area which came out to be 5882 and we used that area for our predictions. Figure 11b shows the predicted production of maize in Visakhapatnam by SVR in the years 2018 and 2019.

From Fig. 11, we can infer that our model did not give accurate predictions; this is due to the randomness in data, for instance the year 2000 which was used as test data,

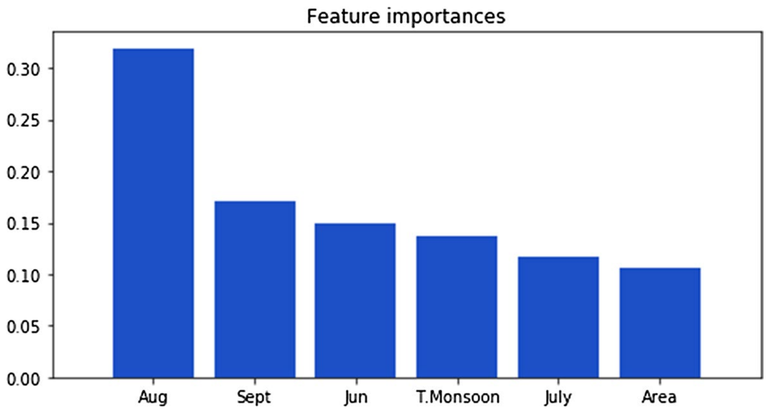
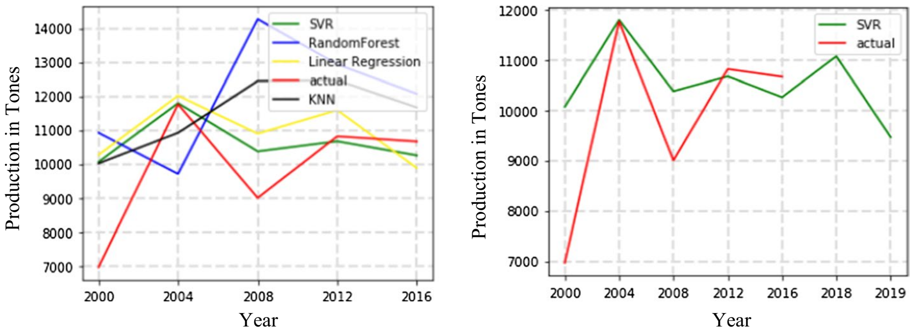


Fig. 10 Features that are most important in determining the yield of maize



(a) Performance of various algorithms

(b) Predicting yield of Bajra by SVR

Fig. 11 Predicting yield of bajra in Visakhapatnam

in that year rainfall in August and September was 311.5 and 102.9 and production was 6979 t, whereas in the year 2011 which was used for training data, the rainfall in August and September in that year was 215.8 and 129.7. We can see that both the data are comparable to each other, but in the year 2011 the production of maize was 25,112 t. This randomness in data causes errors in predicting. Nevertheless, our model gave the following predictions for production of maize in Visakhapatnam:

- 2018 → 11,083 t production → 1884 kg/ha yield
- 2019 → 9478 t production → 1611 kg/ha yield

4.7 Predicting yield of rice in Visakhapatnam

To predict the yield of rice in Visakhapatnam, first, we apply feature selection method so that only important features can be selected in order to give the predictions. Figure 12 shows the important features in determining the yield of rice. From Fig. 12, we can infer that area and rainfall in the months of September and June are the most important features in determining the yield of rice; hence, we use only these features in our prediction.

Figure 13 shows the performances of the various regression algorithms used for making predictions for production of rice. From Fig. 13a, we can infer that SVR gave the most optimal performance with 10,749 t root mean square error. Then, for predicting the yield of rice in the years 2018 and 2019 we take mean of last 5 years' area which came out to be 98,558 ha and we used that area for our predictions. Figure 13b shows the predicted production of rice in Visakhapatnam by SVR in the years 2018 and 2019.

From Fig. 13, we also infer that there is a decrease in production of rice in Visakhapatnam. This is because of the increasing population in the city and less area being given to the rice crop. Nevertheless, the predictions of production of rice made by our model are as follows:

- 2018 → 165,096 t production → 1675 kg/ha yield
- 2019 → 163,645 t production → 1660 kg/ha yield

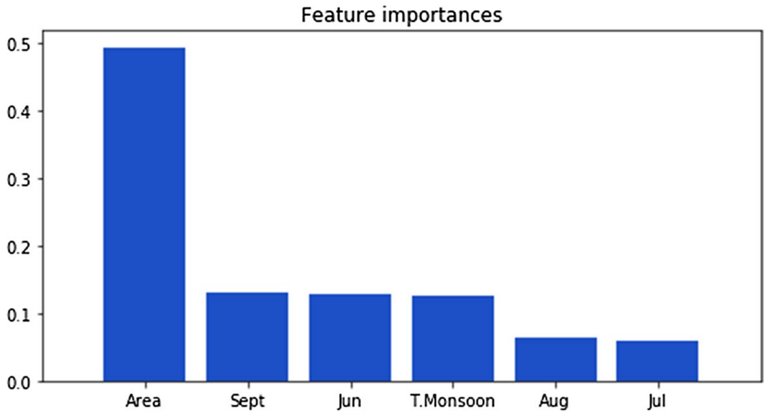


Fig. 12 Features that are most important in determining the yield of rice

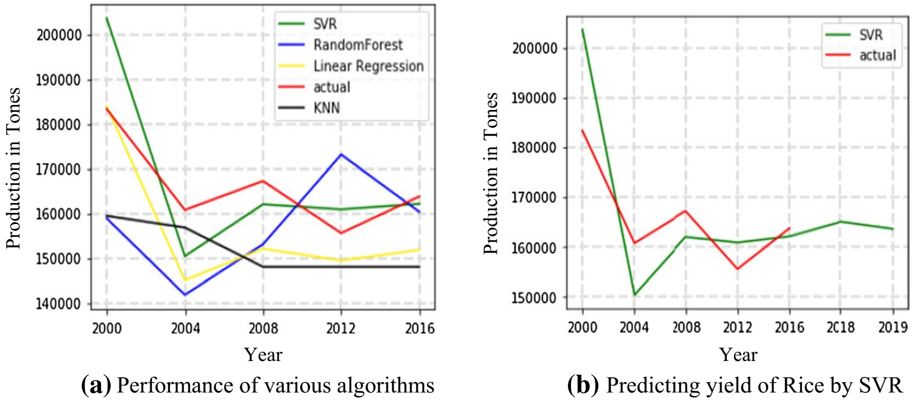


Fig. 13 Predicting yield of rice in Visakhapatnam

4.8 Predicting the yield of ragi in Visakhapatnam

To predict the yield of ragi in Visakhapatnam, first, we apply feature selection method so that only important features are selected in order to give the predictions. Figure 14 shows the important features in determining the yield of ragi. From Fig. 14, we can infer that area and rainfall in the months of June and September are the most important features in determining the yield of rice; hence, we use only these features in our prediction.

Figure 15 shows the performances of the various regression algorithms used for making predictions for production of ragi. From Fig. 15a, we can infer that SVR gave the most optimal performance with 1731 t root mean square error. Then, for predicting the yield of ragi in the years 2018 and 2019 we take mean of last 5 years' area which came out to be 22,709 ha and we used that area for our predictions. Figure 15b shows the predicted production of ragi in Visakhapatnam by SVR in the years 2018 and 2019.

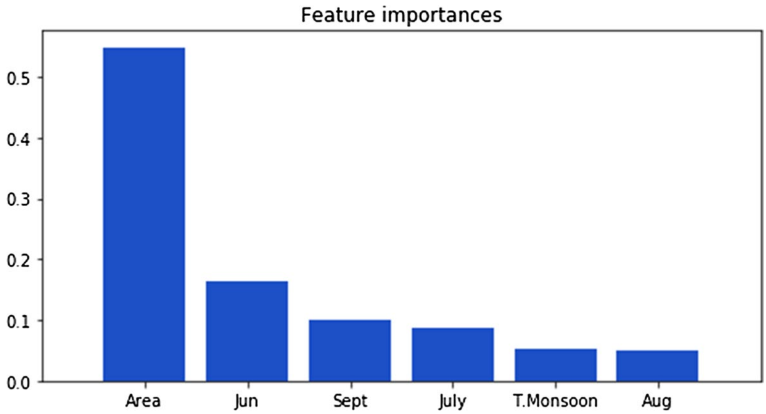
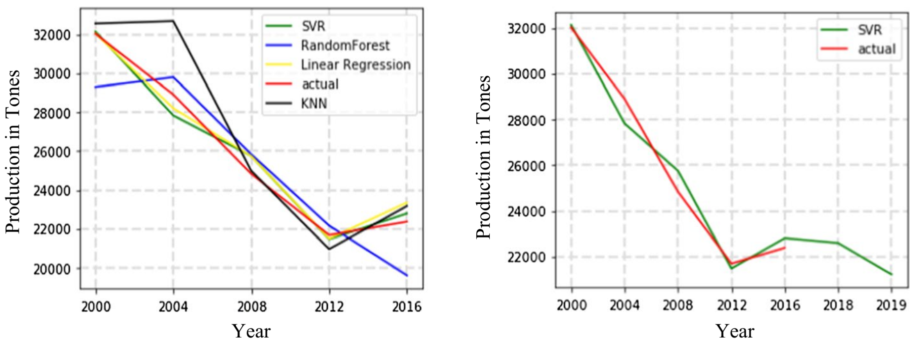


Fig. 14 Features that are most important in determining the yield of ragi



(a) Performance of various algorithms

(b) Predicting yield of Ragi by SVR

Fig. 15 Predicting yield of ragi in Visakhapatnam

From Fig. 15, we infer that there is a general decrease in production of ragi in Visakhapatnam; this can be because of the increasing population in the city and less area being given to the ragi crop. Nevertheless, the predictions of production of ragi made by our model are as follows:

- 2018 → 42,060 t production → 1852 kg/ha yield
- 2019 → 36,959 t production → 1627 kg/ha yield

Finally, from the predicted results, we infer that there is general decrease in production of kharif crops in Visakhapatnam from 1997 to 2017. This is because of the increasing population in the city and less area being given to the crops. Also, we have kept the area same for the years 2018 and 2019, we got the similar predictions, but as the rain predicted in 2018 is more than 2019, we can see the decrease in production of crops in the year 2019.

5 Conclusions and future scope

In the present study, we have applied modular artificial neural networks (MNNs) in order to predict the amount of rainfall that can occur in monsoon season in Visakhapatnam in the upcoming years. After predicting rainfall, we did feature selection in order to select only important features in predicting various kharif crops in Visakhapatnam. Finally, by using only those features, we predicted yield bajra, maize, rice and ragi. It can be seen from the results that in recent years less area is given to crops if we compare area given to crops in the year 1997; this is because of the increasing population in the city as area which was given to crops earlier has now been used as a residential area. In this study, we tried to predict the yield of various kharif crops in Visakhapatnam. As we have used only rainfall and area attribute in predicting the yield of crops, the yield of the certain crop depends on many other factors like fertilizers used, irrigation and many more. If those parameters are considered, we can increase the accuracy of our predictions. Hence, there is still substantial space for improving the predictions although by only considering the area and rainfall we got good results. Since the model is build using the climatic parameters, only the values of weather attributes like temperature and rainfall need to be passed, where the model can predict the correct output. Hence, the same working strategy can be extended to similar or different climates.

Funding Funding was provided by Science and Engineering Research Board (SERB) (Grant No. ECR/2017/001273), Govt. of India.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interests.

References

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews, 11*(10), 203–224.

Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Berlin: Springer.

Borlaug, N. E. (2002). The green revolution revisited and the road ahead. Stockholm: Nobelprize.org.

Bornn, L., & Zidek, J. V. (2012). Efficient stabilization of crop yield prediction in the Canadian Prairies. *Agricultural and Forest Meteorology, 152,* 223–232.

Campolo, M., Soldati, A., & Andreussi, P. (2003). Artificial neural network approach to flood forecasting in the River Arno. *Hydrological Sciences Journal, 48*(3), 381–398.

Cannas, B., Fanni, A., Pintus, M., & Sechi, G. M. (2002). Neural network models to forecast hydrological risk. In *Proceedings of the 2002 international joint conference on neural networks, 2002. IJCNN'02* (Vol. 1, pp. 423–426).

Cantelaube, P., & Terres, J. M. (2005). Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A, 57*(3), 476–487.

Cheng, B., & Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science, 9,* 2–30.

Curtin, R. R., Cline, J. R., Slagle, N. P., March, W. B., Ram, P., Mehta, N. A., et al. (2013). MLPACK: A scalable C++ machine learning library. *Journal of Machine Learning Research, 14*(Mar), 801–805.

Dabrowska-Zielinska, K., Kogan, F., Ciolkosz, A., Gruszczynska, M., & Kowalik, W. (2002). Modelling of crop growth conditions and crop yield in Poland using AVHRR-based indices. *International Journal of Remote Sensing, 23*(6), 1109–1123.

Dawson, C. W., & Wilby, R. L. (2001). Hydrological modelling using artificial neural networks. *Progress in Physical Geography, 25*(1), 80–108.

De Vos, N. J., & Rientjes, T. H. M. (2005). Constraints of artificial neural networks for rainfall-runoff modelling: Trade-offs in hydrological state representation and model evaluation. *Hydrology and Earth System Sciences Discussions, 2*(1), 365–415.

De Wit, A. D., & Van Diepen, C. A. (2007). Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts. *Agricultural and Forest Meteorology, 146*(1), 38–56.

Fortin, V., Ouarda, T. B., & Bobée, B. (1997). Comment on "The use of artificial neural networks for the prediction of water quality parameters" by HR Maier and GC Dandy. *Water Resources Research, 33*(10), 2423–2424.

Fraser, A. M., & Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A, 33*(2), 1134.

Gaddeyya, G., & Kumar, P. R. (2014). Studies on weed infestation of some agricultural fields at Visakhapatnam district, Andhra Pradesh. *Journal of Crop and Weed*, *10*(2), 419–429.

Giustolisi, O., & Savic, D. A. (2006). A symbolic data-driven technique based on evolutionary polynomial regression. *Journal of Hydroinformatics, 8*(3), 207–222.

Haboudane, D., Miller, J. R., Tremblay, N., Zarco-Tejada, P. J., & Dextraze, L. (2002). Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sensing of Environment, 81*(2), 416–426.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., et al. (2008). Big data: The future of biocuration. *Nature, 455*(7209), 47–50.

Ines, A. V., Das, N. N., Hansen, J. W., & Njoku, E. G. (2013). Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. *Remote Sensing of Environment, 138,* 149–164.

Juang, C. F., & Hsieh, C. D. (2012). A fuzzy system constructed by rule generation and iterative linear SVR for antecedent and consequent parameter optimization. *IEEE Transactions on Fuzzy Systems, 20*(2), 372–384.

Kang, Y., Khan, S., & Ma, X. (2009). Climate change impacts on crop yield, crop water productivity and food security—A review. *Progress in Natural Science, 19*(12), 1665–1674.

Kannan, E., & Sundaram, S. (2011). *Analysis of trends in India's Agricultural Growth*. Bangalore: Institute for Social and Economic Change.

Kennel, M. B., Brown, R., & Abarbanel, H. D. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A, 45*(6), 3403.

Kumar, A., & Bhattachrya, S. (2015). Crop yield prediction using Agro Algorithm in Hadoop. *International Journal of Computer Science and Information Technology & Security (IJCSITS), 5*(2), 271–274.

Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology, 150*(11), 1443–1452.

Lobell, D. B., & Field, C. B. (2007). Global scale climate–crop yield relationships and the impacts of recent warming. *Environmental Research Letters, 2*(1), 014002.

Maiti, S., & Agrawal, P. K. (2005). Environmental degradation in the context of growing urbanization: a focus on the metropolitan cities of India. *Journal of Human Ecology, 17*(4), 277–287.

May, R. J., Maier, H. R., Dandy, G. C., & Fernando, T. G. (2008). Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling and Software, 23*(10), 1312–1326.

McBratney, A., Whelan, B., Ancev, T., & Bouma, J. (2005). Future directions of precision agriculture. *Precision Agriculture, 6*(1), 7–23.

Mo, X., Liu, S., Lin, Z., Xu, Y., Xiang, Y., & McVicar, T. R. (2005). Prediction of crop yield, water consumption and water use efficiency with a SVAT-crop growth model using remotely sensed data on the North China Plain. *Ecological Modelling, 183*(2), 301–322.

Moriondo, M., Giannakopoulos, C., & Bindi, M. (2011). Climate change impact assessment: The role of climate extremes in crop yield simulation. *Climatic Change, 104*(3–4), 679–701.

Morshed, A., Dutta, R., & Aryal, J. (2013, April). Recommending environmental knowledge as linked open data cloud using semantic machine learning. In *2013 IEEE 29th international conference on data engineering workshops (ICDEW)* (pp. 27–28).

Oettli, P., Sultan, B., Baron, C., & Vrac, M. (2011). Are regional climate models relevant for crop yield prediction in West Africa? *Environmental Research Letters, 6*(1), 014008.

Panda, S. S., Ames, D. P., & Panigrahi, S. (2010). Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing, 2*(3), 673–696.

Parent, B., & Tardieu, F. (2014). Can current crop models be used in the phenotyping era for predicting the genetic variability of yield of plants subjected to drought or high temperature? *Journal of Experimental Botany, 65*(21), 6179–6189.

Rajurkar, M. P., Kothyari, U. C., & Chaube, U. C. (2002). Artificial neural networks for daily rainfall-runoff modelling. *Hydrological Sciences Journal, 47*(6), 865–877.

Ramachandran, V. K., Rawal, V., & Swaminathan, M. (2010). *Socio-economic surveys of three villages in Andhra Pradesh: A study of Agrarian relations*. New Delhi: Tulika Books.

Rosenzweig, C. E., Antle, J., & Elliott, J. (2015). *Assessing impacts ofclimate change on food security worldwide*. Eos97EO047387.

Salehnia, N., Hosseini, F., Farid, A., Kolsoumi, S., Zarrin, A., & Hasheminia, M. (2019). Comparing the performance of dynamical and statistical downscaling on historical run precipitation data over a semi-arid region. In *Asia-Pacific Journal of Atmospheric Sciences*, pp. 1–13.

Sudheer, K. P., Gosain, A. K., & Ramasastri, K. S. (2002). A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrological Processes, 16*(6), 1325–1330.

Sujatha, R., & Isakki, P. (2016, January). A study on crop yield forecasting using classification techniques. In *International conference on computing technologies and intelligent data engineering (ICCTIDE)* (pp. 1–4).

Ullah, A., Salehnia, N., Kolsoumi, S., Ahmad, A., & Khaliq, T. (2018). Prediction of effective climate change indicators using statistical downscaling approach and impact assessment on pearl millet (*Pennisetum glaucum* L.) yield through genetic algorithm in Punjab, Pakistan. *Ecological Indicators, 90,* 569–576.

Wang, W., Van Gelder, P. H., Vrijling, J. K., & Ma, J. (2006). Forecasting daily streamflow using hybrid ANN models. *Journal of Hydrology, 324*(1), 383–399.

Wu, C. L., Chau, K. W., & Fan, C. (2010). Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. *Journal of Hydrology, 389*(1), 146–167.