Prediction of Agriculture Yieldsusing Machine Learning Algorithm

Jagannath Ray[1], Rajesh Kumar Ojha[2], Surabika Hota[2]
[1]Associate Professor, [2]Assistant Professor, [1,2]Dept. of CSE
[1,2]Gandhi Institute for Technology, Bhubaneshwar, India

Abstract: In recent years, great efforts have been carried out on the challenging task of predicting different crop yields. Developing exact models for crop yield estimation utilizing Information and Communication Technologies may support farmers and different stakeholders to improve decision making about national food import/export and food security. Most of the crops are selected based on the economic range. In our proposed work also we have consider the economical crops and they provide better prediction compared with the existing classifiers. The proposed ensemble classifier provides an efficient crop yield and crop disease forecasting model. Our proposed work provides knowledge to the farmers about the climatic conditions of the probability of crop disease and the climatic conditions for better crop yield. Even it discovers the crop yield and crop diseases, but does not concentrate on the solution to solve the productivity issue caused by crop diseases. Further, our future work concentrates on the above issue with different algorithms.

Keywords: Agriculture, crop prediction, regression, random forest algorithm.

1. INTRODUCTION

According to the records of the previous year 2018 and 2019, there are approx. 145 million landholdings in India. We may assume that India has about 130 million farmers. In a country like India which has increased demand for food due to the increasing population in the country. The most disadvantaged situation is that farmers who have access to irrigation are better placed but those who are in rain-fed and drone-prone areas are most vulnerable. A single crop failure due to flood, lack of soil fertility, drought, climatic changes, lack of underground water and some other factors may destroy the crop and this affects the farmers. There is no commodity-based farming in India till now. While in other countries the organizations advise farmers to grow specific crops according to the locality of the area and some other factors. Every farmer who produces the crop always tries to know how much yield will get from his expectations. So we want to help farmers by creating a machine learning model that predicts the crop yield. Although there are models that help in yielding they are hardware-based which is expensive and difficult to maintain. We can also increase the yield of the crop by systematic study of different methods like planting, fertilization,

irrigation and some other methods which help to optimize the production of crop but most of the farmers in India are illiterate to study them and follow the optimizing methods and follow them. So we came across an idea to implement a machine learning model which calculates the history of the field and suggest whether the crop should be planted in that area or not, basically it benefits the farmers and saving them lot of trouble. Although our proposed system is limited only to a certain crop which is Rice, cause most of the farmers in India rely on farming Rice.

The proposed system predicts the crop yield accurately using SVM to produce accurate results and helps farmer to choose the right crop according to the area and climatic conditions because in prediction process of the system we include the data of soil nitrogen, underground water, temperature, rainfall which may produce the accurate results in recommending farmer to invest in farming that crop or not.

2. LITERATURE SURVEY

There are a variety of approaches to the classification of land use efficiency indicators in agriculture, and a number of expert scholars have conducted their research on this, including I.A. Artamonova thinks about the plowing of agricultural lands, the relative indicators of the total area of fertilized land, the total area of agricultural lands. In addition, this author proposes to take into account the organizational and legal form of land management [2]. In this regard, in our opinion, the author has taken into account the fact that individual indicators of economic efficiency of land use in the personal subsidiary farms of peasants and citizens cannot be used.

Bhanu.K.N'et al., 2020, proposed a paper related to the advancements of smart agricultural crop monitoring system with respect to Internet of Things (IoT) [1]. In this paper [1], the authors illustrated the interconnectivity benefits of multiple sensors associated on the smart device with the remote server, in which the remote server connectivity enables the agricultural people to monitor the status of the crops on the field instantly without any delay at anytime from anywhere in the globe. This paper describes the power of Internet of Things association in agricultural field as well as this kind of approaches provides better yielding in crop maintenance and modernize the field and nature of agriculture in good way. In this paper [1], the authors considers the major factors such as sunlight, temperature control, moisture level of the soil, required protein supplements to the soil, climate conditions and so on. Multiple category of sensors are associated over the smart device and provide the proper

communication services with the help of Internet of Things and enable the farmers to monitor the crops without any hurdle. The major advantage noticed in the paper [1] is the systematic monitoring and accumulating the modern technologies over the approach to provide support to the farmers in an efficient way. However, the limitations need to resolve in this paper is the remote monitoring of crops in visualized way as well as the missing of Artificial Intelligence association on the proposed approach. As well as the external third party support is accumulated for remote server maintenance such as Think Speak and so on, which will be not possible to customize easily on client end and complex in working.

"An important condition for the development of agricultural production based on land use is the support of the state, an important element of which is preferential taxation," said O.Ya. Starkova [3].

In the efficient use of land, O.D. According to Ermolaenko, the large area of land and the lack of transport capacity, in turn, necessitate the development of infrastructure, which is impossible without state participation [4]. It is expedient for the state to act as the main reformer in the development of the agricultural economy. This is because the state supports the reclamation of lands and the application of fertilizers, thereby contributing to maintaining the quality of the land.

Yu.D. According to Bakhteev and ZA Kudyusheva, the profitability of agricultural enterprises depends on the productivity of arable land [5]. T.G. Khanbaev and L.S. Daibova noted that "the overall indicator of land use efficiency is the production of comparable products for 100 hectares of agricultural land or arable land and sold at current prices" [6].

Currently, the N.A. on the topic of rational use of land resources, identification of problems in their implementation, development of guidelines and implementation of measures to improve the economic use of land resources. Frieva conducted scientific research [7].

In the research of Vita Cintina and Vivita Pukitit, land use efficiency is based on agricultural production, and through proper and efficient use of land, it is possible to solve several problems such as food production, welfare and social sustainability [8].

A.L.Zheliaskov and N.S.Denisova conducted research on the optimal volume of agricultural land use, the concept of optimal land use, the rational size of agricultural land and the impact of the organization of the municipality on the optimal use of land [9].

In general, the study conducted economic analysis of factors and developed recommendations and recommendations on the results, but did not talk about land reclamation and their impact

on crop types, correlation-regression analysis of changes in efficiency and future prospects. This, in turn, requires more in-depth research and studies in this area.

3. PROPOSED METHOD

The proposed model will mainly focus on crop production based on four factors and one Machine Learning algorithm called SVM (support vector machine). SVM is used to classify whether rice can grow in that area based on the data from soil, temperature, underground water and rainfall. And also implementing a web application with HTML, CSS, and JavaScript. The web application can be used to interact with the Machine Learning model and by providing inputs we can get the prediction output. The application also uses the weather API from yahoo to get the current temp at that location, which will be one of the factors.

Support Vector Machine

SVM is a machine learning algorithm that comes under the supervised category and is used for binary classifications problems. The objective of this algorithm is to plot a hyper plane in an N-dimensional space, where N is the number of features that are going to be in a dataset, that distinctly classify the data points.

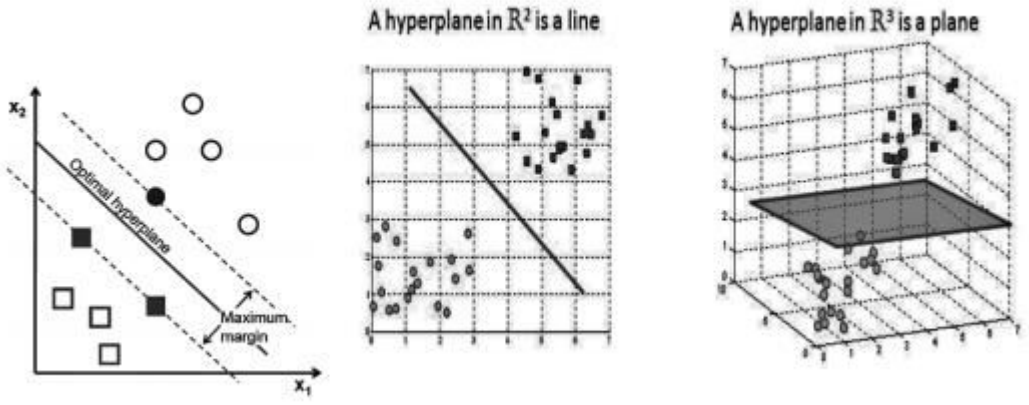

Figure 1: Maximum Margin and Hyper planes

There can be any number of hyper planes plotted, but the algorithm's main target is finding the plane that has a Maximum Margin i.e the maximum distance between data points of the features being plotted. The more the distance more accurate will be the classification. As shown in fig 1, that the data points are far from all the other points from the Optimal hyper plane, making it as a Maximum margin.

Cost Function: The main objective is to maximize the margin. So hinge loss is used to do that, the cost is zero if the predicted value and actual value are of the same sign, if they are not we can calculate hinge loss value. And adding a regularization parameter will balance the margin maximization and loss.

$$J(\theta) = C[\sum_{i=1}^{m} y^{(i)}\, Cost_1(\theta^T\,(x^{(i)}) + (1 - y^{(i)}) Cost_0(\theta^T(x^{(i)})] + \frac{1}{2}\sum_{j=1}^{n} \theta_j^2$$

In this proposed model we used Linear SVM which suited our kind of prediction. In Linear SVM the loss function is as similar to that of Logistic Regression. The x-axis here is the output i.e θT x. Just like the Sigmoid function, the hypothesis used here is when θT x >= 0, predicts 1, otherwise, predicts 0.

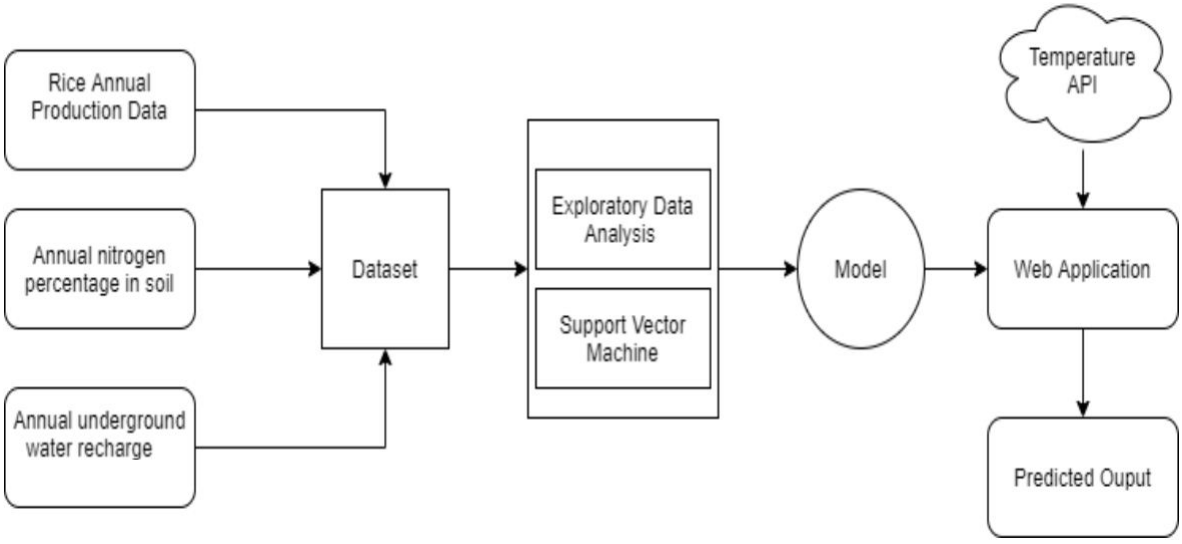

Figure 3: Architecture Diagram

Exploratory Data Analysis: The important part before building a model is to analyze the data first and extract the features which are causing the output target variable. The steps include are filling null values, dropping the features which are not necessary, visualizing the data, normalization.

Splitting data into Training and Test: The dataset is split into two Training and test. We can also select the proposition of their division metric. In this model, the training set is 70% of the dataset and 30% is the test set. The training of a model also depends on this proportion as more training of data more chances of better accuracy.

Training the Model: The next step is to train the model using our preferred algorithm. We choose trial and error method and wanted to select the best algorithm which gives us better accuracy, first we selected Non-Linear SVM which got us the precision of 0.62 i.e 62% accuracy, which is very low and not suitable for our dataset. Then we choose Linear-SVM which got us the precision value of 0.93 i.e 93% accuracy.

Parameter Tuning: To increase the accuracy or the precision score we can tune the parameters which define the training model. We got 93% of accuracy with Linear-SVM with

a cross-validation score of 10 (k=10). So by changing the value of the cross-validation score to 5 (k=5) got us better results with the accuracy percentage of 96.3.

K-fold cross-validation: It is a type resampling technique used to evaluate or estimate the Machine Learning model skill or accuracy on unseen data. It divides the data into k groups and trains each group separately and the precision score from all the groups are averaged to the final precision or accuracy value.

Prediction: From the finalized model we can start predicting our values. We need to provide five parameters, rainfall in mm, average underground water recharge, nitrogen in soil, area in sq.ft and temperature. The result will be of two types one is 0 and the other is 1. If 0 is the output then it is not recommended to grow rice in that area. If the output is 1 then it is recommended to grow rice in that area.

4. EXPERIMENTAL RESULTS

4.1 Data Set:A proper dataset is required for proper training of a model. There are four data factors used in this proposal which are Nitrogen percentage in soil, the annual rainfall in mm, the annual underground water recharge and the annual temperature. The dataset is focused on the Indian State Tamil Nadu, which has over 32 districts. So we have collected the data from the year 2010 to 2020 which contains all the annual values of the data of the factors mentioned above. The rainfall data is obtained from the Indian Gov website, the underground data and the soil nitrogen percentage data is obtained from self-research and the remaining data is obtained from Kaggle which includes the annual production value of the wheat crop from the year 2010 to 2020 and the area in sq. ft. The target variable in this dataset is “output” which has two values which are 0 and 1. 0 represents whether rice farming in that area is suitable or not and vice versa. This dataset contains a total of 12 features and 500 observations.

To maximize the harvest yield, the determination of the suitable crop that will be planted is an essential job. It relies upon different variables like the kind of soil and its organization, atmosphere, the geography of the area, crop yield, market costs and so forth. In our implementation of proposed work, we consider the agriculture information for the crop yield anticipating the parameters, for example, state, year, precipitation, temperature, humidity, season, harvested region, production, pesticide use, since all the above parameters additionally assume a significant job in crop production.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	District	Rainfall (year) (mm)										Temperature (year) (Celsius)						
2		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2010	2011	2012	2013	2014	2015	2016
3	Kancheepuram	482.1	490.8	334.7	641.8	16.5	29.1	14.3	86	847.6	1227.7	39.6	36.9	24.6	22.8	31.4	30	19.6
4	Cuddalore	346	383.1	130.6	697.8	114.3	44.1	23.4	81.7	614.1	1206.7	41	38.2	23.9	26.6	31.3	30.2	20.3
5	Salem	346	448.6	180.3	370.3	12.2	16	167.9	170.8	628.4	997.9	37.6	35.4	23.1	26.9	33.1	31.1	20
6	Namakkal	239.8	338.3	57.5	291.6	8.3	13.9	141.7	148.6	445.3	793.4	38.9	34.9	23.5	27.4	35.1	33.1	22
7	Dharmapuri	289.2	393.4	99.4	330.1	8.4	18.2	240.9	160.4	617.9	902.1	35.7	33.2	24.3	25.4	39.7	35.4	26.3
8	Krishnagiri	326.8	399	120.1	289.4	5.7	10.7	283.2	151.6	737.8	850.7	32.6	29.8	22.4	24.5	31.4	30	19.6
9	Coimbatore	222.6	188.8	142	328.9	18.4	20.5	178.5	150.3	545.5	689.3	32.4	30.1	21.5	23.5	31.5	30.2	20.3
10	Erode	174.4	229.8	74.9	314.6	10.8	16.1	258.4	142.4	518.5	792.9	31.3	30.4	20.4	21.7	33.1	31.1	20
11	Karur	125	213.6	69.5	314.7	17.2	17.5	127.8	109.2	339.3	655	33.4	32.1	21	23.9	35.1	33.1	22
12	Perambalur	270.9	290.7	127.2	440.9	24.3	21.4	122.2	108.9	544.6	861.9	36.1	33.6	24.1	26	39.7	35.4	26.3
13	Pudukkottai	233.8	350.6	159.5	406.2	55.2	33.1	54.4	97.5	502.9	887.4	40.8	35.6	26	28.5	42.6	37.6	27.5
14	Thanjavur	288.6	318.4	210.3	550.3	103.5	42.3	68.4	102.1	662.8	1013.1	40.8	36.6	25	28.9	40.2	36.8	23.8
15	Nagapattinam	245.2	286.1	248.6	941	146.3	85.7	51.4	88.5	691.8	1393.3	39.1	35.5	23.7	27.5	40.4	37.9	23
16	Madurai	211.7	335.9	228.2	419.1	33.2	28.1	177.3	144.8	658.4	927.9	37.6	35.1	24.4	27.4	37.3	35.1	21.6
17	Dindigul	165.6	295.4	227.8	436.4	38.9	30.9	151.3	168	583.6	930.7	38.2	36.1	24.9	27.8	37.1	35.2	22.2
18	Ramanathapuram	100.7	148.3	191.3	491.7	67	51.3	69.3	115.5	428.3	807.8	36.6	34.1	23.6	26.6	36	33.3	22
19	Sivagangai	180.1	301	159.1	422.7	74.6	27.9	107.3	121.2	725.1	872.8	36.9	34.5	23.5	26	32.5	29.7	21.2
20	Ariyalur	318.8	382	128.5	545.5	108.3	32.5	43.7	101.8	598.3	1071.8	34.7	31.7	20.4	23.6	32.9	30	19.5

Figure 2: Sample data set based on the season and crop yield.

Above Fig. 2 shows a sample data set utilized for our implementation with the normal rainfall and temperature of each area for as long as ten years extending from 2010 to 2020. The variation in temperature and rainfall not only affects the productivity.

4.2 Results for Training and Testing Dataset of our Proposed Methodology

Here utilized 70% of information for training and 30% of information for the testing stage. Around the pieces of information from the year 2010 to 2016 and 2017 to 2019 for the training and testing. During training, a pattern comprising of known sources of input and yields is presented to the system. The input sources are taken care of through the system and yield is determined. The error is determined and the loads between the hidden layer and yield layer are balanced. Likewise, the loads between the input layer andhidden layers are additionally balanced. This procedureis repeated for all samples in the training set. The patterns are persistently presented and loads are balanced until the error is adequately low. When the preparation was finished for each dataset, the testing set was presented to the proposed system to anticipate the yield productivity and diseases for the 3 years of the test set.

4.3 Predicted outcome using the proposed methodology

Our proposed works consider the issue of predicting the average yield of a type of harvest for a region of concern dependent on historical information. In particular, our proposed work is concentrating on the average yield per unit zone in a given geographical region, e.g., Country or district. Our proposed structure forecasts the crop productivity for the various regions of Tamil Nadu. Below table 1 shows the sample predicted result of our proposed structure. This fundamentally concentrates on the rainfall and temperature of each area. Hotter temperatures expected due to environmental change and the potential for progressively extraordinary temperature occasions will affect plant productivity. Fertilization is one of the extremely

penetrating phonological stages to temperature extremes across all species and during this developmental stage temperature extremes would greatly affect production. It additionally decreased the harvest yield by as much as 80 to 90% from a typical temperature system.

Table 1 Forecasted crop production by our proposed framework

District	Rainfall Prediction (mm)	Temperature Prediction (Celsius)	Crop	Predicted Production
Karimnagar	847.6	38.4	Coconut	5.13E+07
Hyderabad	614.1	41.8	Sugarcane	2.29E+04
Warangal	626.4	41.2	Guar seed	1.94E+06
Vijaywada	445.3	39.4	Pulses	1.79E+09
Tirupathi	617.9	37.7	Jowar	1.67E+05
Kadapa	737.8	38.4	Carrot	1.26E+08
Kurnool	545.5	36.5	Potato	4.33E+05
Vizag	518.5	36.8	Rice	2.29E+02
Amaravathi	339.3	34.8	Wheat	3.34E+07
Nizamabad	544.6	31.3	Onion	4.79E+03
Krishna	502.9	32	Coriander	3.57E+05
guntur	662.8	22.5	Maize	1.54E+09

4.4 Comparison Analysis of Proposed and Existing Work

Different data mining strategies are executed on the input information to evaluate the best performance yielding strategy. The current work utilized data mining procedures to get the optimal climate requirement like the optimal scope of temperature and rainfall to accomplish higher production of yields. Clustering techniques are analyzed utilizing quality metrics. In our research work, the performance comparison is done with KNN, Decision Tree, Naïve Bayes and Adaboost with our proposed SVM ensemble classifier based on its, accuracy, recall and f-measure.

Table 2: Performance analysis, comparison with existing methodologies

Name of the Classifier	Accuracy	Precision	Recall	F-measure	MSE
SVM	81.62	67.3	86.03	72.1801	0.2997
Decision Tree	85.66	77	89.49	81.1452	0.3775
Naïve Bayes	90.53	94.31	92.16	80.0425	0.1449

Adaboost	94.69	94.18	66.12	75.0446	0.2072
Proposed SVM	95.48	95.16	93.33	87.71	0.105

Table 2 demonstrates the performance of various classifiers along with the proposed forecasting model. The accuracy of SVM is 95. 48%, which is higher than the other classifiers. The precision rate of a Naive Bayes classifier is 94.31% and our proposed framework achieves 95.16%, which is higher than the other classifiers. The recall rate for Naïve Bayes is 92.16%, which is also higher than the other classifiers and our proposed ensemble learner attains the value of 93.33%. F-measure and Mean Square Error also relatively provide 87.71% and 0.105. From this analysis, it is clear that both Naïve Bayes and Adaboost outperforms the other classifiers, and also our proposed ensemble classifier (SVM) provides better performance than the other classifiers.

5. CONCLUSION

There is so much to explore in machine learning yet, as there can be new algorithms, new techniques in the future. Our paper is a simple crop prediction recommendation systemwhich is only limited to one state Tamil Nadu, as we hope to do more papers on other Indian states and encourage other fellow researchers to also persue research in the agriculture field, as this is our main source of food all over India. It alone contributes 60% of the entire GDP. But since 2018 it is gradually decreasing, the per capita water availability is also decreasing, which will result in a lot of crop production failures. And also there are multiple numbers of suicides of farmers all over India, who just work very hard and don't get the expected results due to many factors. This paper is a small contribution to the agriculture field and dedicated to all the farmers, to help them in their farming, so that they can get profits and benefits of the new technologies which they don't have any idea of. So finally we want to conclude that as an Engineer we should take responsibility and contribute our knowledge to the betterment of our society or country

REFERENCES

[1] Avinash Kumar, Sobhangi Sarkar & Chittaranjan Pradhan 2019, 'Recommendation System for Crop Identification and Pest Control Technique in Agriculture', IEEE International Conference on Communication and Signal Processing, vol. 37, pp. 0185-0189, Apr 2019.

[2] Christopher Brewster, Ioanna Roussaki, Keith Ellis, Kevin Doolin & Nikos Kalatzis, 'IoT in Agriculture: Designing a Europe-Wide Large-Scale Pilot', IEEE Communication Magazine, vol. 22, issue 7, Sept 2017.

[3] David G Michelson, Maziyar Hamdi & Pooyan Abouzar, 'RSSI-Based Distributed SelfLocalization for Wireless Sensor Network Used in Precision Agriculture', IEEE Transactions on Wireless Communication, vol. 15, issue 10, pp. 125-131, Oct 2016.

[4] Dutta, Ritaban, Morshed, Ahsan, Aryal, Jagannath, D'Este, Claire, Das & Aruneema 2016, 'Development of an intelligent environmental knowledge system for sustainable agricultural decision support', Research Gate, Environmental modelling & software 2014, vol.52, pp. 264- 272.

[5] Fan-Hsun Tseng, Hsin-Hung Cho & Hsin-Te Wu, 'Applying Big Data for Intelligent Agriculture-Based Crop Selection Analysis', IEEE Access, vol. 7, 2019.

[6] Federico Viani, Michael Bertolli, Marco Salucci & Alessandro Polo, 'Lowcost wireless monitoring and decision support for water saving in agriculture', IEEE Sensors Journal, vol. 99, pp. 1–1, May 2017.

[7] Fransisco Yandun Narvaez, Giulio Reina & Miguel Torres 2017, 'A Survey of Ranging and Imaging Techniques for Precision Agriculture Phenotyping', IEEE/ASME Transactions on Mechatronics, vol. 22, issue 6, pp. 2428-2439, Oct 2017.

[8] Giritharan Ravichandran, & R S Koteeshwari , 'Agricultural Crop Predictorand Advisor using ANN for Smartphones', IEEE International Conference on Emerging Trends in Engineering, Technology and Science, vol. 45, pp. 138-145, Oct 2016.

[9] Ihsan Ali, Muhammad Zakarya & Rahmin Khan, 'Technology-Assisted Decision Support System for Efficient Water Utilization: A Real-Time Testbed for Irrigation Using Wireless Sensor Networks', IEEE Access, vol. 6, no. 6, pp. 2342-2350, May 2018.

[10] Johan.Estrada-Lopez, AlejandroA. Castillo-Atoche, Javier Vazquez-Castillo & Edgar Sanchez-Sinencio,'Smart Soil Parameters Estimation System Using an Autonomous WirelessSensor Network with Dynamic Power Management Strategy', IEEE Sensors Journal, vol.18, no.21, pp. 8913–8923, Nov 2018.

[11] Narongsak Lekbangpong, Jirapond Muangprathub, Theera Srisawat & Apirat Wanichsombat, 'Precise Automation and Analysis of Environmental Factor Effecting on Growth of St. John's Wort', IEEE Access, vol .7, pp.112848 - 112858, Aug 2019.

[12] Nurzaman Ahmed, Debashis De & Hussain, Md. Iftekhar Hussain, 'Internet of things (iot) for smart precision agriculture and farming in rural areas', IEEE Internet of Things Journal, vol. 5, no. 6, pp. 4890–4899, Dec 2018.

[13] Rakesh Kumar, M.P.Singh, Prabhat Kumar & J.P.Singh, 'Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique', IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, vol. 25, pp. 138- 145,May 2015.

[14] S.Pudumalar, E.Ramanujam, R.Harine Rajashree, C.Kavya, T.Kiruthika & J.Nisha, 'Crop Recommendation System for Precision Agriculture', IEEE International Conference on Advanced Computing, pp. 645-650.,June 2017.

[15] Sk Al Zaminur Rahman, Kaushik Chandra Mitra, S.M. Mohidul Islam "Soil Classification using Machine Learning Methods and Crop Suggestion Based on Soil Series", 2018 IEEE International Conference of Computer and Information Technology (ICCIT).

[16] S. K. S. Raja, R. Rishi, E. Sundaresan and V. Srijit, "Demand based crop recommender system for farmers," 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), Chennai, 2017, pp. 194-199, doi: 10.1109/TIAR.2017.8273714.

[17] Suyash S. Patil, Sandeep Thorat , "Early Detection of Grapes Diseases Using Machine Learning and IoT", International Conference on Cognitive Computing and Information Processing (CCIP),Jan 2017.