Crop Yield Estimator for Smart Agriculture using Machine Learning

Hemamalini S
Dept. of Computer Science and Engineering
Panimalar Engineering College
Chennai, India

Suraksha. M
Dept. Of Computer Science and Engineering
Panimalar Institute of Technology,
Chennai, India

Sreya. S
Dept. Of Computer Science and Engineering
Paniamalar Institute of Technology, Chennai, India

Sowmiya. R
Student, Dept. Of Computer Science and Engineering
Panimalar Institute of Technology, Chennai, India

***Abstract*— Agriculture practices will have a significant impact on the well-being of Indian citizens in the future. A sustainable food supply can be ensured using machine learning, and farmers can stay informed about the most recent weather patterns. In a recent study, researchers looked into how to discover agricultural yield prediction models using artificial intelligence approaches. Because of the nonlinear relationship between the input and output variables, the system of supervised learning, widely used to evaluate fruits, was ineffective. Even though these methods improve the accuracy of crop yield prediction, they are expensive. The application of ML approaches to predict crop productivity has been discovered in an increasing number of studies. This paper provides a thorough analysis of the precision of various ML models to investigate their efficacy. This study intends to investigate how machine learning can be used to forecast crop yield. Various machine learning (ML) algorithms such as Regression, Decision trees, etc are evaluated based on their performance.**

Keywords*— *Crop yield prediction, Machine learning (ML), Agriculture, Long Short-Term Memory Networks (LSTM), Random Forest, Stochastic Gradient Descent (SGD,) Decision Tree, Gradient Boosting, K-nearest Neighbour (KNN).

I. INTRODUCTION

Agriculture is an important human activity since it supplies fundamental necessities such as food, clothes, and shelter. Agriculture and related industries such as forestry and fisheries contribute 15.4 percent of GDP (gross domestic product) and employ approximately 31 percent of the labor force[1]. With the increasing population specifically in Asian countries, there is a strong demand for agricultural products. With such a large number of farmers and rising suicide rates, we need to help farmers comprehend the necessity of crop prediction and to expand their basic knowledge of soil quality and location-wise meteorological limits to produce crop yields using a technology solution [2]. Sadly, an increasing frequency of adverse weather incidents influences agricultural productivity. Therefore, the availability of water and elements like temperature and air pollution has a big impact on agriculture. A single crop failure caused by a flood, a lack of soil fertility, a drought, climate change, a lack of subsurface water, or other reasons might kill the crop, affecting farmers. Currently, there is no commodity farming in India [3]. The effects of global warming are worsening the situation for farmers, whose crops are routinely ravaged by unfavorable weather conditions [4]. The biggest problem confronting agriculture today is the rising demand for food. It is crucial to have a prediction model to anticipate crop production since there are high demands to satisfy agricultural demand, maintain livelihoods, and ensure economic progress. Farmers and policymakers need reliable crop yield estimations to prepare and handle these issues. Over the last few years, there has been a lot of analysis on predicting crop estimation using machine learning methods. ML models can learn patterns from historical data and make predictions on new data, providing a valuable tool for decision-making in agriculture. Agriculture must be monitored and optimized to support a country's food security and economic prosperity [5]. Applying more fertilizer than necessary causes hazardous ailments in crops with increased fertilizer content [6]. Farmers can improve agricultural yield output rates when there are unfavorable environmental circumstances by employing machine learning algorithms. They help harvesters minimize losses by identifying the crops that will yield the greatest in a specific climate. A model makes use of Support Vector Machines (SVMs) to predict crop yield depending on the shape, texture, and color of patterns on the infected surface [7]. Another study uses neural network models like Feedforward neural networks and Recurrent neural networks [8]. The study proposed by Tiwari and Shukla used Convolutional neural networks (CNNs) to lower crop yield prediction error and the relative error rate [9]. Machine learning techniques have been used in agriculture to enhance crop disease prediction, smart irrigation systems, yield prediction, and crop selection approaches. Farmers will gain from these ML techniques because they will increase productivity while requiring less input. Furthermore, improvements in instruments and technology should be accurate as they make predictions and decisions based on a greater amount of data. This research work examines the benefits and drawbacks of various ML-based agricultural techniques. The findings of this study will contribute to the growing body of research on the use of ML in agriculture and provide insights into the potential of ML models to improve crop yield prediction. Ultimately, the goal is to develop accurate and reliable crop yield prediction models that can help farmers and policymakers make informed decisions to improve agricultural productivity and food security.

II. LITERATURE REVIEW

Tiwari and Shukla [9] utilized CNN to forecast crop yield. The difficulty with the current paradigm was that agricultural drifts for crop growth were constantly breaking down since they were not ideal for environmental elements including weather, soil quality, and temperature. The model decreased crop yield prediction accuracy while also reducing relative error.

Shreya et al. [10] utilized K-means Clustering and Naïve Bayes algorithm for predicting agricultural yield estimation. K-means clustering constructed clusters. Based on the clustered hypothesis, data kept in clusters enabled quick searches in minimal time, which aided the farmers in predicting the yield. Predictions are, however, unachievable since this model gives a zero probability to each category of a variable if it is absent from the training data set.

Shruti Kulkarni et al. [11] utilized Neural Network. This constructed model considered the most relevant environmental factors and soil characteristics that impact crop development, giving each factor equal weight in the outcome prediction. This model produced error rates, and it is data-centric. Its precision depends on the dataset. The larger the data, the better accuracy. The latest datasets help predict real-time fluxes in soil and climate conditions.

Krutika Hampannavar et al. [12] developed a model using Support Vector Machine (SVM) to forecast the quantity of fertilizer intake together with histogram analysis. Histogram analysis is used for identifying the lack of nitrogen and the quantity of nitrogen fertilizer consumption. There is no probabilistic justification for the classification because the SVM simply positions data points above and below the hyperplane of classification.

Tanhim et al. [13] utilized Deep Neural Network for the crop selection and estimation of agricultural production. The study considered the training set as 80% and the testing set as 20% and showed the accuracy and Mean Square Error (MSE) of each model. Despite showing better accuracy, farmers were trailing behind in their agricultural and irrigation processes.

Suresh et al. [14] utilized K-Means and Modified KNN. This study considered factors like area, groundwater, and rainfall for the analysis and tested the precision of each algorithm. Although it showed better accuracy, only a small portion of such high-resolution data is available.

Versteeg [15] used simple calculations for crop production. This study analyses crop yield depending on growth rate. This is the oldest approach to crop yield prediction and it is not very efficient. This tool is used sparingly and is only pertinent in a few unusual circumstances.

Balakrishnan et al. [16] used AdaSVM and AdaNaive as the proposed ensemble model for crop production. This ensemble model is evaluated with SVM and Naïve Bayes. From the findings, it can be inferred that both of the proposed strategies have a fair degree of improvement in prediction accuracy and a good degree of decline in the percentage of classification mistakes.

Paudel et al. [17] utilize Machine Learning tools. The limitation of this study is that it does not identify a specific model or set of parameters that may consistently produce high performance and instead heavily relies on the ongoing measurement of numerous models.

Tseng et al. [18] utilized IoT for predicting crop yield. The created model made use of an IoT sensor device. The goal of big data analysis in IoT aimed to examine environmental aberrations also to assess and comprehend the farmers' crop-growing practices. However, this model presented an uncommon distribution if revealed to probable threat in soil moisture content, temperature, and air humidity.

III. METHODOLOGY

The establishment of a crop production estimation model is possible through machine learning. In addition to estimating yield, this study also analyses the most suitable model among several ML models. In this study, we made use of agricultural data to predict crop yield in the first step. The agriculture dataset was then cleaned up of any noisy information. To create a suitable model for predicting crop yield, we extracted attributes from the pre-processed data in the following step. To keep track of the environmental conditions of farming operations, a sensor device was created. Data was gathered and analyzed to comprehend the techniques farmers employ for growing crops and the variables that influence agriculture growth. To evaluate the interactions between environmental factors and risks, a 3D clustering model which is a circular curve analog to a Euclidean space was used, and it arranges data point in a circular pattern depending on their closeness to each other. However, for the model to operate as intended, proper calibration is required. Crop yield was predicted by Tiwari and Shukla (2011) using an artificial neural network (CNN) and a geographic information system (GIS) [9]. However, their model did not perform well with real-time data. A fresh technique for estimating crop yields has been created, and it can be used with almost any kind of crop. This new model is adaptable for use in a range of agricultural circumstances and is simple to use. A method for a global and local period explanation, and feature extractors offered a way to recognise numerous pests and illnesses. This model can handle difficult scenarios from a nearby region. Less false positives are produced in training as a result of more accurate identification.

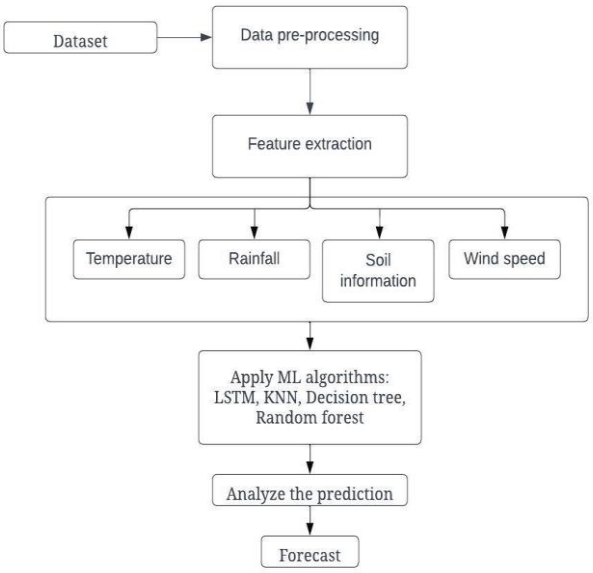

Figure 1: Block Diagram

Figure 1 pictorially depicts the methodology. The dataset is collected and pre-processed to remove missing and null values. Then feature extraction is performed on the data. Parameters like temperature, rainfall, and soil extract the feature. Then we applied ML algorithms like Decision Tree, KNN, and so on to predict crop yield. The prediction is analyzed using regression then the forecast is displayed.

IV. PROPOSED MODEL

In this section, we discuss the precision and effectiveness of various Machine Learning models by meticulously analyzing the error rates and forecasts crop yield based on environmental factors. By taking into account a variety of variables, including weather conditions, this study also determines and forecasts the region where the crop will be grown. Machine learning models can be trained on historical data related to crop yield and various factors that impact it, such as weather patterns, soil conditions, irrigation, fertilization, etc. The trained models can then be employed to forecast future crop yields depending on current and forecasted conditions. Some of the commonly utilized machine learning algorithms for crop yield prediction are regression models, Decision trees, Neural Networks, and SVMs. The precision of the prediction relies on the quality and quantity of data used for training the model and the selection of appropriate features. These crop yield prediction models can be applied to various crops, including cereals, vegetables, fruits, and oilseeds. These models can be useful for farmers to make informed decisions about crops.

A. Data collection:

The parameters that could affect crop yield were collected. The datasets were collected from various sources. It includes rainfall data, temperature data, yield data, and pesticide data from the past century.

B. Data processing:

The collected data is pre-processed to remove any missing or incorrect data points, and to transform the data into a format suitable for ML models. This involves data cleaning, normalization, and feature engineering. The data are cleaned to remove any error or inconsistencies.

C. Feature extraction:

Estimating a crop's production is a difficult process that depends on a wide range of variables. Temperature, precipitation, soil information, humidity, and wind speed are a few of these.

D. Feature selection:

Relevant features that are most strongly correlated with crop yield are selected using statistical and machine learning techniques.

E. LSTM:

Time-based data can be categorized, processed, and forecasted using an LSTM network.

F. Linear regression:

Linear regression is used to estimate crop production. The collected data is split into two sets called training data and testing data. The trained data identifies the strongly correlated inputs. The test data was used to find the precision of the model.

G. Decision tree:

This algorithm creates a tree-like structure that is generated from the train and test data. It starts from the root of the tree, divides data depending on the feature, and produces the result. This model divides the data based on feature values into nodes or branches as far as prediction is made.

H. Random forests:

It is a predictive model. Random forests can examine crop growth and biophysical alterations concerning the present climatic circumstances. The Random Forest algorithm develops numerous decision trees and generates a final prediction. This yields success rates of more than 92.81%. Random forest is an approach that trains data using the bagging method. Compared to other popular forms of machine learning techniques, random forests often offer a better level of precision. This model shows a precision of about 91% and is considered to be the most accurate model.

I. K-nearest Neighbors (KNN):

Crop production predictions are made using past information such as rainfall, temperature, and groundwater level. The groundwater level dataset is classified using the KNN model is beneficial for analyzing past groundwater levels and forecasting future levels.

J. Stochastic Gradient Descent regression (SGD):

SGD is a popular optimization technique frequently employed in machine learning applications to determine the model parameters that best fit the predicted and observed outputs. SGD is a variant of gradient descent algorithm. In logistic regression, two optimization functions are commonly used to locate the most suitable regression weights—gradient descent and SGD. These algorithms iteratively edit a set of parameters to minimize an error function. In a single iteration, the gradient descent algorithm

updates the parameters depending on the gradient of loss function evaluated on the training data. SGD updates the model parameters by randomly selected set of training data which then estimates the yield. This algorithm accelerates training process.

V. RESULTS AND DISCUSSION

Table 1: Dataset sample of features

	Year	Country	Item	Rainfall (mm)	Temperature (Celsius)	Pesticides (tonnes)	Yield (hg/ha)
0	1990	Albania	Barley	812.23450	12.051221	121	10000
1	1990	Albania	Carrots and turnips	812.23450	12.051221	121	150000
2	1990	Albania	Cauliflowers and broccoli	812.23450	12.051221	121	171429
3	1990	Albania	Garlic	812.23450	12.051221	121	65000
4	1990	Albania	Maize	812.23450	12.051221	121	36613
...	...	...	...	...	...	...	...
25224	2016	Zimbabwe	Maize	455.13295	22.310167	2185	4405
25225	2016	Zimbabwe	Oats	455.13295	22.310167	2185	20505
25226	2016	Zimbabwe	Potatoes	455.13295	22.310167	2185	51792
25227	2016	Zimbabwe	Sweet potatoes	455.13295	22.310167	2185	27283
25228	2016	Zimbabwe	Wheat	455.13295	22.310167	2185	19013

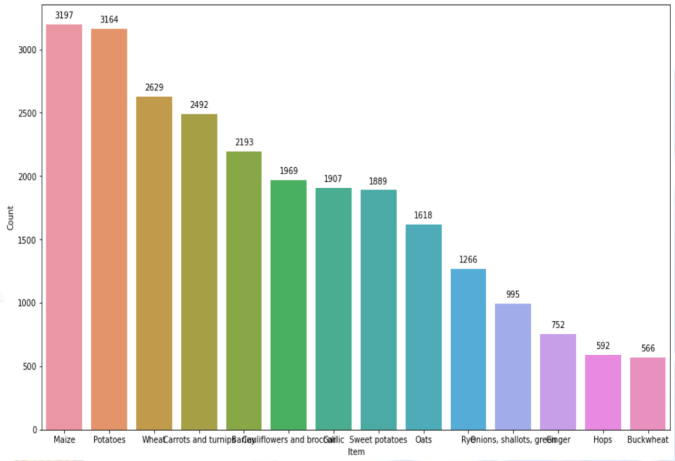

Figure 2: Frequency distribution of various crops

Figure 2 shows the frequency count of various crops. This result was brought by considering several parameters such as rainfall, temperature, yield of previous years, and pesticide.

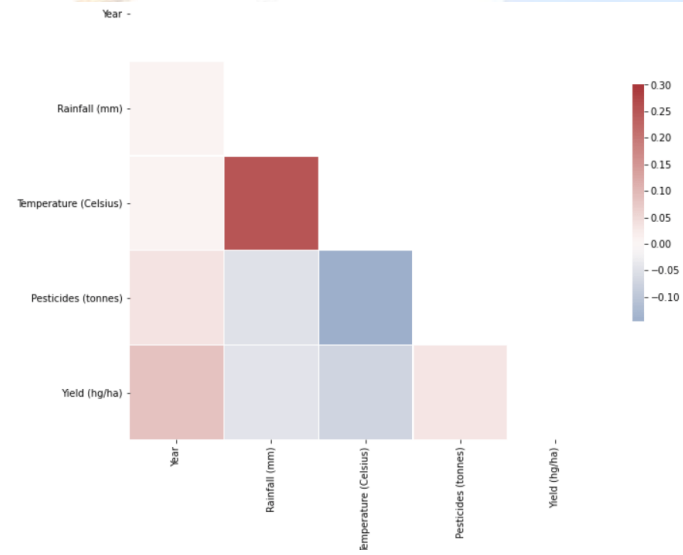

Figure 3: Analysis using heatmap

Figure 3 describes the correlation of temperature, rainfall, and pesticides by visualizing them in a heatmap.

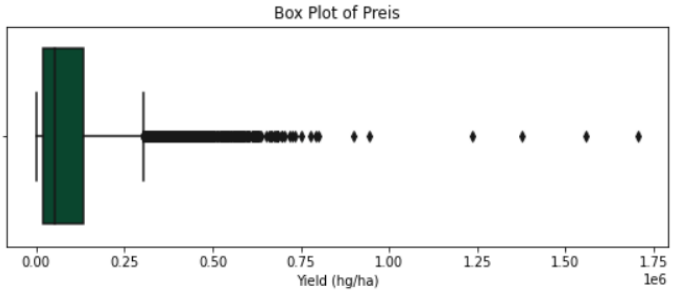

Figure 4: Visual representation using box plot

Figure 4 illustrates the box plot representation. The yield data was plotted in a box plot and it was used to find the range of yields and outliers. We visualized the distribution of past crop yields to predict future yields.

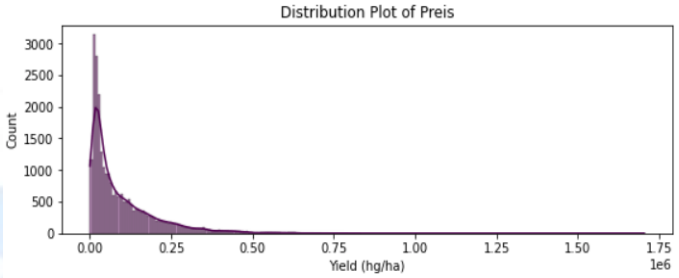

Figure 5: Visual representation using distribution plot

Figure 5 represents the distribution plot representation. It is used for the same purpose as a box plot. It was used to understand the distribution of crop yield and to identify the outliers. This is an alternate method of the box plot.

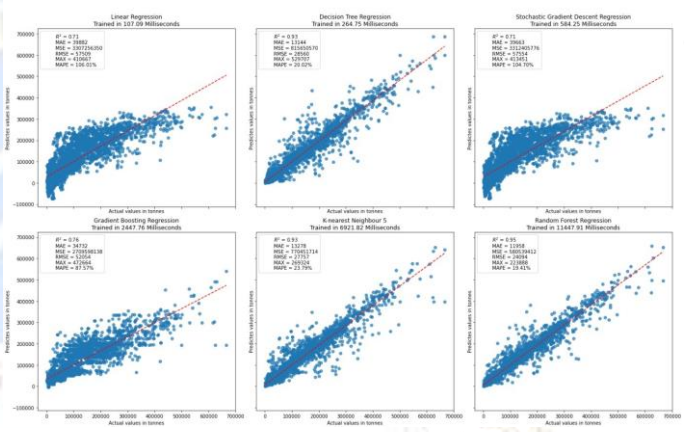

Figure 6: Regression analysis of ML models

Figure 6 depicts the regression analysis of ML models like decision tree, linear regression, SGD, random forest, etc. The result also shows the error rates of the ML models from which it is evident that the random forest has the lowest error rate.

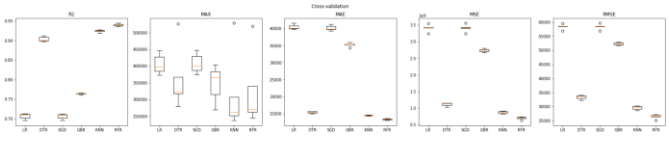

Figure 7: Cross-validation of ML models

Figure 7 describes the result of cross-validation. It is used to assess the performance of a model. It guaranteed a model's accuracy by recognizing overfitting or underfitting. It estimated the performance of each ML model. The data was split into a training set and a validation set. The training set was used to develop the model and the validation set is used to evaluate its performance.

Table 2: Error rates of ML models

Model	R^2	MSE(tones)	RMSE(tones)
Linear regression	**0.71**	**3307256350**	**57509**
Decision tree	**0.93**	**815650570**	**28560**
SGD	**0.71**	**3312405776**	**57554**
KNN	**0.93**	**770451717**	**27757**
Random forest	**0.95**	**508539412**	**24094**
Gradient Boosting regression	**0.76**	**2709598138**	**52054**

Table 2 shows the error rates of R^2, Mean Square Error (MSE), and Root mean square error (RMSE) in tones. Of all ML models, the Random forest has the lowest MSE and RMSE.

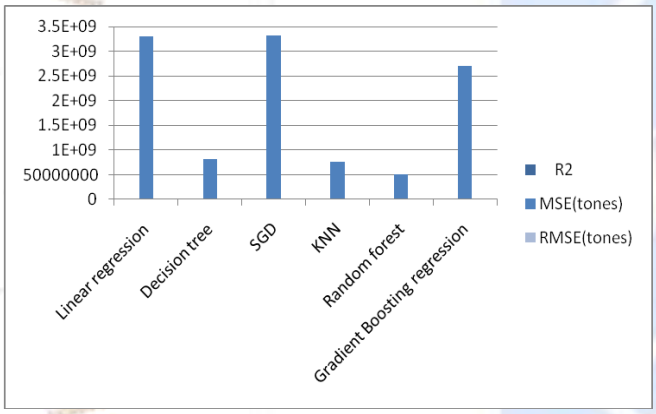

Figure 8: Analysis of error rates using bar chart

Figure 8 represents the error rates of each ML model in the bar chart. From the result, it is transparent that the Random forest has the lowest error rate.

VI. CONCLUSION AND FUTURE WORK

In conclusion, our study used various machine learning algorithms, including linear regression, decision trees, random forest, and SGD, to predict crop yield. The findings of this study have important implications for farmers and researchers, as they can use this information to optimize their farming practices and increase crop yield. This study also goes over how an algorithm's output is impacted by features and data accessibility. We also provide a list of features that although not tested in this study, may be pertinent to subsequent research. The findings of this research show that the Random Forest algorithm is efficient among other Machine Learning algorithms in regression analysis. The random forest algorithm showed over 85% accuracy rates. Figure 5 depicts regression with the error rates of each ML model used in the research. Of all ML algorithms, the random forest algorithm showed fewer error rates which can be observed in Figure 8. In the experimental study, researchers used ML techniques for crop prediction in the context of agriculture. We can estimate crop yield accurately and open doors for better agricultural practices using this technology. Future research can focus on additional features such as temperature, soil, humidity, and so on to provide more stable and precise estimation models. We can assist farmers in increasing their profitability and making a positive impact on the world's food security by applying ML algorithms to estimate crop yields.

REFERENCE

[1] Aruvansh Nigam, Saksham Garg, Archit Agrawal, Parul Agrawal, "Crop Yield Prediction Using Machine Learning Algorithms", IEEE International Conference on Image Information Processing, 2019.

[2] R. Ghadge, J. Kulkarni, P. More, S. Nene, and R. L. Priya, "Prediction of crop yield using machine learning," Int. Res. J. Eng. Technolgy, vol. 5, 2018.

[3] Suresh, N. Manjunathan, P. Rajesh, and E. Thangadurai, "Crop Yield Prediction Using Linear Support Vector Machine," European Journal of Molecular & Clinical Medicine, vol. 7, no. 6, pp. 2189- 2195, 2020.

[4] F. H. Tseng, H. H.Cho, and H. T. Wu, "Applying big data for intelligent agriculture-based crop selection analysis," IEEE Access, vol. 7, pp. 116965-116974, 2019.

[5] M. Alagurajan, and C. Vijayakumaran, "ML Methods for Crop Yield Prediction and Estimation: An Exploration," International Journal of Engineering and Advanced Technology, vol. 9 no. 3, 2020.

[6] R.Pradeep Sarvana Kumar, V.K.S Bhallaji, "A Novel Approach towards the design of an Efficient System for Optimizing the usage of Fertilizers", International Conference on Embedded Systems, 2014.

[7] S. D. Kumar, S. Esakkirajan, S. Bama, and B. Keerthiveena, "A microcontroller based machine vision approach for tomato grading and sorting using SVM classifier," Microprocessors and Microsystems, vol. 76, pp.103090, 2020.

[8] P. Sivanandhini, and J. Prakash, "Crop Yield Prediction Analysis using Feed Forward and Recurrent Neural Network," International Journal of Innovative Science and Research Technology, vol. 5, no. 5, pp. 1092-1096, 2020.

[9] P. Tiwari, and P. Shukla, "Crop yield prediction by modified convolutional neural network and geographical indexes," International Journal of Computer Sciences and Engineering, vol. 6, no. 8, pp. 503-513, 2018.

[10] Shreya V. Bhosale, Ruchita A. Thombare, Prasanna G. Dhemey, Anagha N. Chaudhari, "Crop Yield Prediction Using Data Analytics and Hybrid Approach", IEEE International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.

[11] Shruti Kulkarni, Shah Nawaz Mandal, G. Srivatsa Sharma, Monica R. Mundada, Meeradevi, "Predictive Analysis to Improve Crop Yield using a Neural Network Model", IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018.

[12] Krutika Hampannavar, Vijay Bhajantri, Shashikumar. G. Totad, "Prediction of Crop Fertilizer Consumption", IEEE International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.

[13] Tanhim Islam, Tanjir Alam Chisty, Amitabha Chakrabarty, "A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh", IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2018.

[14] A Suresh, P. Ganesh Kumar, M. Ramalatha, "Prediction of major crop yields of Tamilnadu using K-means and Modified KNN", IEEE International Conference on Communication and Electronics Systems (ICCES), 2018.

[15] M.N. Versteeg, H. van Keulen, "Potential crop production by some simple calculation methods, as compared with computer simulations", Centre for Agrobiological Research (CABO), 1986.

[16] Narayanan Balakrishnan, Dr.Govindarajan Muthukumarasamy, "Crop Production-Ensemble Machine Learning Model for Prediction", International Journal of Computer Science and Software Engineering (IJCSSE), vol. 5, pp. 148-153, 2016.

[17] Dilli Paudel, Hendrik Boogaard, Allard de Wit, Marijn van der Velde, Martin Claverie, Luigi Nisini, Sander Janssen, Sjoukje Osinga, Ioannis N. Athanasiadis, "Machine learning for regional crop yield forecasting in Europe", 2022.

[18] F. H. Tseng, H. H.Cho, H. T. Wu, "Applying big data for intelligent agriculture-based crop selection analysis," IEEE Access, vol. 7, pp. 116965-116974, 2019.