

A Machine Learning Approach for Predicting Crop Yield in Precision Agriculture

Santosh Kumar Upadhyay
CSE Department
AKGEC, Ghaziabad, India
upadhyaysantosh@akgec.ac.in

Priyanshi Kushwaha
CSE Department
AKGEC, Ghaziabad, India
priyanshi2010047@akgec.ac.in

Prakhar Garg
CSE Department
AKGEC, Ghaziabad, India
prakhar2010196@akgec.ac.in

Abstract— Agriculture is the main earnings-producing field as well as a cause of livelihood in India. Different biological variables and seasonal and financial factors affect yield growth, but unexpected variations in these variables result in a major loss of crops. When adequate mathematical or statistical techniques are applied to information related to soil, climate, and previous yield, these hazards can be quantified. With the advancement of machine learning, crop yields may be anticipated by extracting helpful information from crop fields that assist the government in deciding import/ export in advance. This study provides a machine-learning approach based on Random Forest Regression to predict crop yield with an R-square of 0.95. This agricultural yield prediction helps farmers to make plans for shortage/surplus of production in well advance to get significant benefits.

Keywords— Agriculture, Machine learning, Prediction, Crop, Decision Tree, Random Forest Algorithm

I. INTRODUCTION

Prediction of Crop yield is useful in the nation's food management and development. Agriculture has a direct influence on domestic and global GDP, and yield prediction plays an important role in the procurement of food. India's economy relies mainly on the growth of agricultural yields and the products of its associated agro-industry. Prediction of agriculture yield before harvest is quite convenient for peasants to decide their future activities. This helps the farmers to do the planning, formulate import/export choices, manage prices, and procure crops. People working in this domain have created predictive methods based on climatic factors and previous crop yields that might suggest a pre-harvest practice. The majority of crop yields rely on rainwater, which is extremely unpredictable. Crop development relies on various factors such as climate, nitrogen, water, soil characteristics, rainwater, surface temperature, soil moisture, plant rotation, etc. Various areas of crop remote sensing have utilized the Artificial Neural Network, particularly for the classification of crop type and the assessment of crop region. Many machine learning (ML) approaches such as Bayesian network, Decision tree, regression analysis, random forest, support vector regression, and convolutional neural network (Deep learning) are also used to solve issues in Precision agriculture. Advanced ML approaches like deep learning particularly CNN became a boon for crop disease detection [1,2]. Transfer learning approaches also proved to be a good choice for plant disease classification [3,4,5,22].

One of the primary objectives of agricultural production is to obtain the highest crop yield at a minimum price in a healthy environment. Yield growth and plant development are based on several parameters like climatic conditions, soil characteristics, irrigation, topography, and fertilizer choices. The need for timely and appropriate sensing of such

characteristics for big farming areas has resulted in increased acceptance of remote and proximal sensing systems[6].

Several machine-learning methods have been introduced in recent years to obtain precise yield forecasts for various plants[7]. The most effective research in the area of yield prediction using machine learning methods up to the year 2010 were ANN [8,9,10], M5-Prime Regression Trees [11,12], Support Vector Regression, and KNN, etc.

The primary outcome of this study is the proposal of an effective and efficient yield prediction method for Precision Agriculture.

The rest of this paper is organized as follows: A study of the reviewed publications is presented in section 2. Dataset and Methods are discussed in section 3. Section 4 is dedicated to a discussion of the result analysis. The conclusion is presented in the final section (section 5).

II. LITERATURE REVIEW

Agricultural yield prediction is essential for planning policies, ensuring the availability of food, and making agricultural decisions. Crop production estimates have been more accurate and dependable in recent years because of the use of methods based on machine learning (ML), especially the Random Forest algorithm. In this study, we have reviewed several methods of agricultural yield prediction based on Decision tree and Random Forest machine learning algorithms. This section presents a review in two subsections: Decision tree algorithm-based methods and Random Forest algorithm-based methods.

A. Decision tree algorithm-based methods

This subsection describes various crop yield forecasting methods based on the decision tree algorithm.

Veenadhari et al.[13] presented a model for Soybean Productivity using Decision Tree Algorithms. The current research is focused on a decision tree to evaluate the impact of climatic variables on the output of soybean crops. Interactive Dichotomizer 3 (ID3) decision tree algorithm was used to specify the most affecting climatic parameter on soybean production in the Bhopal district. The decision trees stated there is a correlation between environmental factors and productivity of soybean crops and these variables have been verified by the rule accuracy and Bayesian classification. Decision tree results were structured into distinct rules for the end-users to better understand the impact of climate variables on the output of soybean crops. The salient findings of this study are: 1)The analysis of the decision tree suggested that soybean crop productivity was mostly affected by relative humidity followed by temperature and rainfall 2)This study's decision tree is quick to implement and much to be required as depictions of interpretations of understanding 3)The

guidelines created from the decision tree are useful in anticipating circumstances under specified environmental parameters that are accountable for elevated or poor soybean crop productivity. One of the limitations is that the model only predicts low or high yield, but cannot predict the quantity of yield output.

D.M. Johnson [14] used a decision tree algorithm for yield prediction of soybean and corn crops in the United- States. An evaluation of pre- and in-season remotely sensed factors for the yield prediction was done by using four predictors. These predictors were precipitation, nighttime land surface temperature (LST), daytime LST, and Normalized Difference Vegetation Index (NDVI). Nighttime LST, daytime LST, and NDVI were collected from the MODIS sensor. Precipitation was acquired from the National Weather Service (NWS). During the summer and most significantly in early August, NDVI was discovered to be strongly and favorably associated with yields of soybean and corn whereas early-season NDVI was adversely associated with crop yields. It was observed that Daytime LST was adversely associated with crop yields during the summer. The investigation on precipitation and Nighttime LST suggested that there was no correlation between crop yields and these parameters. A decision rule-based predictive model was prepared by incorporating the daytime LST variables and the full-time series of NDVI. The proposed predictive model illustrated very nice results ($R^2 = 0.93$) for both crops.

Mahesh et al. [15] provide a prediction of Crop Yield using ML Algorithms based on Soil Moisture. Crops including rice, maize, kidney beans, chickpeas, and so forth were employed for the forecast, according to Zenodo (CERN European Organisation for Nuclear Research). Accuracy increased by estimating crop yield based on soil moisture using a decision tree machine learning technique. The first dataset was a soil moisture dataset that offers soil information. The second dataset was the crop dataset. Important components in the crop dataset include crop yield prediction type based on N, P, K, and pH levels; input nutrient attribute values are included in the soil moisture dataset. When crop output is forecast using a machine learning decision tree approach based on soil moisture data, better outcomes are achieved. In this proposed system, an accuracy of 95% was gained.

This work [16] used the ID3 algorithm to forecast whether the yield of soybeans will be high or low. ID3's output was contrasted with the Naïve Bayes (NB) classifier. The results showed that compared to the NB classifier, the ID3 algorithm performed 7% better.

B. Random Forest algorithm-based methods

This subsection describes various crop yield forecasting methods based on the Random Forest algorithm.

Fukuda et al.[17] presented a model for the prediction of yields of mango fruit under separate irrigation systems using a random forest algorithm in Northern Thailand. Additionally, the information obtained from the RF models may be helpful for water management under field circumstances. Rainfall data and irrigation information from 10 days were combined to develop 4 RF models. The ranges of mango fruit yield were accurately estimated by RF models. The assessment of the result showed that the estimation of the production of mango primarily depends on the timing of the water supply. The use of various input factors emphasized the significance of

irrigation information and rainfall on crop production. This research on the assessment of the production of mango shows the significance of Random Forest in precision agriculture, with particular emphasis on irrigation strategy.

Everingham et al.[18] presented a random forest model to predict sugarcane production accurately. The Random Forest algorithm was used for each prediction date to define variable significance for all available predictor variables. The Out-of-bag (OOB) prediction error was used as a measure of variable significance. For the optimization of the random forest model, a forward selection method was used. In other words, the models were reconstructed with the most significant feature and then sequentially added more variables to optimize the OOB R-squared for regression models and the OOB classification error rate for classification models. This provides the model with the chance to enhance efficiency while maintaining a reduced set of features to avoid the problem of overfitting. To demonstrate year-wise variability in regional sugarcane yields at Tully in northwestern Australia, the input provided to the Random Forest algorithm contained six factors. These factors were seasonal weather prediction indices, sugarcane crop model, Simulated biomass from the APSIM, observed rainfall and recorded highest and smallest temperature and radiation. The findings of this research stated that it was feasible to determine whether output would be above the average as soon as September in the year before harvest in 86.36 percent of years. This precision was enhanced by January in the harvest year to 95.45 percent. It was observed that the R-squared parameter of the random forest regression model consistently increased in the duration of the pre-harvest season to the same harvest season. This enhancement was recorded as 66.76 to 79.21 percent. Observed weather conditions like rainfall, highest and smallest temperature and radiation, simulated biomass indices, and seasonal climate prediction indices were typically presented at different phases in the models.

Crop prediction using machine learning techniques, more especially, the Random Forest classifier is covered in [19]. The study builds a crop production prediction model by analyzing historical data on soil properties and climatic patterns. Random Forest is used to effectively capture the complex relationships and interactions among several data components. By analyzing historical data on soil qualities and meteorological trends, the study creates a model for predicting crop yield. Depending on the characteristics, dataset, and particular decision tree technique used, the accuracy of agricultural production forecasts made with decision tree models may differ. The accuracy according to the author's dataset was 97.69%.

The potential of the Random Forest algorithm for predicting tea output in the Indian state of Assam was examined by Deka et al. [20]. Data on climatic variables, fertilizer application, irrigation, and other agronomic aspects were included in the study along with data on tea output. The results showed that when it came to tea yield prediction, the Random Forest model performed better than other machine learning methods like Decision Trees and Support Vector Machines.

Machine Learning algorithm plays a crucial role in crop yield prediction and crop recommendation [23,24,25].The evaluation of the existing research leads to the conclusion that using ML (particularly the Random Forest algorithm) in agricultural yield forecasting might improve agricultural

decision-making. This is accomplished by giving accurate agricultural production predictions and identifying the best crops to grow while accounting for environmental factors. The studies that have been evaluated highlight the potential benefits of these systems for farmers and underline the need for machine learning algorithms in building accurate crop yield prediction systems.

III. MATERIAL AND METHODS

The Layout of the crop yield prediction system is shown in Fig. 1. The dataset that is utilized to train and evaluate the suggested model is described in this section. This section also covers Decision tree regression and Random Forest regression ML techniques that are employed in the creation of yield prediction systems.

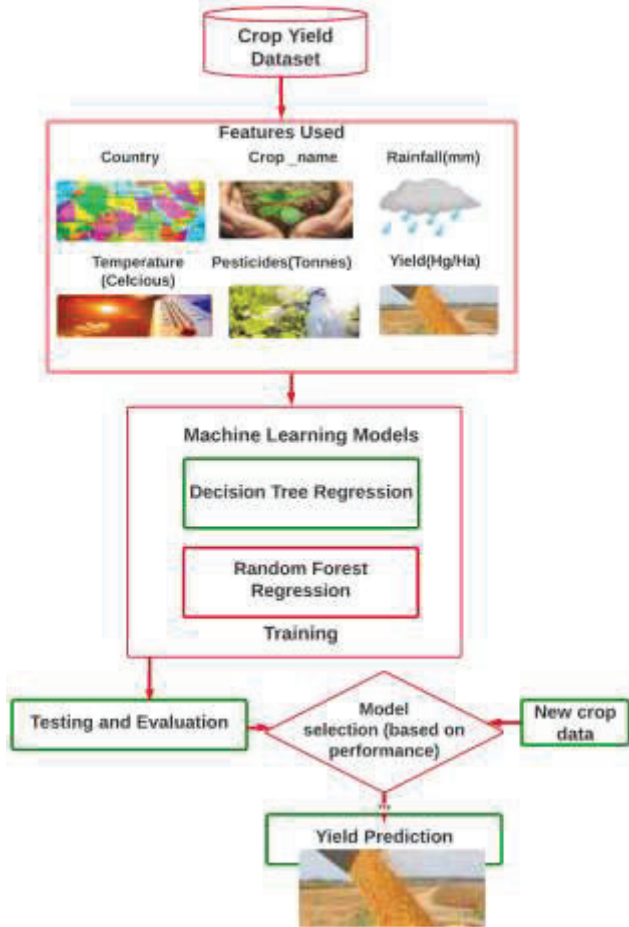


Fig. 1. Yield Prediction System Layout

A. Experimental Setup

Using a laptop with an Intel-Core i5-1035G1 CPU running at 1.00GHz, 4GB of RAM, a 250 GB SSD, and Windows 10, the model was constructed and validated. The programs are implemented and the dataset is analysed using Python Jupyter Notebook.

B. Dataset Collection

Features in the dataset include Temperature (Celsius), Rainfall (millimeter), Pesticides (Tonnes), Yield (hectograms per hectare i.e. hg/ha), Item (crop name) along with Year and Country. The dataset is collected from the Kaggle. The data set consists of 25229 rows and 7 columns(Features). This analysis uses the dataset [21]. The Food and Agriculture Organisation of the US is the source of data on pesticides and

yield. The World Data Bank is the source of the average temperature and amount of rainfall.

The final dataset was created after pesticides, yield, rainfall, and temperature were cleaned, pre-processed, and combined. An example of the data that was used is shown in Fig. 2.

	Year	Country	Item	Rainfall (mm)	Temperature (Celcius)	Pesticides (tonnes)	Yield (hg/ha)
0	1990	Albania	Barley	812.23450	12.851221	121	10000
1	1990	Albania	Carrots and turneps	812.23450	12.851221	121	150000
2	1990	Albania	Cauliflores and broccoli	812.23450	12.851221	121	171429
3	1990	Albania	Garlic	812.23450	12.851221	121	65000
4	1990	Albania	Maize	812.23450	12.851221	121	36613
...
25224	2016	Zimbabwe	Maize	455.13295	22.310167	2185	4405
25225	2016	Zimbabwe	Onions	455.13295	22.310167	2185	20505
25226	2016	Zimbabwe	Potatoes	455.13295	22.310167	2185	31792
25227	2016	Zimbabwe	Sweet potatoes	455.13295	22.310167	2185	27283
25228	2016	Zimbabwe	Wheat	455.13295	22.310167	2185	19013

25229 rows x 7 columns

Fig. 2. Sample of Dataset

C. Machine Learning Algorithms used in Proposed method

The proposed method used Decision tree regression(DTR) and Random Forest regression(RFR) to forecast crop yield. These algorithms were chosen because they have been used in crop recommendation as well as prediction systems a lot and because previous studies have shown them to be effective.

1) *Decision Tree Regression*: A straightforward and understandable technique that works with both numerical and categorical data is called Decision Tree regression. It is also resistant to overfitting and capable of efficiently handling missing variables and outliers. However, if the tree is too deep or complicated, it cannot generalize effectively to new data and is more prone to high variation and instability. Pruning and ensemble approaches, like Random Forest, can be utilized to get around this. All things considered, DecisionTreeRegressor is a strong and popular method for regression problems, particularly where interpretability and simplicity are sought.

2) *Random Forest Regression algorithm*: Random Forest algorithm is a popular machine learning approach that may be used for both regression and classification. By merging many decision trees, a method known as ensemble learning improves accuracy and reduces the overfitting of the model. Using a random subset of the input characteristics and a piece of the training data, each decision tree is generated by the random forest technique (Fig. 3). Using a method known as bootstrap aggregating, or bagging, where replacement is permitted, the subgroups are chosen at random.

The average of the forecasts is used to calculate the final result in the case of regression. When compared to alternative algorithms, the random forest approach has several benefits. It is very resistant to outliers and missing values, and it performs exceptionally well when processing noisy, high-dimensional data. It also provides a measure for evaluating the importance

of characteristics, which helps in the process of choosing pertinent features and understanding the underlying data.

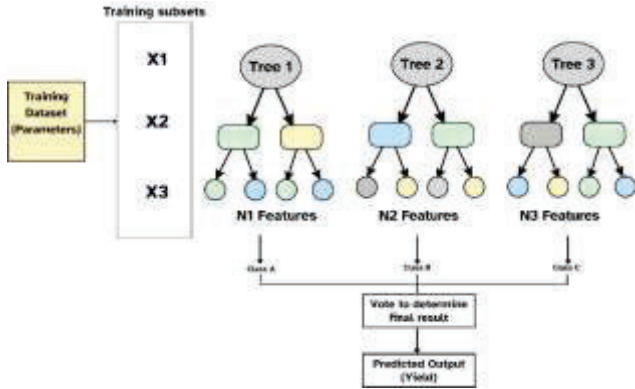


Fig. 3. Random Forest algorithm representation

IV. RESULT ANALYSIS AND DISCUSSION

Experimental result analysis of the suggested method is discussed in this section.

A. Performance Metrics

We used several metrics to evaluate the agricultural yield forecasting system, including R-square(R²), MAX score(MAX), Mean Absolute Error(MAE), Mean Squared Error(MSE), Mean Absolute Percentage Error(MAPE), and Root Mean Square Error(RMSE).

1) *R-square*: A measure of how much of the variance in the dependent variable can be anticipated from the independent variables R-square Score. An R² value closer to 1 indicates a better fit to the data and a higher projected accuracy of the model.

2) *MAX Score*: The largest absolute difference between the expected and actual values is indicated by the MAX Score.

3) *MAE*: The MAE technique is used to calculate the average absolute differences between the predicted and true values.

4) *MSE*: The MSE is the average of the squared differences between the predicted and real values.

5) *RMSE*: The square root of the MSE.

6) *MAPE*: The average magnitude of error generated by a model, or the average deviation from expectations, is measured by MAPE.

TABLE I. PERFORMANCE COMPARISON OF MODELS

ML Approach	R ²	MAE	MSE	RMSE	MAX	MAPE
DTR	0.92	13220	856110050	29259	529707	20.31%
RFR	0.95	11929	573115982	23940	219124	19.31%

TABLE II. COMPUTATION TIME

ML Approach	Computation Time
DTR	246.20 milliseconds
RFR	13138.86 milliseconds

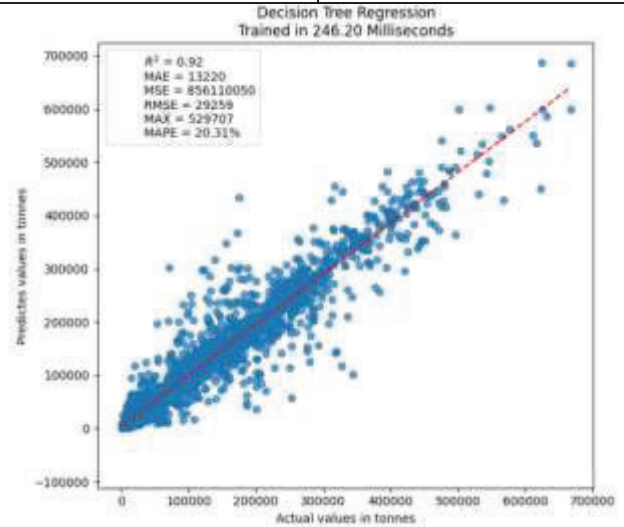


Fig. 4. Decision tree Regression

B. Result discussion and Comparision

Authentic data from Kaggle was used to evaluate the crop yield prediction system. The system attained an R-square of 0.92 with Decision tree regression (DTR), which is a promising result. MAE, MSE, RMSE, MAX, and MAPE values for DTR are noted as 13220, 856110050, 29259, 529707, and 20.31%. Further, the system attained an R-square of 0.95 with Random Forest regression (RFR), which is a better result than DTR. MAE, MSE, RMSE, MAX, and MAPE values for RFR are noted as 11929, 573115982, 23940, 219124, and 19.30% signifying a significant improvement over Decision Tree regression. These combined findings demonstrate the usefulness of Random Forest in our yield prediction system, which offers computational economy together with improved projected accuracy. The Random Forest Algorithm produces the best prediction residuals, with the lowest MAPE value and the greatest R² value, as we see. The random forest also has significantly lower MAX, RMSE, and MSE values than the DTR algorithm. The result prediction for the Decision tree algorithm is shown in Fig.4 whereas the result prediction for the Random Forest algorithm is shown in Fig.5.

The performance of the Decision Tree Regression algorithm and Random Forest Regression algorithm is illustrated in Table 1. The computational time for both algorithms is shown in Table 2.

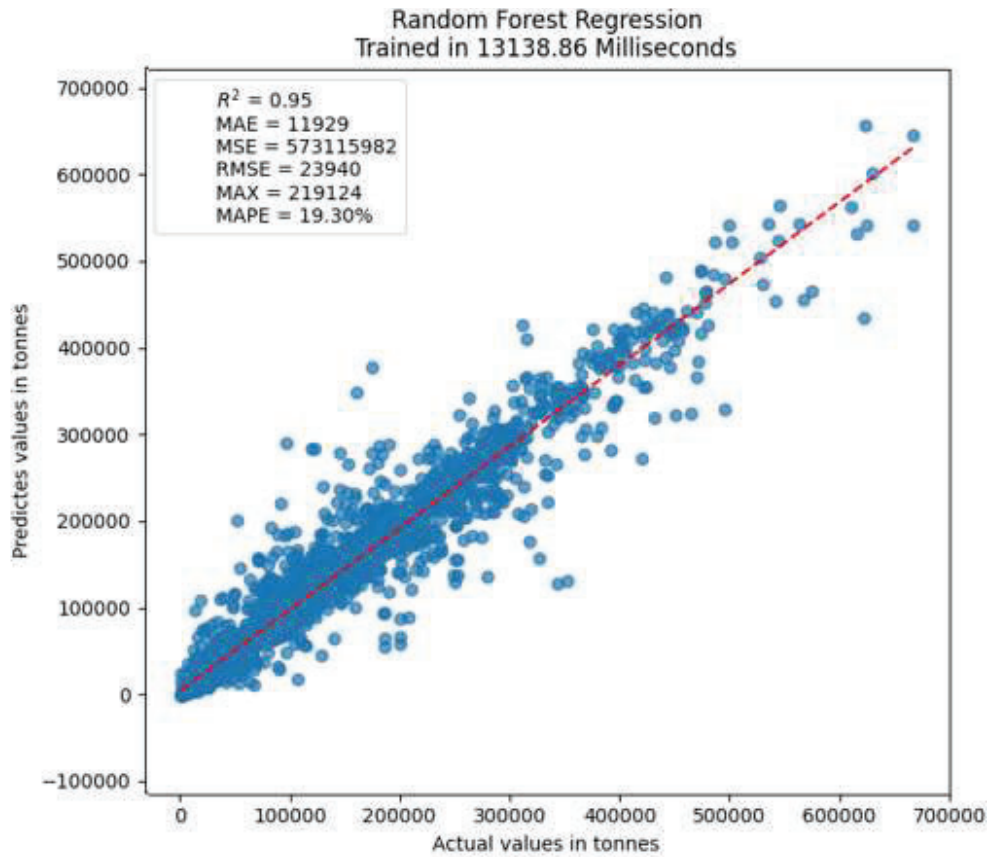


Fig. 5. Random Forest Regression

V. CONCLUSION AND FUTURE WORK

This paper presents a crop yield forecasting system that uses machine learning to determine the yield of production while taking the country, crop name, rainfall, temperature, pesticides, and previous crop yield into consideration. The system employs Decision tree regression and Random Forest regression to predict the yields. Random forest regression is found to be more accurate with an R-square of 0.95. In the future, the proposed system can be embedded with IoT for more automation. Web interfaces and applications can be designed for taking input and displaying output. Proposed system aids in understanding the cumulative impacts of pests, a lack of nutrients and water, the fluctuation of crop yields, and other field variables over the period of growth. We know that crop yield is highly fluctuated by crop disease. As limitation, this system is not capable to include disease parameter in crop yield analysis.

REFERENCES

- [1] S.K. Upadhyay, & A. Kumar, "Early-Stage Brown Spot Disease Recognition in Paddy Using Image Processing and Deep Learning Techniques" *Traitement du Signal*, 2021, Vol. 38, No. 6, pp. 1755-1766. <https://doi.org/10.18280/ts.380619>
- [2] S.K. Upadhyay, & A. Kumar, "A novel approach for rice plant diseases classification with deep convolutional neural network" *Int. j. inf. tecnol.* **14**, 185–199 (2022). <https://doi.org/10.1007/s41870-021-00817-5>
- [3] Rukhsar and S. K. Upadhyay, "Rice Leaves Disease Detection and Classification Using Transfer Learning Technique," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 2151–2156, doi: 10.1109/ICACITE53722.2022.9823596.
- [4] S. K. Upadhyay and A. Kumar, "An Accurate and Automated plant disease detection system using transfer learning based Inception V3Model," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1144–1151, doi: 10.1109/ICACITE53722.2022.9823559.
- [5] Rukhsar and S. K. Upadhyay, "Deep Transfer Learning-Based Rice Leaves Disease Diagnosis and Classification model using InceptionV3", 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2022, pp. 493–499, doi: 10.1109/CISES54857.2022.9844374.
- [6] J.B. Campbell., R.H.Wynne "Introduction to Remote Sensing. Fifth ed. Guildford Press, 2011.
- [7] M. Subhadra, M. Debahuti, S. GourHari, "Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper", *Indian Journal of Science and Technology*, 2016, 9,38.
- [8] J. Liu, C.E. Goering, L. Tian, "A neural network for setting target corn yields", *Transactions of the American Society of Agricultural Engineers*, 2001, 44,705–713.
- [9] S.T. Drummond, K.A. Sudduth, A. Joshi, S.J. Birrel, N.R. Kitchen "Statistical and neural methods for site-specific yield prediction. *Transactions of the American Society of Agricultural Engineers*", 2003, 46,5–14.
- [10] Y. Uno, S. Prasher, R. Lacroix, P. Goel, Y. Karimi, A. Viau, R. Patel, "Artificial neural networks to predict corn yield from Compact Airborne Spectrographic Imager data", *Computers and Electronics in Agriculture*, (2005),47,149–161.
- [11] B. Marinkovic, J. Crnobarac, S. Brdar, B. Antic, G. Jacimovic, V. Crnojevic, "Data mining approach for predictive modeling of agricultural yield data", In *Proceeding of the First International Workshop on Sensing Technologies in Agriculture*, (2009), pp. 1–5, Novi Sad, Serbia.
- [12] Frausto-Solis, A. Gonzalez-Sanchez, M. Larre, "A New Method for Optimal Cropping Pattern" In 8th Mexican International Conference on Artificial Intelligence, (2009), pp. 566–577. Guanajuato, México.
- [13] S. Veenadhari, D.B. Mishra, C.D. Singh, "Soybean Productivity Modeling using Decision Tree Algorithms" *International Journal of Computer Applications*, (2011),27,11–15.

- [14] D.M. Johnson "An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yield in the United States", *Remote Sensing of Environment*, (2014). 141,116–128.
- [15] T.R. Mahesh & G. S. M, "Prediction of Crop Yield Based-on Soil Moisture using Machine Learning Algorithms" Zenodo (CERN European Organization for Nuclear Research), (2023). <https://doi.org/10.5281/zenodo.7780737>
- [16] A. Shastry, H.A.Sanjay, M.C. Sajini, "Decision Tree Based Crop Yield Prediction Using Agro-climatic Parameters", In Shetty, N.R., Patnaik, L.M., Nagaraj, H.C., Hamsavath, P.N., Nalini, N. (eds) *Emerging Research in Computing, Information, Communication, and Applications. Lecture Notes in Electrical Engineering*, (2022), vol 789. Springer, Singapore. https://doi.org/10.1007/978-981-16-1338-8_8.
- [17] S. Fukuda, W. Spreer, E. Yasunaga, K. Yuge, V. Sardud, J. Muller, "Random Forests modeling for the estimation of mango (*Mangifera indica* L. cv. Chok Anan) fruit yields under different irrigation regimes", *Agricultural Water Management*, (2013), 116,142-150.
- [18] Y. Everingham, J. Sexton, D. Skocaj, G. Inman-Bamber, "Accurate prediction of sugarcane yield using a random forest algorithm", *Agronomy for Sustainable Development*, (2016),36,27.
- [19] A. Karim, & J. Benjamin, "Crop Prediction using Random Forest Algorithm", National Conference on Emerging Computer Applications (NCECA), Amal Jyothi College of Engineering(2023). <https://doi.org/10.5281/zenodo.10147295>
- [20] P.C. Deka, A. Das, U.S. Saikia, & R. Borah, (2020). Tea yield forecasting using Random Forest algorithm. *Computers and Electronics in Agriculture*, 175, 105631.
- [21] <https://www.kaggle.com/code/nghianguyen39/crop-yield-prediction/data>, Accessed on 1st Dec.,2023.
- [22] S. K. Upadhyay and A. Kumar, "Automatic Recognition and Classification of Tomato Leaf Diseases Using Transfer Learning Model", *Future Farming: Advancing Agriculture with Artificial Intelligence* (2023) 1: 23. <https://doi.org/10.2174/9789815124729123010005>
- [23] Dolli, P. Rawat, M. Bajaj, S. Vats and V. Sharma, "An Analysis of Crop Recommendation Systems Employing Diverse Machine Learning Methodologies," 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), Dehradun, India, 2023, pp. 619-624, doi: 10.1109/DICCT56244.2023.10110085.
- [24] K. Purohit, S. Vats, R. Saklani, V. Kukreja, V. Sharma and S. P. Yadav, "Improvement in K-Means Clustering for Information Retrieval," 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2023, pp. 1239-1245, doi: 10.1109/ICESC57686.2023.10193031
- [25] V. Sharma, S. Vats, P. Rawat, and M. Bajaj (2023). "Crop recommendation system: A review." *Automation and Computation: Proceedings of the International Conference on Automation and Computation*, pp. 384-396.1st Edition, 2023, CRC Press, doi : 10.1201/9781003333500-44