WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Crop Yield Prediction using Machine Learning

Saravanan. R[1], Arulselvan Gnanamonickam. A[2]

[1] M. Sc. Software Systems, KG College of arts and science, Coimbatore, Tamil Nadu, India
[2]Assistant professor, KG College of arts and science, Coimbatore, Tamil Nadu, India
DOI : https://doi.org/10.55248/gengpi.5.1024.2825

A B S T R A C T

Crop yield prediction plays a pivotal role in enhancing agricultural productivity, particularly in agrarian economies like India, where the livelihoods of a significant portion of the population depend on farming. Given the diversity of climatic zones, soil types, and agricultural practices across the country, accurate crop yield prediction is essential for optimizing resource allocation, managing food security, and supporting sustainable agricultural practices.This research investigates the application of machine learning techniques, focusing on the Random Forest algorithm, to predict crop yields based on multiple influencing factors. These factors include weather conditions (temperature, rainfall, humidity), soil properties, the use of fertilizers, crop type, irrigation practices, and other relevant agricultural data. The Random Forest algorithm, known for its ability to handle large datasets and nonlinear relationships, is employed to build a predictive model using historical agricultural data from various regions in India.

Keywords: Accuracy, Agricultural Productivity

1. Introduction

Agriculture is the backbone of India's economy, contributing significantly to the livelihoods of millions of farmers and the country's food security. With the ever-increasing demand for food due to population growth and the challenges posed by climate change, improving agricultural productivity has become a critical concern. Traditional farming methods, heavily reliant on manual experience and historical knowledge, are no longer sufficient to address the complexities of modern agriculture. The adoption of advanced technologies is essential for enhancing crop production and ensuring sustainable agricultural practices. Crop yield prediction, the process of forecasting the potential output of a crop before harvest, is vital for optimizing resource utilization, planning agricultural activities, and managing risks. However, accurately predicting crop yields is challenging due to the wide range of factors that influence crop growth, including weather conditions, soil properties, input usage (such as fertilizers and pesticides), and local environmental conditions.In India, where agricultural diversity is vast, these variables become even more complex.This study focuses on applying the Random Forest algorithm to predict crop yields in Indian agriculture. By leveraging historical data on climatic conditions, soil properties, and agricultural inputs, this model provides farmers and policymakers with actionable insights to improve decision-making processes. The primary objective of this research is to demonstrate how machine learning can enhance the accuracy of crop yield predictions and contribute to the overall sustainability and efficiency of Indian agriculture.

2.Objectives of the project

The objective of the Crop Yield Prediction project is to develop an accurate predictive model using machine learning to forecast crop yields based on historical agricultural data. This project aims to utilize the Random Forest algorithm for its robust performance in handling various factors such as area, production, rainfall, fertilizer, and pesticide usage. The goal is to build a scalable model that can be applied across different states and regions of India, covering a wide range of crops and seasonal variations. By offering precise yield predictions, the model will aid farmers, policymakers, and agricultural planners in making informed decisions, optimizing resources, and improving agricultural productivity. Additionally, the project seeks to overcome the limitations of existing models, which often focus on specific crops or regions, by offering broader coverage and enhanced accuracy.

3. Proposed system

The proposed Crop Yield Prediction System is designed to address the limitations of existing systems by offering a more comprehensive, scalable, and accurate solution tailored to Indian agriculture. Leveraging the Random Forest algorithm, this system can analyze a wide range of variables such as soil conditions, rainfall, fertilizer, pesticide usage, and seasonal data to predict crop yields for various crops across different states of India.The proposed system integrates historical agricultural data, real-time weather updates, and other relevant factors to provide accurate, region-specific predictions. It incorporates a user-friendly web interface, making it accessible to farmers and policymakers alike. The model can handle diverse crops, unlike previous

systems limited to specific types, and applies robust preprocessing steps such as data cleaning, feature selection, and log transformation to improve the prediction performance. The use of Random Forest ensures better accuracy by minimizing overfitting and handling non-linear relationships in the data. Additionally, the system is designed to scale across multiple regions and crops, providing a broader, more inclusive solution for India's varied agricultural landscape.

3.1 Advantages

- It can predict the yield for a wide variety of crops across different states in India, making it highly adaptable to diverse agricultural conditions.
- By incorporating real-time data, including weather updates, the system ensures more relevant and timely predictions.
- Using the Random Forest algorithm, the system provides more accurate predictions by effectively handling non-linear data and multiple influencing factors like climate, area, and input usage.

4. Methodology

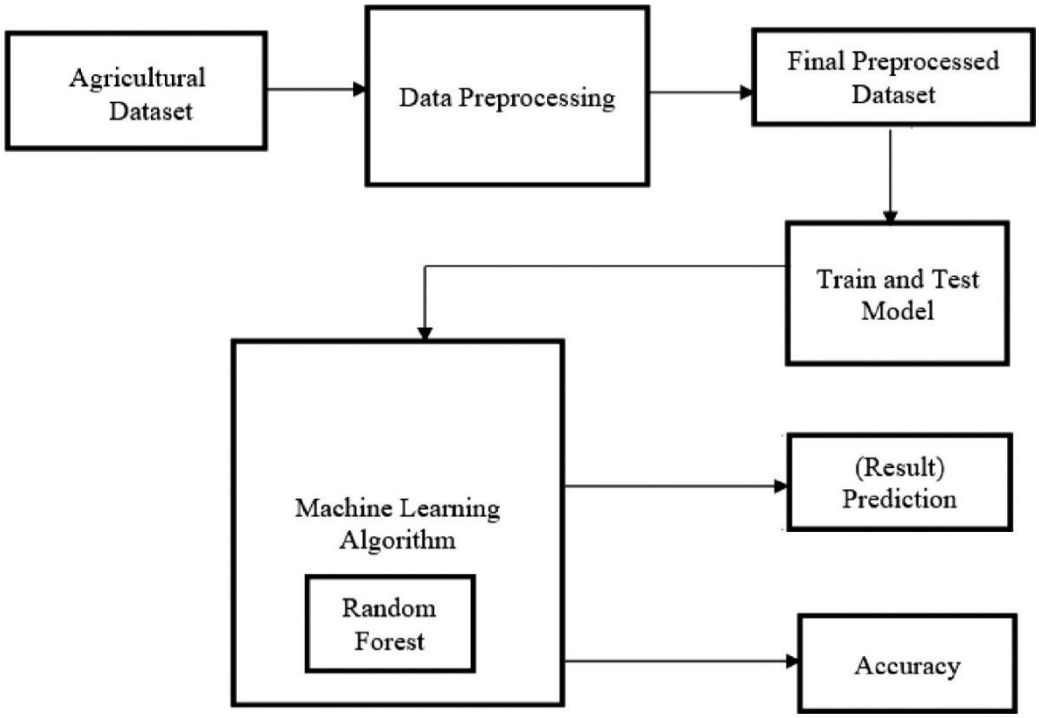

Fig. 1 - System Architecture

The methodology for this crop yield prediction study using Random Forest begins with data collection from reliable sources, including historical crop yield data, weather data (temperature, precipitation, humidity), soil characteristics (pH, moisture, nutrients), and agricultural practices (fertilizer use, irrigation). Preprocessing the data involves handling missing values through imputation techniques like mean or k-NN, normalizing numerical features, and encoding categorical variables such as crop types. After preprocessing, the Random Forest model is trained on this dataset, leveraging its ensemble nature to handle complex relationships between features. The model is evaluated using accuracy metrics like RMSE (Root Mean Square Error) and R^2, and further tuned through hyperparameter optimization to improve prediction accuracy. The model's performance is tested and validated against unseen data to ensure its robustness and reliability in predicting crop yields.

5. SYSTEM IMPLEMENTATION

The system comprises of the following modules:

- Data collection
- Data preprocess
- Model Selection
- Model Training
- Model Evaluation

Data Collection:

A relevant dataset sourced from Kaggle includes key agricultural features crucial for predicting crop yield. The dataset comprises information such as the Area (in hectares) of land used for cultivation, the Pesticide (in tons) and Fertilizer (in tons) applied, and the Annual Rainfall (in mm) recorded for the year. Additionally, it includes the Crop Type (e.g., wheat, rice, sugarcane), the Season (e.g., Kharif, Rabi) in which the crop is planted, and the State where the farm is located.

Data Preprocessing:

Data Preprocessing data preprocessing plays a crucial role in ensuring that the machine learning model can accurately predict crop yield based on various agricultural features. The preprocessing begins with handling missing data and performing imputation if necessary. Numerical features, such as Area, Pesticide Usage, Fertilizer Usage, and Annual Rainfall, are analyzed for skewness, and log transformations are applied to normalize their distributions.

Model Selection:

Model Selection Random forest is a commonly-used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. In this project, Random Forest can be highly effective because it can model complex relationships between various agricultural factors (such as area, rainfall, fertilizer use, etc.) and crop yield. By leveraging its ability to handle both categorical and numerical data, along with its robust performance against overfitting, Random Forest is a suitable algorithm for this type of task

Model Training:

Model Training Model training for the Crop Yield Prediction project involves using historical agricultural data to develop a predictive model capable of estimating crop yields based on various factors such as soil characteristics, climate conditions, and agricultural practices. Initially, the dataset is preprocessed to handle missing values and normalize features, ensuring quality inputs for the model. Several machine learning algorithms, including Random Forest, XGBoost, and Linear Regression, are employed to determine the best-performing model for yield prediction. The training process involves splitting the data into training and testing sets, with the model being trained on the training set while the performance is evaluated on the testing set using metrics such as Mean Squared Error (MSE) and R-squared.

Model Evaluation:

Model evaluation in the Crop Yield Prediction project is a critical step that assesses the performance and reliability of the predictive model. After training, the model is tested using a separate validation dataset to ensure that it generalizes well to unseen data. Key evaluation metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2), are calculated to quantify the model's accuracy and predictive capabilities. MAE and MSE provide insights into the average errors in yield predictions, while R^2 indicates the proportion of variance in crop yield explained by the model. By analyzing these metrics, adjustments and optimizations can be made, ensuring the model is both accurate and effective in providing actionable insights for agricultural decision-making. and MSE provide insights into the average errors in yield predictions, while R^2 indicates the proportion of variance in crop yield explained by the model. By analyzing these metrics, adjustments and optimizations can be made, ensuring the model is both accurate and effective in providing actionable insights for agricultural decision-making

6.Comparison of algorithm

In crop yield prediction, several machine learning algorithms have been utilized, including Decision Trees, XGBoost, Linear Regression, and Random Forest. Decision Trees, while easy to interpret and capable of handling both categorical and numerical data, tend to overfit and perform poorly on complex, large datasets due to their sensitivity to data variations. XGBoost, an advanced gradient boosting technique, offers better performance and generalization, especially with non-linear data, and is highly efficient due to its parallel processing capability. However, XGBoost is computationally intensive and requires extensive hyperparameter tuning, which can be time-consuming and complex for large-scale agricultural datasets. Linear Regression, although simple and interpretable, struggles with non-linear relationships between environmental factors and crop yields, leading to lower accuracy in predictions. In contrast, Random Forest, an ensemble method of decision trees, provides superior performance by reducing overfitting and improving generalization. It can capture complex interactions between multiple variables such as weather, soil, and crop type, making it more robust and accurate for crop yield prediction. Additionally, Random Forest is highly scalable, can handle large datasets efficiently, and offers feature importance insights, making it the best choice for this application.

COMPARISON OF ALGORITHM

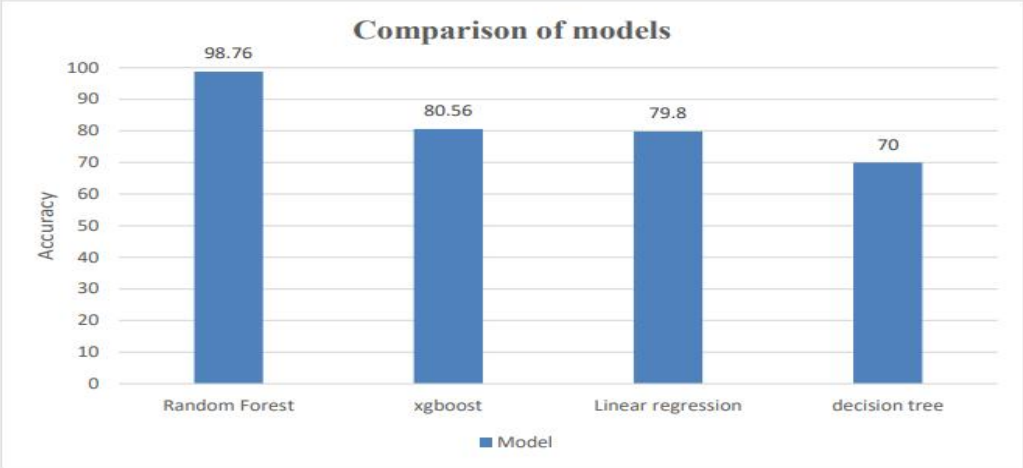

Figure.2 - Comparison of models

7.Execution

The Crop Yield Prediction Using Machine Learning system is designed to provide accurate yield forecasts based on various agricultural inputs entered through a user-friendly graphical interface (GUI). Users input critical parameters such as the cultivated area in hectares, production in tons, annual rainfall in millimeters, fertilizer and pesticide usage in kilograms,weather,State and the specific crop type. These inputs are processed by machine learning models, which analyze the data to predict the expected crop yield. The system leverages algorithms such as Random Forest, Naïve Bayes, Logistic Regression, and Support Vector Machines to ensure robust predictions. This tool can greatly assist farmers in making informed decisions regarding resource allocation, crop selection, and maximizing productivity based on current and historical agricultural data.

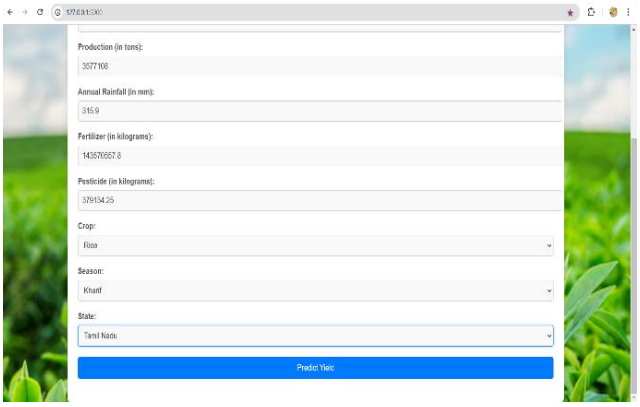

Figure.3-Inputpage

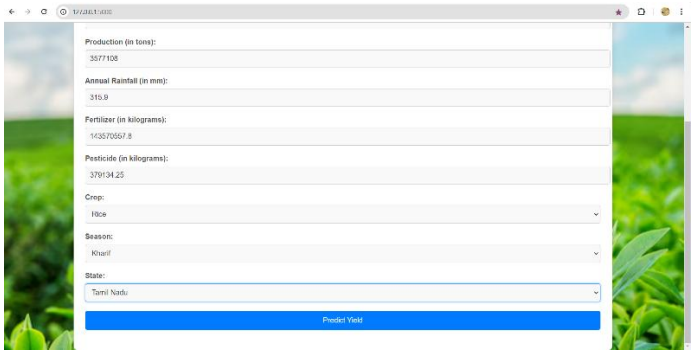

Figure.4-Inputpage2

	A	B	C	D	E	F	G	H	I	J	K	L
1	Crop	Crop_Year	Season	State	Area	Productior	Annual_Ra	Fertilizer	Pesticide	Yield		
2	Arecanut	1997	Whole Yea	Assam	73814	56708	2051.4	7024878	22882.34	0.796087		
3	Arhar/Tur	1997	Kharif	Assam	6637	4685	2051.4	631643.3	2057.47	0.710435		
4	Castor see	1997	Kharif	Assam	796	22	2051.4	75755.32	246.76	0.238333		
5	Coconut	1997	Whole Yea	Assam	19656	1.27E+08	2051.4	1870662	6093.36	5238.052		
6	Cotton(lint	1997	Kharif	Assam	1739	794	2051.4	165500.6	539.09	0.420909		
7	Dry chillies	1997	Whole Yea	Assam	13587	9073	2051.4	1293075	4211.97	0.643636		
8	Gram	1997	Rabi	Assam	2979	1507	2051.4	283511.4	923.49	0.465455		
9	Jute	1997	Kharif	Assam	94520	904095	2051.4	8995468	29301.2	9.919565		
10	Linseed	1997	Rabi	Assam	10098	5158	2051.4	961026.7	3130.38	0.461364		
11	Maize	1997	Kharif	Assam	19216	14721	2051.4	1828787	5956.96	0.615652		
12	Mesta	1997	Kharif	Assam	5915	29003	2051.4	562930.6	1833.65	4.568947		
13	Niger seed	1997	Whole Yea	Assam	9914	5076	2051.4	943515.4	3073.34	0.482353		
14	Onion	1997	Whole Yea	Assam	7832	17943	2051.4	745371.4	2427.92	2.342609		
15	Other Rab	1997	Rabi	Assam	108297	58272	2051.4	10306625	33572.07	0.52087		
16	Potato	1997	Whole Yea	Assam	75259	671871	2051.4	7162399	23330.29	7.561304		
17	Rapeseed	1997	Rabi	Assam	279292	154772	2051.4	26580220	86580.52	0.554783		
18	Rice	1997	Autumn	Assam	607358	398311	2051.4	57802261	188281	0.78087		
19	Rice	1997	Summer	Assam	174974	209623	2051.4	16652276	54241.94	1.060435		
20	Rice	1997	Winter	Assam	1743321	1647296	2051.4	1.66E+08	540429.5	0.941304		
21	Sesamum	1997	Whole Yea	Assam	15765	8257	2051.4	1500355	4887.15	0.487391		
22	Small mille	1997	Kharif	Assam	10490	5391	2051.4	998333.3	3251.9	0.473		
23	Sugarcane	1997	Kharif	Assam	31318	1287451	2051.4	2980534	9708.58	41.89696		
24	Sweet pot	1997	Whole Yea	Assam	9380	32618	2051.4	892694.6	2907.8	3.440435		
25	Tapioca	1997	Whole Yea	Assam	2465	11728	2051.4	234594.1	764.15	4.418261		
26	Tobacco	1997	Whole Yea	Assam	433	26	2051.4	41208.61	134.23	0.38		

crop_yield

Figure.5-Dataset

8.Result

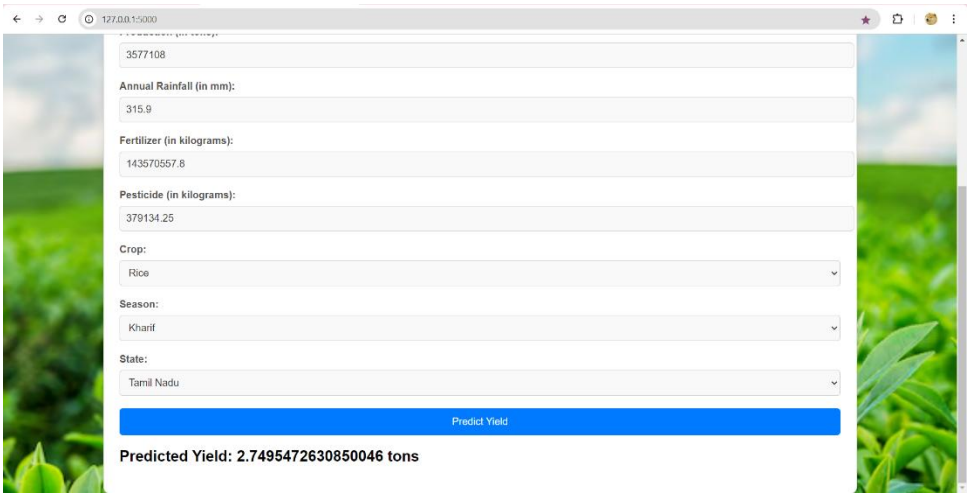

Figure.6 Output page

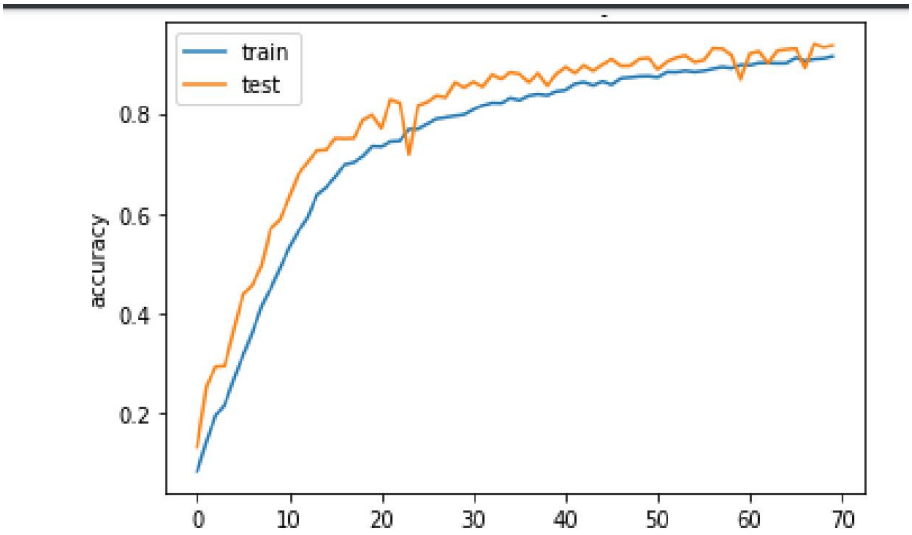

Figure.7Accuracy of train and test data

9. Conclusion

The Crop Yield Prediction project successfully demonstrates the potential of data-driven methods to address agricultural challenges. By leveraging machine learning models, such as Random Forest, combined with preprocessing techniques like log transformation and standard scaling, this project provides a robust solution for predicting crop yield based on key features like area, production, annual rainfall, fertilizer, and pesticide usage. The user-friendly interface allows for easy input of data and provides timely yield predictions, which can assist farmers and agricultural planners in making informed decisions. Despite the absence of a current database integration, the project remains scalable, and future iterations can seamlessly incorporate a database for enhanced data storage and retrieval. Additionally, the project is designed to be adaptable to different regions and crops, enabling broad application in diverse agricultural environments.

10. References

[1] Gandge, Y., 2017, December. A study on various data mining techniques for crop yield prediction. In 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT) (pp. 420-423). IEEE.

[2] Kumar, R., Singh, M.P., Kumar, P. and Singh, J.P., 2015, May. Crop Selection Method to maximize crop yield rate using machine learning technique. In 2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM) (pp. 138-145). IEEE.

[3] Bhanumathi, S., Vineeth, M. and Rohit, N., 2019, April. Crop yield prediction and efficient use of fertilizers. In 2019 International Conference on Communication and Signal Processing (ICCSP) (pp. 0769-0773). IEEE.

[4] Suresh, A., Kumar, P.G. and Ramalatha, M., 2018, October. Prediction of major crop yields of Tamilnadu using K-means and Modified KNN. In 2018 3rd International Conference on Communication and Electronics Systems (ICCES) (pp. 88-93). IEEE.

[5] Singhatiya, S. . and Ghosh, D. S. . (2018) "A Review on Soil Property Detection using Machine Learning Approach", SMART MOVES JOURNAL IJOSCIENCE, 4(8), pp. 6–9. doi: 10.24113/ijoscience.v4i8.152.

[6] Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L. and Mouazen, A.M., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Computers and electronics in agriculture, 121, pp.57-65

[7] Paul, M., Vishwakarma, S.K. and Verma, A., 2015, December. Analysis of soil behaviour and prediction of crop yield using data mining approach. In 2015 International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 766-771). IEEE

[8] Vijayabaskar, P.S., Sreemathi, R. and Keertanaa, E., 2017, March. Crop prediction using predictive analytics. In 2017 International Conference on Computation of Power, Energy Information and Commuincation .(ICCPEIC) (pp. 370-373). IEEE.

[9] Mariappan, A.K. and Das, J.A.B., 2017, April. A paradigm for rice yield prediction in Tamilnadu. In 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 18-21). IEEE.

[10] Sujatha, R. and Isakki, P., 2016, January. A study on crop yield forecasting using classification techniques. In 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16) (pp. 1-4). IEEE.

[11] Fathima, K. Sowmya, S. Barker, S. Kulkarni, "Analysis of crop yield prediction using data mining technique", International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 05, May 2020.

[12] Dahikar, S.S. and Rode, S.V., 2014. Agricultural crop yield prediction using artificial neural network approach. International journal of innovative research in electrical, electronics, instrumentation and control engineering, 2(1), pp.683-686.

[13] Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L. and Mouazen, A.M., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Computers and electronics in agriculture, 121, pp.57-65.

[14] Kim, Y.H., Yoo, S.J., Gu, Y.H., Lim, J.H., Han, D. and Baik, S.W., 2014. Crop pests prediction method using regression and machine learning technology: Survey. IERI Procedia, 6, pp.52-56.

[15] Shakoor, M.T., Rahman, K., Rayta, S.N. and Chakrabarty, A., 2017, July. Agricultural production output prediction using supervised machine learning techniques. In 2017 1st international conference on next generation computing applications (NextComp) (pp. 182-187). IEEE.

[16] Veenadhari, S., Misra, B. and Singh, C.D., 2014, January. Machine learning approach for forecasting crop yield based on climatic parameters. In 2014 International Conference on Computer Communication and Informatics (pp. 1-5). IEEE.

[17] Mishra, S., Mishra, D. and Santra, G.H., 2016. Applications of machine learning techniques in agricultural crop production: a review paper. Indian Journal of Science and Technology, 9(38), pp.1-14.

[18] Rajak, R.K., Pawar, A., Pendke, M., Shinde, P., Rathod, S. and Devare, A., 2017. Crop recommendation system to maximize crop yield using machine learning technique. International Research Journal of Engineering and Technology, 4(12), pp.950-953.

[19] Nigam, A., Garg, S., Agrawal, A. and Agrawal, P., 2019, November. Crop yield prediction using machine learning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 125-130). IEEE.

[20] Liakos, K.G., Busato, P., Moshou, D., Pearson, S. and Bochtis, D., 2018. Machine learning in agriculture: A review. Sensors, 18(8), p.2674.

[21] V. Geetha, A. Punitha, M. Abarna, M. Akshaya, S. Illakiya and A. P. Janani, ”An Effective Crop Prediction Using Random Forest Algorithm,” 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 2020, pp. 1-5.

[22] L. M. Gladence, K. R. Reddy, M. P. Reddy, M. P. Selvan and Refonaa, ”A Prediction of Crop Yield using Machine Learning Algorithm,” 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021, pp. 1072-1077.