ENHANCED CROP YIELD FORECASTING WITH MACHINE LEARNING AND DEEP LEARNING APPROACHES

N.Pavithra[1] R.Ritesh T[2] G.Pradeep[3] M.Prakash Raj[4]
[1] Faculty, Department of IT, Jeppiaar Institute of Technology, Kunnam, Chennai, TN- 631604.
[2,3,4]Student, Department of IT, Jeppiaar Institute of Technology, Kunnam, Chennai, TN- 631604.

ABSTRACT: Agriculture is a cornerstone of the global economy and essential for ensuring food security. Accurate crop yield prediction is vital for boosting agricultural productivity and improving resource management. This project, titled Machine Learning-Based Crop Yield Prediction Using Regression and Deep Learning Techniques, applies advanced computational methods to estimate crop yields by analyzing environmental and soil data. Key parameters such as temperature, humidity, and soil moisture are utilized to forecast yields with precision. The project employs the XGBoost algorithm, known for its efficiency and robustness when handling complex datasets. By integrating regression methods with deep learning models, the system enhances predictive accuracy. The approach includes steps like data preprocessing, feature selection, and model training to identify non-linear relationships in the data effectively. Deep learning enables automated feature extraction, reducing the reliance on manual input and making the model adaptable to a wide range of agricultural conditions. This system provides valuable support to farmers, policymakers, and researchers by delivering reliable yield predictions. Its scalability makes it applicable to diverse regions and various crop types. Moreover, the project promotes sustainable agriculture by improving resource planning, minimizing risks, and optimizing crop management strategies. Results demonstrate that combining machine learning and deep learning techniques outperforms traditional methods, yielding superior accuracy. Future developments aim to incorporate real-time IoT data and explore advanced algorithms to further refine the system's performance. This innovative approach bridges technology and agriculture, paving the way for more informed decision-making and sustainable practices.

Keywords: Agriculture yield prediction, machine learning, regression, deep learning, XGBoost, temperature, humidity, soil moisture, sustainable agriculture, data analysis.

I.INTRODUCTION`

Agriculture is the backbone of many economies and a fundamental Agriculture serves as a foundation for global food security, and with the ever-growing population, the demand for higher agricultural productivity has become increasingly critical. Accurate crop yield prediction plays a vital role in addressing this demand by enabling efficient resource planning, risk management, and optimized crop production. Traditional yield prediction methods, which often rely on historical data and expert insights, can be labor-intensive and less responsive to dynamic environmental changes. This project, titled Machine Learning and Deep Learning-Based Crop Yield Prediction, seeks to modernize the prediction process by combining advanced computational techniques with critical environmental and soil data. Key factors such as temperature, humidity, and soil moisture, which are crucial determinants of crop health and yield, are analyzed to forecast productivity effectively. The XGBoost algorithm, a robust and efficient machine learning method, is employed to process and analyze the collected data. Renowned for its ability to handle large datasets and capture complex relationships, XGBoost enhances the accuracy and reliability of yield predictions. By integrating these modern technologies, the project aims to provide innovative and scalable solutions that can adapt to diverse agricultural conditions, ultimately contributing to sustainable farming practices and global food security.

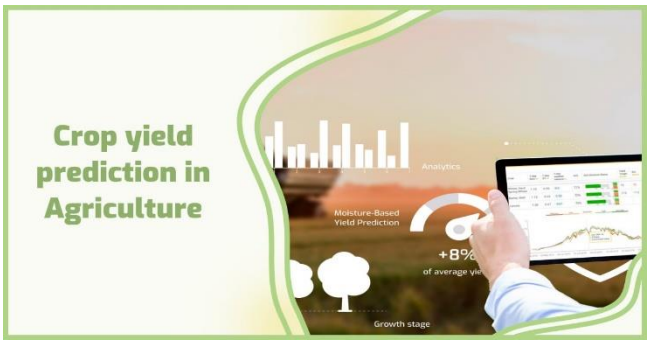

Figure 1. Crop yield prediction

The integration of regression models and deep learning significantly enhances the predictive accuracy of the system. Regression techniques serve as a solid framework for analyzing both linear and non-linear relationships between key variables, while deep learning models streamline the process by automatically extracting relevant features. This combination creates a versatile and powerful framework capable of adapting to the complexities and diversity of agricultural scenarios. This cutting-edge system has far-reaching implications for sustainable agriculture. It provides farmers with precise yield forecasts, empowering them to make data-driven

decisions regarding planting schedules, irrigation strategies, and fertilization practices. Additionally, it supports policymakers in formulating effective plans for food security and resource distribution.

II.RELATED WORK

[1] P. Sharma et al. (2023) proposed a hybrid version combining regression and deep getting to know strategies for predicting agricultural yields. Using environmental statistics inclusive of temperature, humidity, and soil moisture, they implemented device gaining knowledge of algorithms like XGBoost and deep neural networks. Their findings showed that this approach provided a substantial development in prediction accuracy compared to standard models.

[2] Kavita Jhajharia et al. (2023) implemented a combination of machine gaining knowledge of and deep gaining knowledge of strategies to expect crop yields, focusing on algorithms like Support Vector Machines (SVM), Random Forest, and Long Short-Term Memory (LSTM) networks. Their take a look at emphasised the significance of characteristic engineering and integrating a couple of records sources, inclusive of climate forecasts, soil situations, and satellite tv for pc imagery. The authors established that incorporating numerous records stepped forward model performance, improving the accuracy of yield predictions.

[3] N.R. Prasad et al. (2021) used the Random Forest algorithm to expect cotton crop yields at a regional level. They focused on nearby factors, along with soil kind, nearby climate situations, and crop-particular traits, to beautify prediction accuracy. The authors demonstrated that Random Forest models are nicely-perfect for eventualities where facts availability is restricted or where local variations play a massive role in yield consequences. The have a look at found that incorporating these factors advanced model performance, offering treasured insights for nearby-scale crop management.

[4] Bali, N., and Singla, A. (2022) surveyed rising developments in device getting to know for crop yield prediction, inspecting influential elements along with climate, soil high-quality, and pest incidence. They highlighted the combination of environmental information, satellite tv for pc imagery, and system learning algorithms for improved yield forecasting. The take a look at emphasised the need for actual-time information to enhance prediction fashions and decrease uncertainties in crop control.

[5] Na Liu et al. (2020) delivered an ensemble mastering technique to handle imbalanced data in crop yield prediction. This approach is in particular beneficial in agricultural programs, where extreme activities like droughts or pest infestations can dramatically have an effect on crop yields however are underrepresented in datasets. The authors tested that ensemble techniques, which combine a couple of models, can improve prediction accuracy despite imbalanced records.

[6] Snehal S. Dahikar and Sandeep V. Rode (2014) explored using artificial neural networks (ANN) for predicting agricultural crop yields. Their take a look at confirmed that neural networks should version complicated relationships in crop yield information through learning from historical weather statistics and crop-precise characteristics. The authors located that ANN models done properly in forecasting crop yields, mainly in instances wherein information relationships were nonlinear.

[7] P.S. Maya Gopal and R. Bhargavi (2019) proposed an revolutionary approach to crop yield prediction the use of system learning algorithms, together with Random Forest and Gradient Boosting. They integrated sensor facts from the sector, inclusive of soil moisture and temperature, to enhance prediction accuracy. The authors argued that actual-time sensor records, while combined with system gaining knowledge of fashions, provides a extra dynamic and correct forecast of crop yields.

[8] Hayam R. Seireg et al. (2022) implemented ensemble device gaining knowledge of strategies to are expecting wild blueberry yields using simulation facts. Their take a look at highlighted the effectiveness of mixing multiple fashions to improve yield predictions, mainly for vegetation with variable developing situations. The authors established that ensemble methods outperformed person gadget learning models in terms of prediction accuracy and generalization.

[9] Priyanga Muruganantham et al. (2022) carried out a systematic overview on using deep getting to know and remote sensing for crop yield prediction. The authors highlighted how deep mastering fashions, mainly CNNs and RNNs, can extract spatial and temporal capabilities from satellite tv for pc imagery and other environmental statistics. Their evaluate demonstrated the effectiveness of deep mastering strategies in coping with huge, unstructured datasets, permitting extra accurate crop yield predictions.

[10] Khaki Saeed et al. (2020) proposed a hybrid CNN-RNN framework for crop yield prediction, combining convolutional neural networks (CNN) for feature extraction from satellite images and recurrent neural networks (RNN) for modeling temporal relationships in crop facts. Their take a look at showed that this hybrid technique outperformed traditional device studying fashions in predicting crop yields, especially for vegetation like wheat and corn.

III.PROPOSED SYSTEM

The proposed tool for predicting agricultural yields uses system learning (ML) and deep studying (DL) strategies to offer accurate forecasts primarily based totally on environmental factors along with temperature, humidity, and soil moisture. The purpose is to create a robust prediction model that facilitates farmers, researchers, and policymakers in making information-pushed selections for crop management, ultimately enhancing agricultural productiveness and sustainability. Central to the device is using the XGBoost set of guidelines, a powerful and efficient ML version acknowledged for

its excessive ordinary overall performance in predictive responsibilities. XGBoost excels at coping with big datasets, coping with missing values, and offering characteristic significance scores, even as keeping accuracy and speed. The algorithm operates on a gradient boosting framework, which mixes multiple inclined models to create a greater correct and stronger prediction, making it specifically powerful for complicated, non-linear relationships between the enter skills and the target variable—crop yield. [8]

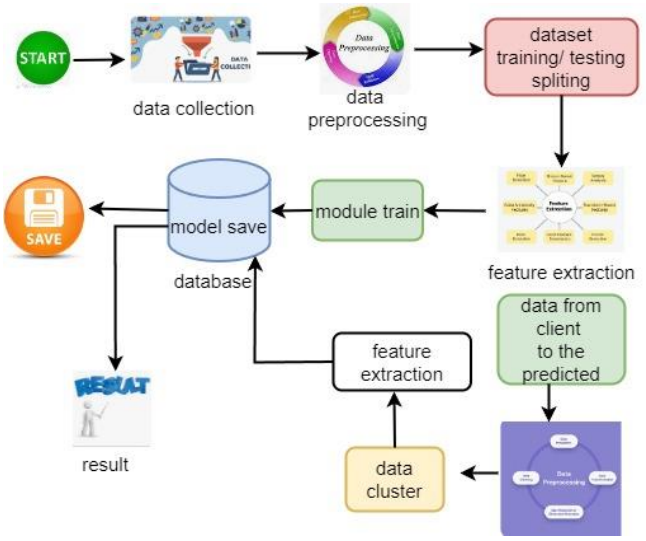

Figure 2. System Architecture Diagram

This ensures that the facts is each contemporary and applicable for analysis. The accumulated information is then cleaned, normalized, and preprocessed to make it appropriate for education device mastering fashions. Missing values are addressed, and outliers are eliminated to make certain the great and consistency of the dataset. Next, applicable functions are extracted, with extra functions like temperature and humidity developments or moisture content variations considered for deeper insights into crop growth conditions. The XGBoost version is then trained the use of this preprocessed information to predict agricultural yield. The version's accuracy is evaluated the usage of metrics which includes suggest squared errors (MSE) and R-squared, making sure that the predictions are each reliable and specific. Once educated, the model is deployed for real-time yield prediction. It constantly gets sparkling environmental data, bearing in mind ongoing yield forecasts. This allows farmers to take proactive movements, inclusive of adjusting irrigation schedules or selecting the satisfactory planting instances, based on anticipated consequences. The gadget's scalability lets in it to be tailored to various crop kinds and regions, and future improvements might also comprise extra environmental elements or integrate IoT technology for more dynamic, actual-time information collection. Additionally, exploring deep learning fashions could in addition enhance the system's predictive accuracy by shooting more complicated styles in the statistics. [6]

IV.METHODOLOGY

A. Dataset Collection

The first step within the technique includes collecting applicable environmental facts, that's vital for ensuring the version's accuracy. The dataset used for predicting agricultural yields usually includes variables including temperature, humidity, soil moisture, and other environmental factors that have an impact on crop boom. These statistics can be sourced from weather stations, satellite tv for pc information, IoT sensors, and agricultural databases. By incorporating each historic and real-time records, the version can account for seasonal versions and precise environmental situations. Ensuring that the dataset is comprehensive, updated, and unfastened from inconsistencies is critical, because it directly influences the accuracy of the predictions, main to higher insights for agricultural management selections. [8]

B. Data Pre-processing

After information series, pre-processing ensures the dataset is easy, dependent, and appropriate for education the version. Raw facts regularly consist of missing values, mistakes, or inconsistencies which can negatively affect version overall performance. Techniques including suggest imputation, ahead filling, or disposing of outliers are applied to smooth the facts. Additionally, specific variables like soil kind or crop range might also want to be encoded into numerical values. Feature scaling techniques, such as Min-Max scaling or Z-score standardization, make sure that no single variable dominates the model, selling balanced gaining knowledge of. Proper facts pre-processing improves the model's potential to generalize, resulting in more accurate predictions of crop yields.[13]

C. Model Selection and Training

In this step, the model for predicting agricultural yields is chosen based on its capacity to deal with complicated records and offer excessive accuracy. XGBoost is selected for its performance, flexibility, and functionality to address large datasets with non-linear relationships. It makes use of a gradient boosting framework, where a couple of vulnerable models combine to create a sturdy predictor.

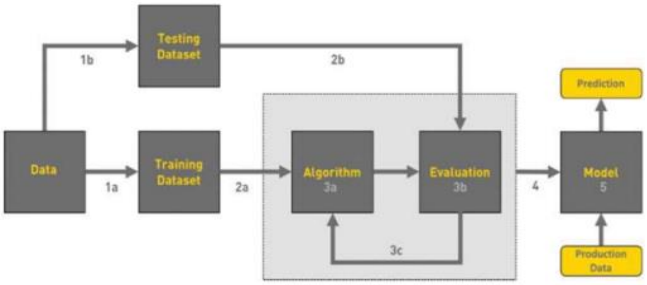

Figure 3. Machine learning Model

D. Hyperparameter Tuning

Hyperparameter tuning is a vital step for optimizing the version's performance. XGBoost offers several hyperparameters, which includes learning charge, number of estimators, tree depth, and subsample ratio, which affect the education procedure. This step includes selecting the exceptional aggregate of hyperparameters to decrease version errors and beautify predictive accuracy. Methods consisting of Grid Search and Random Search are normally used for hyperparameter optimization. These strategies systematically explore a number possible values, figuring out the most appropriate configuration. By great-tuning these parameters, the version can be adjusted to in shape the statistics extra correctly, improving its ability to are expecting crop yields below various environmental situations. [15]

E. Feature Importance Analysis

Feature importance analysis is an critical a part of knowledge the factors riding the version's predictions. XGBoost offers a function importance score for each variable, indicating how a great deal have an effect on every environmental element—inclusive of temperature, humidity, or soil moisture—has at the crop yield prediction. This analysis enables prioritize the largest capabilities and lets in farmers and researchers to attention at the variables that most have an effect on crop increase. By visualizing the significance of various capabilities via charts or graphs, stakeholders can better recognize which environmental conditions are most crucial for optimizing yield results. This perception aids in enhancing agricultural practices and useful resource allocation. [5]

F. Deployment and Integration

After the model is educated and evaluated, it is deployed for real-time prediction. The model is incorporated right into a person-pleasant system that allows farmers or agricultural managers to enter modern-day environmental facts, such as temperature, humidity, and soil moisture, to receive yield predictions. The device techniques the statistics and gives recommendations for most fulfilling crop management practices, which include irrigation schedules, planting times, or fertilizer utilization. Deployment guarantees that the model's predictions are actionable in a practical context, without delay benefiting users by using supporting them make informed choices to enhance crop productiveness and mitigate environmental risks. Real-time integration improves choice-making performance in agriculture.

G. Continuous Improvement

To make certain that the model stays powerful over the years, continuous improvement is essential. As new facts turns into available, the device undergoes regular retraining to alter to shifts in environmental situations, farming practices, or crop sorts. This non-stop learning procedure allows the version live applicable and adaptive. Feedback from real-international effects, including variations between predicted and real yields, is gathered to refine the model. The comments loop lets in for the identification and correction of any biases or inaccuracies, improving future predictions. Incorporating person comments and new environmental factors similarly complements the gadget, making sure that predictions stay correct and aligned with evolving agricultural trends. [4]

V.RESULT AND DISCUSSION

A. Dataset

The dataset used on these studies consists of mortgage software statistics from an online lending platform, masking six years (from September 3, 2012, to September 4, 2018). Originally containing fifty-four fields, many have been excluded due to a excessive range of lacking values, resulting within the selection of 35 key variables for analysis. These selected fields consist of critical info consisting of loan amounts, applicant incomes, credit ratings, mortgage terms, and approval results. Sensitive data, inclusive of telephone numbers and identity details, has been anonymized to guard privateness. The facts statistics are classified to suggest whether each mortgage utility become permitted or rejected. Several system getting to know techniques have been carried out to are expecting loan approvals, with model performance evaluated the use of metrics. These measures provide a complete knowledge of the models' ability to predict loan outcomes successfully, supporting discover the maximum accurate version for selection-making. [8]

	N	P	K	temperature	humidity	ph	rainfall	Encoded_label
count	2100.000000	2100.000000	2100.000000	2100.000000	2100.000000	2100.000000	2100.000000	2100.000000
mean	48.181905	55.060000	48.057143	25.471157	70.488397	6.474750	107.214784	10.000000
std	36.016740	32.772902	51.833864	5.134671	22.303929	0.790175	53.425191	6.056743
min	0.000000	5.000000	5.000000	8.825675	14.258040	3.504752	30.920140	0.000000
25%	20.000000	31.000000	20.000000	22.602437	59.237124	5.948298	67.133894	5.000000
50%	36.000000	53.000000	30.000000	25.373357	79.995776	6.431478	97.659622	10.000000
75%	80.250000	69.000000	47.000000	28.282950	88.018773	6.945725	130.140535	15.000000
max	140.000000	145.000000	205.000000	43.675493	99.981876	9.935091	298.560117	20.000000

Figure 4. Input dataset of the visualization

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is performed to uncover crucial patterns and relationships in the mortgage software statistics. Through EDA, new functions are generated, together with debt-to-income ratios, that can enhance version overall performance. A gender-based evaluation is completed to explore how gender impacts loan approval charges, mortgage sizes, and reimbursement behaviors. Visual tools along with box plots are used to examine the distribution of mortgage amounts throughout different marital status categories, revealing patterns in the quantity of mortgage asked via various groups.[3] Correlation analysis highlights crucial relationships, consisting of the strong correlation between earnings and credit rating, as well as between income and loan quantities. These findings underscore the significance of creditworthiness within the mortgage approval method and the tremendous position of credit scores in predicting loan eligibility.

C. Machine Learning Model

To are expecting loan approvals, system learning algorithms which include Random Forest and Gradient Boosting are employed. Random Forest works by developing more than one decision bushes and mixing their predictions, assisting to reduce overfitting and improve version accuracy. Gradient Boosting, then again, builds fashions sequentially, where each subsequent model corrects mistakes made through the previous ones, enhancing the prediction technique through iterative boosting.

	Precision	Recall	F1-Score	Support
0	0.90	0.85	0.87	100
1	0.82	0.88	0.85	100
Accuracy			0.86	200
Macro Average	0.86	0.86	0.86	200
Weighted Average	0.86	0.86	0.86	200

Figure 5. Report of the classification

Both techniques are properly-appropriate for managing large and complex datasets, together with loan software statistics, and offer sturdy predictions which can inform lending decisions and stumble on ability fraud. These techniques are vital for improving the accuracy of loan approval predictions in dynamic monetary environments.

D. Accuracy Metrics and Evaluation

The performance of the loan prediction model is assessed using metrics such as accuracy, precision, recall, and F1 score. These metrics evaluate the model's effectiveness in correctly identifying loan approvals and fraudulent applications. Accuracy is calculated using the following formula:

$$\text{Accuracy} = TP+TN/\ TP+TN+FP+FN$$

Where:

- True Positives (TP) refer to correctly identified fraudulent loans.
- True Negatives (TN) refer to correctly identified legitimate loans.
- False Positives (FP) refer to legitimate loans incorrectly flagged as fraudulent.
- False Negatives (FN) refer to fraudulent loans incorrectly classified as legitimate.[2]

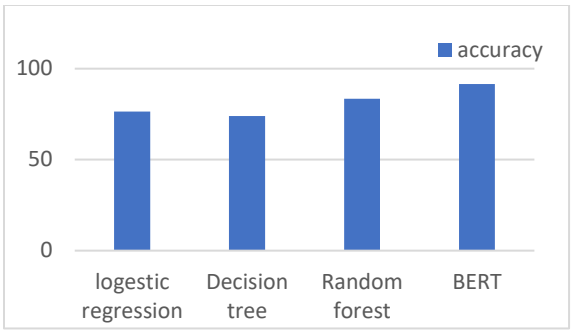

Figure 6. Accuracy comparison Chart

These metrics provide valuable insights into how well the model performs in predicting loan outcomes and detecting fraud. High accuracy, along with a balanced precision-recall tradeoff, ensures that the model makes reliable predictions, minimizing false positives and false negatives, which is crucial for maintaining trust and efficiency in loan approval processes.

Model	**accuracy**	**Precision**	**recall**
Logistic Regression	76.5%	0.75	0.76
Decision Trees	74.0%	0.73	0.74
Random Forest	83.0%	**0.82**	0.82
BERT	91.5%	0.84	0.86

Table 1. Comparison table

Linear Regression Equation

Linear regression can serve as a baseline model for yield prediction:

$$Y=\beta\ 0 +\beta\ 1\ X\ 1 +\beta\ 2\ X\ 2 +\beta\ 3\ X\ 3 +\epsilon$$

Where:

- Y: Predicted agricultural yield
- β 0 : Intercept (constant term)
- β 1 ,β 2 ,β 3 : Coefficients for temperature (X 1), humidity (X 2), and soil moisture (X3) respectively
- ϵ: Error term

Gradient Boosting Loss Function

XGBoost minimizes the loss function for regression, typically the Mean Squared Error (MSE):

$$\text{Loss} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

VI.CONCLUSION

In summary, integrating machine learning models like Random Forest and Gradient Boosting into the loan approval process has demonstrated significant improvements in decision-making within the financial sector. By leveraging comprehensive datasets and

conducting thorough exploratory data analysis, key factors influencing loan approval—such as income, credit score, and loan amounts—have been identified and analyzed. Transforming categorical variables and creating new features, such as debt-to-income ratios, have further enhanced the models' predictive capabilities.[15] Evaluating model performance through metrics like accuracy, precision, recall, and F1 score has highlighted the robustness of these algorithms in accurately predicting loan outcomes and detecting potential fraud. This approach not only improves the accuracy of loan approval predictions but also mitigates risks associated with lending by identifying potentially fraudulent applications. However, it is crucial to recognize that continuous monitoring and periodic retraining of these models are necessary to adapt to evolving economic conditions and borrower behaviors. This ensures that the models remain reliable and effective over time.[14]

VII.FUTURE ENHANCEMENT

Advancements in machine learning (ML) offer promising avenues for enhancing loan approval prediction systems. Integrating advanced algorithms such as CatBoost and LightGBM can improve predictive accuracy by effectively handling categorical data and large datasets. Additionally, incorporating synthetic feature generation techniques can uncover complex patterns, further enhancing model performance. To address the issue of missing data, especially in rejected loan applications, implementing models like the Reject-aware Multi-Task Network (RMT-Net) can provide more reliable predictions by considering both approved and rejected cases. Incorporating explainable AI (XAI) methods can enhance transparency in decision-making processes, allowing stakeholders to understand and trust the model's predictions. This is particularly important in financial sectors where regulatory compliance and ethical considerations are paramount. Furthermore, integrating real-time data streams, such as economic indicators and social media sentiment, can provide a more dynamic and current assessment of loan applicants, leading to more accurate and timely decisions.

REFERENCES

[1] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin and N. Khan, "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction", IEEE Access, vol. 9, pp. 63406-63439, 2021.

[2] P. Sharma, P. Dadheech, N. Aneja and S. Aneja, "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning", IEEE Access, vol. Eleven, pp. 111255-111264, 2023.

[3] Dan Li, Yuxin Miao, k. Sanjay Gupta, J. Carl Rosen, Fei Yuan, Chongyang Wang, Li Wang and Yanbo Huang, "Improving Potato Yield Prediction by way of Combining Cultivar Information and UAV Remote Sensing Data Using Machine Learning", Remote Sensing, vol. 13, pp. 3322-3344, 2021.

[4] Mohsen Shahhosseini, Guiping Hu, Sotirios V. Archontoulis, "Forecasting Corn Yield with Machine Learning Ensembles", Frontiers in Plant Science, vol. Eleven, 2020.

[5] R. S. Renju, P. S. Deepthi and M. T. Chitra, "A Review of Crop Yield Prediction Strategies based totally on Machine Learning and Deep Learning," International Conference on Computing, Communication, Security andIntelligent Systems (IC3SIS), Kochi, India, pp. 1-6,2022.

[6] Kavita Jhajharia, Pratistha Mathur, Sanchit Jain, Sukriti Nijhawan, "Crop Yield Prediction the use of Machine Learning and Deep Learning Techniques", Procedia Computer Science, Volume 218, Pages 406- 417,2023.

[7] S Iniyan ,V Akhil Varma , Ch Teja Naidu ," Crop yield prediction the use of gadget studying techniques",Advances in Engineering Software, extent one hundred seventy five, pages 103326, 2023.

[8] N.R. Prasad, N.R. Patel, Abhishek Danodia, "A. Crop yield prediction in cotton for regional level the usage of random wooded area approach", Spat. Inf. Res, vol. 29, pp. 195 - 206, 2021.

[9] Bali, N., Singla, A. "Emerging Trends in Machine Learning to Predict Crop Yield and Study Its Influential Factors: A Survey". Arch Computat Methods Eng 29, 95–112 ,2022.

[10] Na Liu , Xiaomei Li , Ershi Qi , Man Xu , Ling Li , AND Bo Gao," A Novel Ensemble Learning Paradigm for Diagnosis With Imbalanced Data" in IEEE Access,vol 253,2020.

[11] Snehal S.Dahikar , Sandeep V.Rode," Agricultural Crop Yield Prediction Using Artificial Neural Network Approach", International Journal of Innovative studies in Electrical, Electronics, Instrumentation and Control Engineering,Vol 2, pp 683-686, January 2014.

[12] P.S. Maya Gopal, R. Bhargavi, "A novel approach for efficient crop yield prediction", Computers and Electronics in Agriculture, Volume one hundred sixty five, p 104968, 2019,

[13] Hayam R. Seireg, Yasser M. K. Omar, Fathi E. Abd El-Samie, Adel S. El-Fishaw, Ahmed Elmahalawy," Ensemble Machine Learning Techniques Using Computer Simulation Data for Wild Blueberry Yield Prediction" in IEEE Access,vol 10, 64671-64687,2022.

[14] Priyanga Muruganantham, Wibowo, Santoso; Grandhi, Srimannarayana; Samrat, Nahidul; Islam, Nahina ,". A systematic overview on crop yield prediction with deep studying and faraway sensing." CQ University. Journal contribution, p 1-21, April 2022.

[15] Khaki Saeed , Wang Lizhi , Archontoulis Sotirios V." A CNN-RNN Framework for Crop Yield Prediction", Frontiers in Plant Science quantity 10,2020.