Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Prediction on Crop Yield on Indian based Agriculture using Machine Learning

Yuvasri K.[1], Nagarajan VR.[2]

[1]Student, [2]Associate Professor, Department of MCA, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Tiruchirappalli, India.

E-mail: [1]yuvasrik85@gmail.com, [2]nagarajan.set@dsuniversity.ac.in

Abstract

Crop yield prediction is an important component of modern agriculture that has a significant influence on resource management, policymaking, and food security. Based on data from several sources, including soil, climate, and crop characteristics, this model applies machine learning to develop a prediction algorithm that can analyze crop production in India. Data from government websites and educational resources about traditional agricultural practices was used to train and test the model. This proposed work provides extensive preprocessing, including normalization techniques like Min-Max scaling, filling in missing values, and extracting parameters from correlation and common data. This proposed work aims to use various regression models, such as linear regression, Decision Tree, Random Forest, and XGBoost regressor, based on performance standards like R^2 score, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). XGBoost performs effectively by conducting assessments because it can handle non-linear connections and prevent overfitting using boosting techniques. The model provides policymakers, agronomists, and farmers with valuable data that allows them to identify crops and locations where findings will increase agricultural yields under different climate conditions.

Keywords: Crop Yield Prediction, Machine Learning, Random Forest, XGBoost, Regression Algorithms, Precision Agriculture, Indian Agriculture, R^2 Score, Environmental Data, Data Preprocessing.

DOI: https://doi.org/10.36548/jucct.2025.2.007
Received: 07.05.2025, received in revised form: 11.06.2025, accepted: 03.07.2025, published: 18.07.2025

1. Introduction

Agriculture is a key part of India's socioeconomic landscape, with over 50% of the workforce employed and a major contributor to the country's GDP. Maintaining constant high-quality agricultural output has become a necessary feature in the face of climatic unpredictability, soil degradation, and population growth. One of the main challenges in this situation is predicting crop yield accurately for efficient farm planning, resource allocation, food distribution, and policy formulation. The traditional approaches to crop yield prediction mainly depend on heuristic models or rigid estimating techniques that are unable to capture the complex interactions between various factors, such as temperature, rainfall, crop type, soil conditions, and farming practices. These complicated, non-linear, multi-dimensional connections can be accurately modeled because of advancements in machine learning (ML), which yield data-driven, scalable, and accurate solutions. The project's aim is to develop a crop yield prediction system using machine learning based on previous agricultural data and environmental variables. Data from government websites, such as the Ministry of Agriculture and agricultural research institutes' datasets, are used for further processing. Different regression models, such as linear regression, decision trees, random forests, and XGBoost, are implemented to evaluate the effects of various features and predict crop yield across the regions of India.

1.1 Objectives

The primary objective of this work is to develop a machine learning-based system that will utilize previous agricultural and environmental data to predict crop production with a better accuracy rate. This research also includes the process of collecting and preprocessing datasets that contain essential data like temperature, rainfall, crop type, soil type, and regional type. This project will employ and compare a number of regression techniques, including XGBoost, Decision Tree, Random Forest, and linear regression, to evaluate changes and predict the production of different types of crops and their geographical areas. Each model is evaluated using performance metrics, including R^2 Score, Mean Squared Error (MSE), and Mean Absolute Error (MAE) to identify the effective prediction strategy. This method aims to produce useful inputs for farmers, agronomists, and policymakers enable them to make efficient planning and resource allocation decisions and offering proper crop prediction The

main goal is to increase agricultural sustainability and production through data-driven strategies that will reduce the risks related to climatically unpredictable variables.

1.2 Challenges in Traditional System

The statistical methods, manual calculations, or heuristics that are commonly used in traditional crop yield projection systems are usually inaccurate and unaffected by changing conditions. The complex nonlinear connections between many environmental and agricultural factors, such as temperature, rainfall, crop type, and soil health, are often overlooked. Additionally, previous research fails to handle large amounts of diverse data from several places and cannot adapt effectively. Crop planning, resource allocation, and food security are compromised by inaccurate predictions based on reliance on previous records and expectations of those research efforts. Real-time decision-making is not achieved with traditional methods because they cannot analyze and provide data in a timely manner. These approaches require extensive human involvement because they use minimal automation. These difficulties highlight the urgent need for automated and data-driven approaches, such as machine learning, to provide accurate and secure yield predictions. The major role of agriculture in India is explained in the introduction, along with the need for accurate crop yield prediction to ensure food security, optimize resource allocation, and support informed policymaking. Previous research yield prediction techniques are often incorrect because it is difficult to measure the complex connections between farming and environmental factors. Data-based models are used to analyze previous records, and the different elements affecting those records produce more accurate predictions due to advancements in machine learning techniques. This model proposes using regression techniques such as linear regression, decision trees, random forests, and XGBoost to predict crop yield based on variables such as rainfall, temperature, soil composition, and crop type. The technology will help users make informed agricultural decisions and advance sustainable agricultural practices by providing accurate data.

2. Literature Survey

Using ML algorithms, this research [1] showed that both field data and satellite imaging data can be used on a citrus farm to enhance production prediction. Their research highlights the importance of using diverse data sources to improve model ubiquity and

usability. This method was expanded in another research work [5] by adding vegetative indices (VIs) and actual evapotranspiration (AET) to ground data. Their research demonstrates the value of capturing spatiotemporal variation in wheat production prediction using environmental factors obtained from satellites. Extreme Gradient Boosting (XGBoost) and deep learning models were compared in this study [2]. They discovered that XGBoost offered equivalent accuracy as well as quicker training times, especially in low-data conditions, given DL approaches are data-driven. Using USDA datasets, this research [8] conducted a data-driven analysis with traditional machine learning algorithms. Based on their findings, feature engineering and data preparation are essential for improving ML model performance in yield prediction tasks. Convolutional neural networks (CNNs) are the main focus of this work [3] to estimate winter wheat production by analyzing phenological and environmental data. By identifying hierarchical patterns in time-series input, CNNs were able to increase durability across geographies and seasons. Similarly, a hybrid model called Graph Neural Network-Recurrent Neural Network (GNN-RNN) was suggested by this work [10], which utilized of both temporal and spatial connections in the data. For accurate predictions, their methodology highlighted the need for modeling time-series patterns and geographic autocorrelation. The proposed model [9] implements a hybrid ML-DL model that includes deep as well as extensive architectures with the aim of predicting agricultural yield. This method increased prediction accuracy across a variety of crop types classified using DL's abstraction skills and ML's generalization. This study [7] uses several machine learning models to build a combined smart system. Their model continually applies rule-based heuristics, constantly identifying the optimal prediction, providing a more flexible and accurate prediction pipeline for different data settings. To maximize performance in resource-constrained scenarios, this work [6] also used a hybrid learning approach that combined statistical learning with deep learning. Their methodology shows improved accessibility and scalability for implementation in particular rural areas. This research [4] customized their deep neural network (DNN) technique for crop choice and yield prediction o in Bangladesh to allow for localized agricultural practices. The adaptation of this model for increased value in certain geographic and socioeconomic situations is highlighted in their work.

2.1 Limitations of Existing Surveys

From the above studies, the use of freely available data was unable to accurately represent a variety of agricultural methods and environmental conditions. The performance

of the prediction models can be significantly affected by the condition of the data, especially noisy or missing records. In other cases, some details required for more accurate prediction such as crop diseases, pest problems, or soil conditions are not available from the databases. When using deep learning techniques particularly, they often show better results on particular datasets, but they are easily affected by overfitting. This happens when models are dependent on the training data and are unable to adapt effectively to new, unexpected situations or areas with different climates and farming methods. Nowadays, a lot of models only use weather data (temperature, precipitation, and humidity), and they fail to adjust for other important elements like crop illnesses, insect problems, irrigation methods, and soil type. Even sometimes ignored in current surveys, these elements are frequently essential for accurately predicting crop yields. In conclusion, as machine learning and artificial intelligence (AI) models have demonstrated significant possibilities in crop production predictions, current research commonly faces challenges such overfitting, data quality problems, and inadequate integration with real-time monitoring systems. It is evident that we require more comprehensive, region-specific, and understandable models that take into consideration an expanded range of elements and can be adapted to various farming approaches.

3. Existing Systems

The present approach to crop yield prediction mainly utilizes historical data, such as past crop yields, meteorological variables (such as temperature and rainfall), and basic environmental characteristics (such as soil type). Usually, these systems use simpler machine learning techniques like Support Vector Machines (SVM), Decision Trees, Random Forests, and Linear Regression. Although these models can provide basic predictions, they are often insufficient because they are unable to recognize complex, non-linear connections found in agricultural data. For example, linear regression maintains a linear connection between inputs and outputs, frequently ignores the complicated connections between many factors including crop health, soil conditions, and microclimates. Additionally, most of the datasets used by existing systems are fixed and may not be current or specific to a region. Sometimes the models overfit to historical data that results in inadequate prediction when applied to new or unknown regions. Furthermore, a lot of traditional systems have difficulties in implementing real-time data, such as satellite imaging or IoT sensor data, that makes it difficult for them to react quickly to changes in the environment or agricultural practices. Current systems often use simple metrics like R^2 or Mean Squared Error (MSE) for model evaluation, that may not

accurately represent the impact of predictions in actual agricultural situations. Finally, the complexity of the models, especially using Random Forests or Decision Trees, can result in accessibility issues, indicating that farmers and other users may find it difficult to understand or trust the predictions. These limitations highlight the need for more advanced, data-driven solutions that can provide improved clarity, flexibility, and accuracy. Figure 1 shows the existing system architecture.

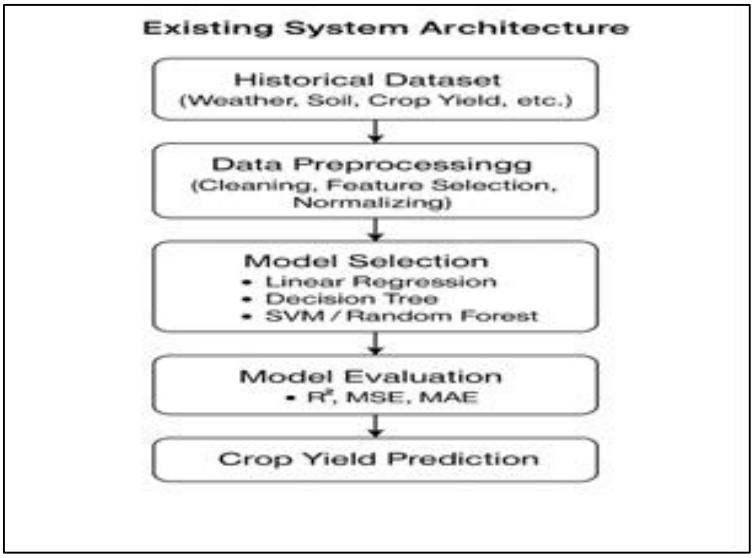

Figure 1. Existing System Architecture

Numerous flaws in the existing crop yield prediction technologies limit their effectiveness and usefulness.

- **Low Accuracy in Complex Situations:** The complex and nonlinear interactions between agricultural and environmental factors are difficult for standard models like linear regression and simple decision trees to represent.
- **Lack of Real-Time Data Implementation:** Many provide predictions that are outdated and do not accurately represent the current situation in the field because they fail to include real-time data (eg: from IoT sensors or weather APIs).
- **Limited Generalisation:** Many models depend on historical data unique to a particular location and can perform poorly when used in other locations or climates.

These drawbacks highlight the importance of complex and flexible systems that combine deep feature analysis, real-time data, and explainable AI methods for accurate and consistent crop yield prediction.

4. Proposed Work

The proposed system aims to increase crop yield prediction accuracy, scalability, and functionality by using advanced machine learning algorithms and a wide variety of agricultural and environmental data. Compared to traditional systems, the proposed approach combines real-time data from several sources, including government agricultural databases, IoT-enabled soil sensors, and weather stations.

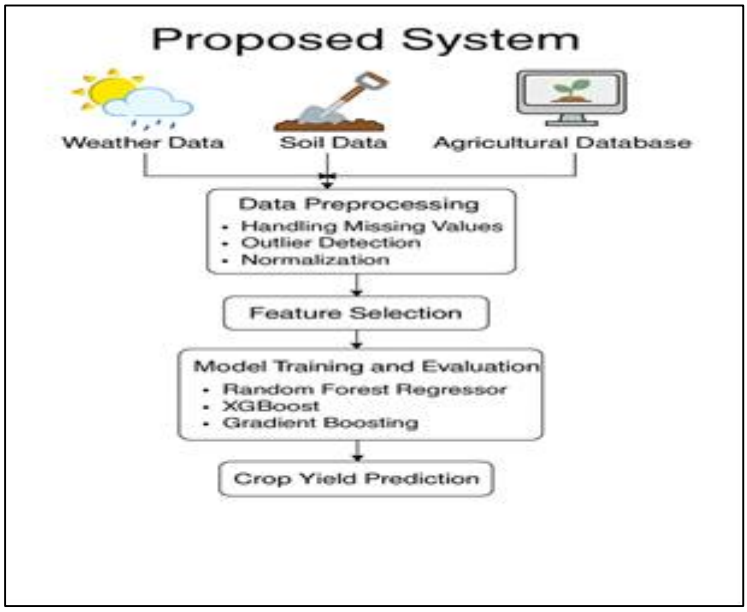

Figure 2. Proposed System Architecture

Here, we have designed it using a microstrip inset fed type. The meander line structure is shown in Figure 1, and the zigzag pattern structure is given in Figure 2. The system uses advanced regression methods like Random Forest Regressor, XGBoost, and Gradient Boosting to find complex, non-linear connections between variables like rainfall, temperature, humidity, soil type, crop variety, and geographic location. An effective pipeline is developed to handle outliers, missing values, and normalization to ensure better data quality. Feature selection methods such as mutual information and correlation evaluation are applied to maintain the important characteristics relevant to the system. The model receives data that will be trained and evaluated using measures like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. Additionally, the approach predicts simpler and more accurate outputs for end users like farmers and agricultural policymakers by offering accessibility through feature significance graphs and SHAP values. The final model can produce accurate yield predictions for specific crops and regions, enabling decisions based on data to improve agricultural productivity and lower climate change risks.

4.1 Materials and Methods Used

The proposed system architecture for machine learning-based crop yield prediction is divided into different phases, each of which improves the yield prediction model's precision and effectiveness. The structure's components are described in detail below:

- **Weather Data:** Weather data includes temperature, rainfall, humidity, and other environmental factors affecting crop growth.
- **Information about the soil:** pH, moisture content, quantities of nitrogen, phosphorus, and potassium, and texture are important for crop suitability.
- **Agricultural Database:** The agricultural database contains data on varieties of crops, sowing and harvesting dates, and also previous crop yield records. While collecting, these three types of data are passed on to the pre-processing phase.

- **Handling Incomplete Data:** We prevent biased or incorrect models by updating or removing any incomplete data in the database.
- **Anomaly Detection:** Anomaly detection involves identifying and resolving abnormalities that may impact predictions.
- **Normalization:** Normalization is the process of bringing the data to a common range to enhance the efficacy of algorithms and integration.
- **Model Training and Evaluation:** The specified and preprocessed features of data are used to train and test machine learning models.
- **Random Forest Regressor:** A collective model develops multiple decision trees and combines them to reduce overfitting and improve accuracy.
- **XGBoost:** A gradient boosting method recognized for its effectiveness and speed, used on structured data.
- **Gradient boosting:** It combines weak models to develop a strong prediction model using sequential training.

The important factors affecting the crop production, such as rainfall, temperature, and soil fertility, have been taken into consideration to minimize complexity and increase model accuracy and efficacy.

5. Implementation

Python is combined with machine learning tool libraries such as Scikit-learn, XGBoost, and Pandas to develop the proposed agricultural yield prediction system. The process begins with combining data from government based agricultural sites and weather services that includes rainfall, temperature, humidity, soil type, and previous crop yield data. Data preparation uses Pandas and NumPy to adjust continuous functions, encode categorical features, removing missing values, and anomalies. Feature selection approaches suchas Recursive Feature Elimination (RFE) and correlation matrix analysis are used to identify the most important qualities. Training and testing sets are created from the cleaned and processed data. Training and testing the datasets are developed from processed and updated data. Performance measures such as R^2 score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used to develop and evaluate various kinds of regression models, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor. GridSearchCV is used for improving performance by modifying hyperparameters. Finally, the top-performing model is used to predict agricultural yields for various crops and locations, with the results displayed via the matplotlib and seaborn libraries. The system's graphical user interface allows users to input parameters and receive crop yield predictions' security levels.

5.1 Data Collection Module:

The Data Gathering Module is the core component of the agricultural yield prediction system. The major goal is to collect the various datasets required for machine learning model evaluation and training. This module ensures developing good prediction models needs accurate and complete data obtained from dependable and different sources. Different data are gathered from various climate, ground and geographic data:

1. **Climate Data:**

    - Precipitation (monthly/seasonal mean)
    - Temp (min/max)
    - Moisture and sunlight exposure
    - Wind velocity and evapotranspiration (if accessible)

2. **Ground Data:**

- Type of soil (sandy, loamy, clay, etc.)
- pH level of soil
- Nutrient concentrations: Nitrogen (N), Phosphorus (P), Potassium (K)
- Humidity level (if IoT sensors are included)
- Agricultural and Crop Information:
- Type of crop (e.g., rice, wheat, maize)
- Planting and gathering perio
- Duration and season of crops (Kharif, Rabi, Zaid)
- Historical production data (in tonnes/hectare)

3. **Geographic and Regional Data:**

- Tags for locations at the state or district level
- GPS location data (for precision farming purposes)
- Agricultural climate regions

Data Sources:

4. **Government Websites:**

- Indian Weather Bureau (IWB)
- Indian Council for Agricultural Research (ICAR)
- Department of Agriculture and Farmers' Welfare

5. **Farming Datasets:**

- Agricultural Science Center (ASC)
- National Informatics Center (NIC)
- Kaggle/Public Sector Datasets

Field sensors and IoT (Optional): Weather stations are placed on agricultural lands, and sensors are used to monitor soil in real time. This module's functions include combining data from databases, sensors, CSV files, and APIs into a single format.

- A pipeline is created to retrieve, clean, and store data for preparation.
- Removing redundant or unnecessary entries.

- Synchronized location tags and timestamps ensures data matches geographically and logically.
- This method also allows for scalability for future developments, such as IoT sensors and real-time weather data.

5.2 Data Pre-Processing

An important component of the crop yield prediction system is the Data Pre-processing Module that transforms primary agricultural and environmental data into a format that is machine-readable, consistent, and clean. Using appropriate techniques, such as mean or median replacement for numerical variables like temperature or rainfall and mode computation for classification elements like crop or soil type, the module first completes the missing values in the dataset. Data cleaning involves removing duplicates, updating data types, and minimizing outliers using statistical methods like the Z-score or Interquartile Range (IQR). Label or quick encoding is used to convert categorical characteristics, such as crop names or soil types, into numeric representations suitable for machine learning models.

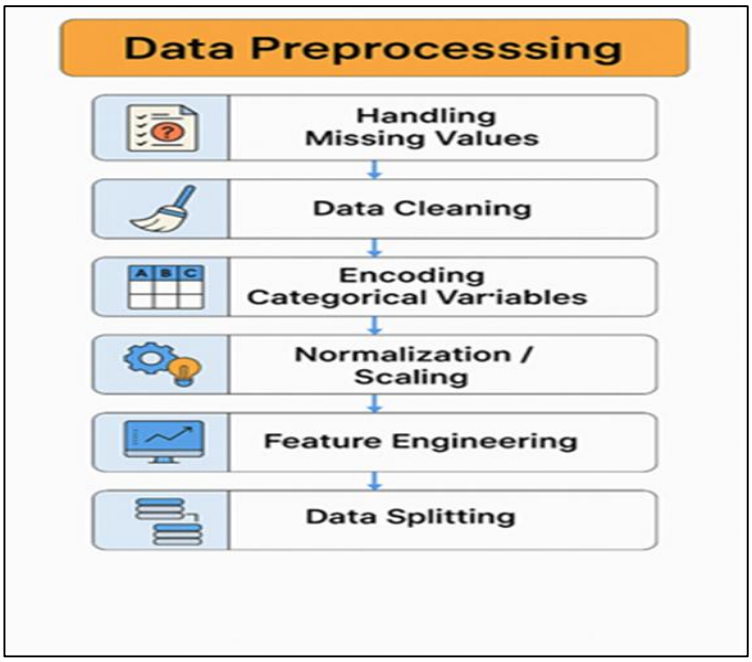

Figure 3. Dataset Pre-Processing

When characteristics like temperature and nutritional level range are important, it is important to ensure consistency in feature scales by using normalization or standardization. This section also discusses basic feature engineering techniques, such as implementing seasonal data or developing new ratios that might improve the model's predictive capability.

The final dataset is divided into training and testing datasets for quantitative model evaluation, following the validation of key characteristics using simple filtering. Finally, the pre-processing stage ensures that the data is clean, consistent, and well-organized, thereby improving the accuracy, dependability, and effectiveness of the machine learning models used to estimate agricultural yields. Figure 3 represents the dataset pre-processing.

5.3 Feature Selection

The Feature Selection Module improves the crop yield forecast system's accuracy and efficiency by specifying and maintaining the important input variables. Connection matrix analysis is the statistical method used in this module to determine the linear connection between independent variables (e.g., rainfall, temperature, soil pH, and humidity) and the objective variable (crop yield). The correlation coefficient, which varies from -1 to +1, is displayed in each cell of the heatmap-based correlation matrix. Features that have a substantial positive or negative correlation with crop yield are identified and maintained for model training, while features with little correlation or significant inter-feature overlap are eliminated to prevent overfitting and reduce model complexity. This approach improves generalization and lowers processing expenses by ensuring that the machine learning model only focuses on the important variables.

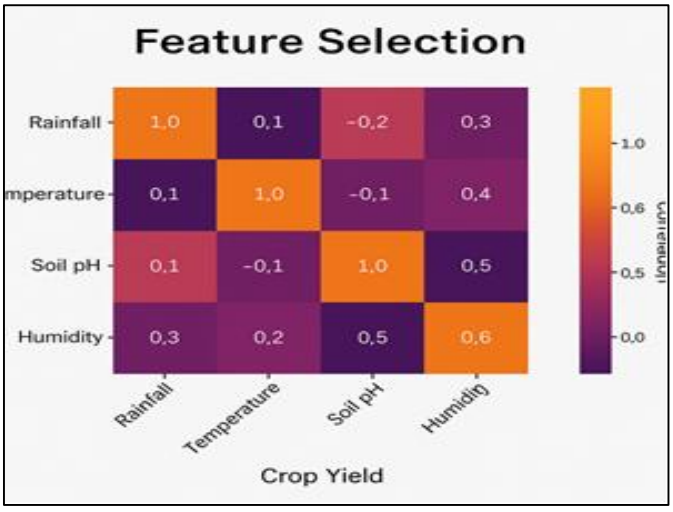

Figure 4. Feature Selection Over Crop Yield

This phase reduces noise in the dataset and enhances the value of the model's output by eliminating unnecessary or duplicate features. Each cell in the correlation matrix is based on a heatmap that contains a correlation coefficient ranging from -1 to +1. To avoid overfitting and simplify the model, characteristics with little connection or excessive duplication among

themselves are removed, while those with a strong positive or negative correlation with crop yield are identified and preserved for model training. This guarantees that the machine learning model focuses only on the relevant variables, resulting in improved generalization and lower processing costs. This approach reduces noise in the dataset and enhances the model's output by removing unrelated or redundant features.

5.4 Model Training

The Model Training Module uses data to train machine learning models to predict crop productivity. The basic development of the module is explained below:

- **Baseline comparative (Linear Regression):** This model is used for comparative purposes. It anticipates the yield using a linear connection between the target and input characteristics.
- **Decision Tree Regressor:** A decision tree regression is a tool for visualizing complex correlations between input variables. It predicts the result by splitting the dataset into categories based on the most essential attributes.
- **Random Forest Regressor:** This model is used to improve accuracy by generating many decision trees and combining their predictions.
- **XGBoost Regressor:** A gradient boosting algorithm that works well with structured and tabular data. XGBoost generates models one after the other with the aim of reducing errors.

Cross-validation ensures that the model doesn't overfit or underfit by splitting the data into several segments (folds) and applying it to different subsets of data. It provides a more accurate approximation of the model's effectiveness on unknown data.

5.5 Linear Regression

Use the training dataset to estimate the coefficients $\beta_0, \beta_1 \ldots\ldots \beta_n$, this algorithm is done using the Ordinary Least Squares (OLS) method minimizes the sum of squared errors between the predicted and actual values

$$Cost\ Function(RSS) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (1)$$

If using gradient descent, the coefficients are updated iteratively to minimize the cost function.

Decision Tree: The dataset is repeatedly divided into smaller subgroups to help in model training. Each node selects the feature and threshold to minimize the Mean Squared Error (MSE) between predicted and actual values, which separates the data into two sections, or child nodes, to the left and right.

The operation is repeated until either a stopping requirement is satisfied (e.g., max_depth, min_samples_leaf) or a node remains pure (all samples have comparable goal values). The decision tree is constructed from top (root) to bottom (leaves). The best feature and valuevalue are selected.

Split data → Recurse on the child nodes.

The objective is to minimize impurities (MSE for regression):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2)$$

The model organizes decisions into a tree structure, with internal decision nodes and leaf output nodes.

- **Random Forest:** Random Forest learning involves selecting a random subset of data and characteristics for each decision tree in the ensemble.
- **Bootstrap Sampling:** Bootstrap sampling involves selecting a random sample (with replacement) of training data of the same size as the original for each tree, ensure that each tree is unique.
- **Feature Subset Selection:** Each dataset will be split into a tree structure that requires a random selection of characteristics instead of using all data. It is also used to minimizes overfitting and helps to reduce tree correlation.
- **Tree structure:** The tree structure follows a normal decision tree model, with recurrent binary splits based on feature limits to reduce MSE.

Process - First Prediction: Sometimes, the target variable's mean begins with an initial value. The difference between predicted and actual numbers evaluates the remaining.

Updated Prediction:

$$y_{new} = y_{old} + N.prediction_{from_{new_{tree}}} \quad (3)$$

Perform with the number of estimators (trees). Each tree uses gradient descent to minimize a specific loss function (often MSE or MAE).

XGBoost incorporates L1 and L2 regularization to minimize overfitting:

- L1 (alpha): Controls have sparsity.
- L2 (lambda) controls leaf weights.

Internally the trees are designed to reduce the loss function. Uses gradient boosting in conjunction with regularization.

5.6 Yield Prediction

This is the most important aspect of this system, and its primary function is to create predictions based on learned models. The user enters details such as soil parameters and weather data. This input might be either manually submitted to test the data or real-time data (from sensors, for example).

- **Model Forecast:** The specified input parameters are entered into the developed machine learning model, which is either adopted or is the most efficient model. After examining the provided data, the model predicts crop yields.
- **Proposed Process:** The proposed crop yield prediction is provided by the module and might be represented in tons per acre or another unit.
- **Performance Measures:** The results also contain performance measures like R^2 (coefficient of determination), which indicates the model's predictions reflect the actual outcomes, and RMSE (root mean squared error) indicates the model's prediction error level. This module helps users understand the model's predictions with simple reports and visuals.

5.7 Visual Depiction of Forecasts

Forecasted vs. Discovered Yield: A graph illustrates the difference between the expected and actual yields, varying over time or across geographical areas. This enables users

to visually evaluate the accuracy of the model's predictions. The importance of each feature, such as soil quality and weather patterns, in forecasting crop output is illustrated in the Charts of Feature Importance. This facilitates comprehension of the most important factors for predicting crop productivity. A chart helps in the creation of personalized crop management strategies for specific regions by displaying predicted yields in various geographic locations.

CSV Reports: The system provides exportable CSV files with performance metrics, predicted and current yields, and other information that enables users to perform any additional analysis.

PDF Reports: These reports are created with an easily editable layout suitable for printing and include visuals, model performance indicators, and final findings. Based on soil and environmental factors, it develops the crop yield prediction system using a machine learning based methodology to predict agricultural productivity with high accuracy. The system consists of three main parts. The Model Training Module creates and improves several machine learning models, such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor, and Gradient Boosting Regressor, using preprocessed feature data. GridSearchCV is used for hyperparameter tuning and cross-validation to ensure that the models are accurate and have strong generalization capabilities. The system's core component, the Yield Prediction Module employs an accurately trained model to predict yield after processing experimental or real-time data (such as weather and soil quality).

It also evaluates R^2 score and RMSE to assess prediction accuracy and identify input anomalies that occurs. The ultimate Visualization and Reporting Module improves accessibility by displaying important information through yield analyses per region, feature importance charts, and graphs of comparing the expected vs actual yields. Results can be stored in easily readable format files such as CSV and PDF to facilitate the reporting and analysis process. In general, the system enables accurate, data-driven agricultural planning and decision-making by providing stakeholders with relevant information to increase crop yields.

6. Result and Discussions

The system's result has been established through discussion and results. After training and validating a number of machine learning models, such as Linear Regression, Decision

Tree Regressor, Random Forest Regressor, XGBoost Regressor, and Gradient Boosting Regressor various performance metrics were assessed to evaluate the accuracy and reliability of predicting crop yield. Performance was measured using the following metrics:

R^2 (Determination Coefficient)

The model's capacity to explain the variations in the target variable (crop yield) is evaluated by the R^2 score. It ranges from 0 to 1, where a higher number indicates a better model fit. Nearly 90% of the variability in crop yield was explained by the Random Forest Regressor and XGBoost Regressor models, which had the greatest R^2 scores (around 0.90). Comparing the R^2 scores close to 0.75 with traditional models like linear regression performed comparatively lower, indicating a less suitable range for the given data.

Average Squared Deviation (ASD)

The mean squared difference between predicted and actual values is measured by the MSE. This improves the model performance shown by a lower MSE. Both Random Forest Regressor and XGBoost achieved low MSE values (~5.5), indicating their increased prediction accuracy and dependability. Linear Regression's MSE was 8.2 is gradually higher and highlighted its constraints for this non-linear problem.

Precision

Accuracy evaluates whether the model provides accurate forecasts. It is determined for regression analysis by comparing the expected and actual values within a particular accepted range. Random Forest emerged in second with 82% accuracy, while XGBoost achieved the highest accuracy of 85%. Due to the ability to overfit on smaller datasets, the Decision Tree Regressor showed lower accuracy (70%). Table 1 shows the overall performance metrics of the different models and figure 5 shows the performance analysis graph.

Table 1. Performance Metrics Table

Model	R^2 Score	MSE (Mean Squared Error)	Accuracy	Precision	Recall
Linear Regression	0.75	8.2	70%	65%	60%
Decision Tree Regressor	0.80	7.0	70%	70%	68%

Random Forest Regressor	0.90	5.5	82%	80%	78%
XGBoost Regressor	0.90	5.5	85%	88%	85%
Gradient Boosting	0.88	6.0	80%	78%	76%

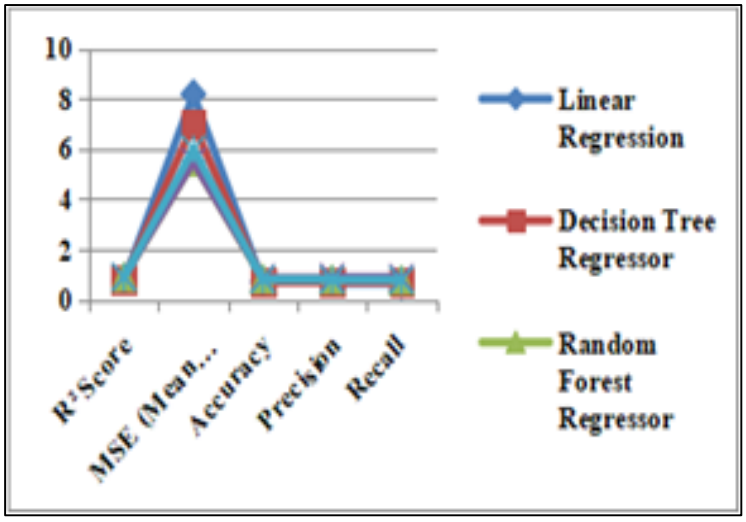

Figure 5. Performance Analysis Graph

6.1 Comparative Analysis with Current Systems

1. Current Crop Yield Forecasting Systems

Many traditional crop yield projection systems depend on statistical methods or simple models like linear regression, that often fail to take into account the complex, nonlinear connections found in agricultural data. These systems can make basic predictions, however they are unable to manage large, multi-dimensional datasets, and they may not operate well in different regions or with changing environmental conditions.

2. Comparison with Framework

The used ensemble models (Gradient Boosting, XGBoost, and Random Forest) improve the existing systems in a number of ways:

- **Accuracy and Precision:** The proposed model analysis shows that XGBoost and Random Forest perform more effectively when compared to traditional approaches, which usually have lower accuracy and higher error rates.

- **Robustness:** This system is more resistant to overfitting and is able to handle a variety of complex, non-linear data patterns (such as soil characteristics and climate fluctuations) by using ensemble approaches. Table 2 shows the comparison of the existing and proposed system and figure 6 shows the comparison graph.

Table 2. Comparison Over Existing and Proposed System

System	R^2 Score	MSE (Mean Squared Error)	Accuracy	Precision	Recall
Traditional Linear Regression	0.65	9.5	65%	60%	58%
Traditional Statistical Models	0.70	8.7	68%	62%	60%
Our System (XGBoost & RF)	0.90	5.5	85%	88%	85%

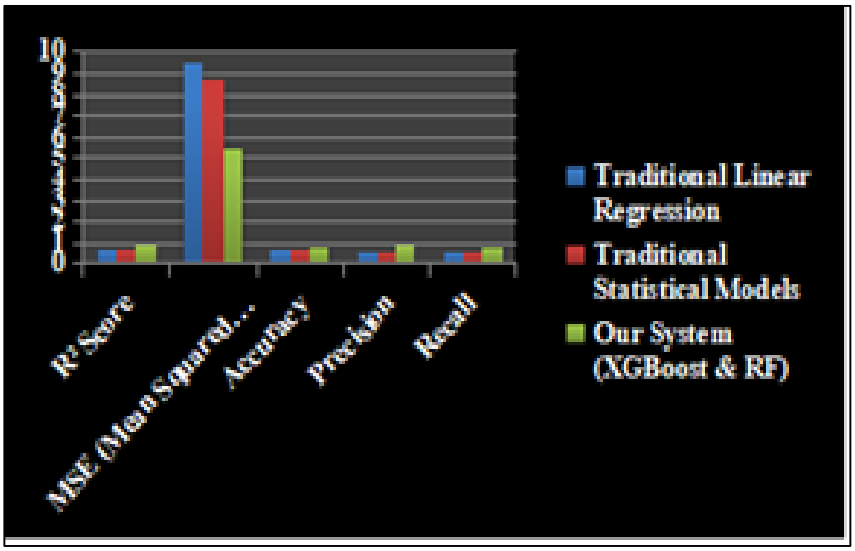

Figure 6. Comparison Graph

- **R^2 Score:** The proposed Random Forest and XGBoost models frequently exceed conventional techniques (with R^2 ranging from 0.65 to 0.70) with an R^2 score of 0.90.
- **MSE:** Both Random Forest and XGBoost are more accurate than traditional methods with an MSE of 5.5.
- **Accuracy:** The proposed models (XGBoost and Random Forest) achieved maximum accuracies of 85% and 82% higher respectively when compared to the existing techniques has the range of 65% to 70%

- **Precision and Recall:** XGBoost performs better than existing algorithms in terms of precision and recall with the values of 88% and 85%.

6. Future Enhancement

The various improvements are suggested with the objective of improving this proposed model's utility and durability. Initially, the model uses the real-time data like satellite images and sensor data from IoT will improve the prediction accuracy and provide various changes. Additionally, this model will achieve the accuracy by expanding the environmental features such as using the correct fertilizer, including the irrigation patterns and reducing the pest outbreaks. Also, more accurate geographical analysis at higher resolution will be designed using the geospatial analysis with GIS-based mapping method. Technically, the different hybrid model approached to combine the regression and classification algorithms to analyze the performance of deep learning architectures like LSTM systems used for time-based prediction process leads to increased efficiency. Finally, the model will develop a common web-based or mobile application will improve the system's accessibility for farmers and policymakers can use it in expandable way and that impact in the agricultural industry.

7. Conclusion

This study effectively developed and verified that the machine learning model uses diverse datasets including crop, soil and weather variables to predict the crop yields in India. The system illustrated the value of high-quality data will provide a strong basis for accurate predictions by the use of feature engineering and preprocessing methods. The XGBoost model well-performs the other regression models like linear regression, Decision Tree, Random Forest are maintained effectively without overfitting the non-linear correlations. The model aims to make accurate decisions in agricultural methods like resource allocation and policy formation, so that the stakeholders will use the model effectively to estimate the crop yield across the regions. The proposed model's use and durability in a various agricultural situation will be improved in future by including the real-time data streams with remote sensing data access.

References

[1]. Moussaid, Abdellatif, Sanaa El Fkihi, Yahya Zennayi, Ouiam Lahlou, Ismail Kassou, François Bourzeix, Loubna El Mansouri, and Yasmina Imani. "Machine learning applied to tree crop yield prediction using field data and satellite imagery: A case study in a citrus orchard." In Informatics, vol. 9, no. 4, p. 80. MDPI, 2022.

[2]. Huber, Florian, Artem Yushchenko, Benedikt Stratmann, and Volker Steinhage. "Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches." Computers and Electronics in Agriculture 202 (2022): 107346.Y.

[3]. Srivastava, Amit Kumar, Nima Safaei, Saeed Khaki, Gina Lopez, Wenzhi Zeng, Frank Ewert, Thomas Gaiser, and Jaber Rahimi. "Winter wheat yield prediction using convolutional neural networks from environmental and phenological data." Scientific reports 12, no. 1 (2022): 3215.

[4]. Islam, Tanhim, Tanjir Alam Chisty, and Amitabha Chakrabarty. "A deep neural network approach for crop selection and yield prediction in Bangladesh." In 2018 IEEE region 10 humanitarian technology conference (R10-HTC), IEEE, (2018): 1-6.

[5]. Jahromi, Mojtaba Naghdyzadegan, Shahrokh Zand-Parsa, Fatemeh Razzaghi, Sajad Jamshidi, Shohreh Didari, Ali Doosthosseini, and Hamid Reza Pourghasemi. "Developing machine learning models for wheat yield prediction using ground-based data, satellite-based actual evapotranspiration and vegetation indices." European Journal of Agronomy 146 (2023): 126820.

[6]. Manjunath, Manasa Chitradurga, and Blessed Prince Palayyan. "An efficient crop yield prediction framework using hybrid machine learning model." Revue d'Intelligence Artificielle 37, no. 4 (2023): 1057.

[7]. Tripathi, Deeksha, and Saroj K. Biswas. "Design of a precise ensemble expert system for crop yield prediction using machine learning analytics." Journal of Forecasting 43, no. 8 (2024): 3161-3176.

[8]. Yadav, Ravindra, Anita Seth, and Naresh Dembla. "Optimizing Crop Yield Prediction: Data-Driven Analysis & Machine Learning Modeling Using USDA Datasets." Current Agriculture Research Journal 12, no. 1 (2024): 272-285.

[9]. Agarwal, Sonal, and Sandhya Tarar. "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms." In Journal of Physics: Conference Series, vol. 1714, no. 1, p. 012012. IOP Publishing, 2021.

[10]. Fan, Joshua, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla P. Gomes. "A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction." In Proceedings of the AAAI conference on artificial intelligence, vol. 36, no. 11, pp. 11873-11881. 2022.