SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I1P114

Volume 13 Issue 1, 157-173, January 2026

Original Article

Ensemble-Driven Machine Learning Regression Models for Climate-Sensitive Crop Yield Prediction: A Comparative Performance Analysis

Siva Subramanian R[1*], M Elumalai[2], B.Saratha[3] , K.Ramesh[4], K.Sudha[5], J.Gnana Jeslin[6]

[1]*Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Tiruchirappalli Campus, Tamil Nadu, India.*
[2]*Department of CSE, Asan Memorial College of Engineering and Technology, India.*
[3]*Department of Artificial Intelligence and Data Science, R.M.K.Engineering College, Kavarpettai, India.*
[4]*Department of Information Technology, Panimalar Engineering College, Chennai, India.*
[5]*Department of CSBS, Associate Professor, R.M.D Engineering College Kavarpettai, India.*
[6]*Department of CSE, R.M.K College of Engineering and Technology, Puduvoyal, India.*

[1]*Corresponding Author : sivamr8@gmail.com*

Received: 08 November 2025 Revised: 10 December 2025 Accepted: 09 January 2026 Published: 14 January 2026

Abstract *- Precise forecasting of crop yields is the key to food security, resource management, and sustainable food farming. This paper will examine how different Machine Learning (ML) models can be used to predict crop yield in relation to climatic and other environmental conditions, like rainfall, temperature, and the use of pesticides. Multiple performance metrics, such as R^2, Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) were used to train and evaluate seven ML models which were Linear Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), Gradient Boosting (GB), XGBoost, Random Forest (RF), and Bagging. The experimental findings showed that the ensemble-based models were very effective compared to the traditional regression and distance-based algorithms. The Bagging recorded the best prediction accuracy in terms of R^2 score, closely followed by the RF. The two models were effective in capturing nonlinear relationships and high generalization in varied climatic and crop conditions. On the other hand, the simplicity of models like LR and KNN demonstrated low predictive abilities. The results highlight the scalability and the strength of the Ensemble Learning(EL) techniques in crop yield forecasting. The paper concludes with a set of recommendations on how to incorporate Explainable AI, real-time data that uses IoT, and region-specific hybrid deep learning systems to improve the interpretability, adjustment, and accuracy of agricultural forecasting systems in the future.*

Keywords *- Bagging, Crop Yield Prediction (CYP), Ensemble Learning, Machine Learning, Random Forest.*

1. Introduction

1.1. Background

One of the most important sectors that contributes to the survival of human beings and the economy is agriculture, which supplies food, raw materials, and jobs to billions of people across the world. Proper prediction of crop yield is an issue of focus in agricultural research and management since it allows agricultural policy makers, farmers, and agribusiness to make good decisions about food security, food prices, and allocation of resources. Nevertheless, crop yield is a complicated process that is affected by a great variety of environmental, agronomic, and socio-economic factors. The interaction of variables like rain, temperature, use of pesticides, soil fertility, and climatic patterns in the region is not linear, and in fact, interacts in complex nonlinear manners to influence crop productivity [1]. The world agricultural industry has been experiencing increasing challenges in recent years owing to climate change, erratic weather patterns, and soil erosion.

The changes in rainfall and the increase in temperature have had a significant impact on crop production and food availability, especially in areas that rely on rain-fed crops. These issues reveal the pressing necessity to have more data-driven, adaptive, and reliable prediction systems that can help reduce risk and guarantee sustainable agricultural production [2]. Historical trends, statistical methods that are based on regression, as well as domain knowledge, are key elements in the traditional methods of predicting crop yield. In spite of a certain degree of accuracy, such models do not always represent the sophisticated relationships between climatic, biological, and geographical variables. Moreover,

these traditional methods have limitations of linear assumptions, lack scalability, and are sensitive to missing or noisy data. Machine learning ML) has become an effective tool for improving the accuracy of crop yield prediction in this respect [3]. ML methods can be trained to capture the latent patterns and nonlinear relationships in large datasets and can provide superior generalization and flexibility to changing agricultural conditions. Using ML algorithms, it is possible to process agricultural data, including temperature measurements, precipitation, pesticide application, and soil properties in large volumes to produce high-quality predictive models that are both interpretable and accurate [4].

As a result, prediction models based on ML have become a part of the development of precision agriculture systems that should maximize the use of the resources, enhance their efficiency, and minimize their environmental impact [5, 6]. ML-based crop yield prediction has, therefore, tremendous opportunities to transform the existing agricultural processes and allow making decisions based on data, and play a significant role in the overall objectives of global food security [7]. This paper discusses and analyzes the performance of various ML algorithms to determine the most dependable method of predicting crop production based on climatic and environmental data.

1.2. Problem Definition

The forecasting of agricultural production is, by its nature, a challenging task because of the dynamic and multivariate character of influencing factors. Conventional yield estimation techniques, including regression-based prediction, time series forecasting, or expert-based evaluation, tend to be restricted by their assumption of fixed assumptions and simplistic relations.

These models do not usually account for the nonlinear interaction between the major environmental factors, such as temperature, rainfall, and pesticide use, which, in combination, define the outcome of crop yield [7, 8]. Moreover, the fact that the agricultural conditions of various regions, soils, and crops vary complicates the accurate modeling of yield. As an illustration, two areas that receive the same amount of rainfall can yield vastly different amounts of yield because of differences in soil fertility or pesticides used.

Moreover, climatic fluctuations have increased the uncertainty of the agricultural results, and thus, the use of conventional linear models is not practical in contemporary times. Another weakness of conventional methods is that they cannot effectively analyze large and heterogeneous data sets. Agricultural data typically includes thousands of observations gathered across multiple regions, crops, and years. Such high-dimensional data cannot be handled with algorithms that can only learn more complicated relationships, but also be generalized to previously unknown conditions. Thus, it is increasingly necessary to have automated and scalable solutions capable of forecasting crop yields with a high degree of accuracy, taking into account a combination of several interacting variables simultaneously.

The alternative to traditional prediction methods is the use of machine learning ML) models, including ensemble techniques, regression trees, and boosting algorithms. They can understand the past trends in agriculture, adjust to the shifting climatic trends, and produce exact forecasts of Yield [9].

The primary issue this study aims to address is the enhancement of the accuracy, strength, and explainability of crop yield forecasting based on various machine learning ML) methods trained on actual agricultural data [10]. The study will identify the best-performing algorithm by systematically comparing various models under different geographical and climatic conditions to understand which algorithm is most effective at modeling crop production in nonlinear conditions.

1.3. Research Motivation and Objectives

Efficient CYP is crucial in addressing some of the most pressing issues in global agriculture, including food insecurity, resource scarcity, and climate-related production losses. The rationale of the study is the growing demand for data-driven intelligence in agricultural decision-making processes.

As the Artificial Intelligence industry rapidly evolves and open agricultural data sets are provided, it has become a possibility to use advanced ML methods to model yields with an even greater degree of accuracy than before.

1.3.1. Three Main Reasons Drive this Research

First, it will offer a comparative analysis of several ML models, i.e., LR, RF, GB, XGBoost, KNN, DT, and Bagging, on a large, real-world dataset, obtained on Kaggle. The comparative approach enables the identification of models that offer the best trade-offs in terms of accuracy, complexity, and interpretability.

Second, the research aims to understand how major environmental elements, such as rainfall, average temperature, and pesticide use, impact agricultural productivity. Through the analysis of the feature importance of various ML models, the study identifies which variables are the most important in explaining the variation in the yield across the countries and crop types.

Third, the study has been inspired by the desire to promote sustainable agriculture and policy formulation. With improved and more precise predictive systems, governments and farmers can plan crop production more effectively, allocate resources more efficiently, and predict potential yield deficits resulting from climate change.

The following are the objectives of this study,

1. To pre-process and analyze a multi-country crop yield dataset of climatic and agricultural characteristics.
2. The work aims to deploy and train several ML models to predict crop yield.
3. To measure and compare the model performance based on the statistical measures like R2 score, MSE, MAE, and MAPE.
4. To determine the most performing model according to predictive accuracy and generalization.
5. To examine how the environmental variables impact yield results.
6. The study achieves these objectives in relation to a broader objective of incorporating AI in agricultural forecasting and decision-making.

1.4. Scope and Contribution

This paper aims to apply, compare, and evaluate the use of various ML algorithms in predicting crop yield based on a real-world dataset of various regions and crops. The data consists of the key climatic and environmental data, as well as specific and regional data.

Specific information about soil or irrigation is not considered in the research because these variables are not available in the data set; instead, it focuses on the effects of climatic and environmental variables on yield outcomes. The researchers employ seven ML-based models to compare their performance and find the most effective model to use in predicting yield accurately.

The main contributions of this paper can be presented as follows,

1. Creation of an efficient ML pipeline to predict crop yields, such as the processing of data, encoding of features, and evaluation of the model.
2. Extensive comparative study of seven ML models (LR, RF, Gradient Boost, XGBoost, KNN, DT, and Bagging).
3. Detection of the Bagging and the RF as the most successful models with the highest R^2 scores and the lowest error scores.
4. An in-depth investigation into feature significance reveals that rainfall, temperature, and pesticide use are the most significant factors in predicting yield.
5. Further, the cause of sustainable agriculture is achieved by showing how ML-driven systems can be used to improve the accuracy of forecasting, facilitating informed decision-making, and resource optimization.
6. The given paper thereby contributes to the expanding literature on the topic of AI-based agricultural analytics by offering an evidence-based assessment of the model performance on a large and heterogeneous dataset.

1.5. Paper Organization

The remainder of the paper is structured in the following way:

Section 2 (Literature Review) gives a summary of previous studies that have been conducted on predicting crop yield using traditional and ML models, with their key variables and gaps in the research.

Section 3 (Dataset Description) describes the dataset that was utilized in this study, including its features, statistical features, and pre-processing stages.

Section 4 (Methodology) includes the ML models used, model configurations, training plan, and performance assessment measures.

Section 5 (Results and Analysis) expounds on the comparative findings of all models, article feature significance, and best-performers.

Section 6 (Discussion) explains the findings in connection to the practical application in agriculture, challenges, and implications on sustainability.

Section 7 (Conclusion and Future Work) recaps the overall findings, contributions, and future research directions in ML-based agricultural forecasting.

2. Literature Survey

ML-based crop yield prediction has become one of the significant areas of research in precision agriculture. Different researchers have investigated the use of traditional, hybrid, and deep learning methods to predict yield by using climatic, soil, and environmental factors. This section is a summary and analysis of the current literature on CYP, with the emphasis on datasets employed, dependencies between features, models, and important conclusions.

2.1. Machine Learning for Crop Yield Prediction

ML is a critical component of agricultural decision support systems that determines the pattern of yield and crop management optimization.

[11] ML is the key decision support tool in predicting crop yields and informing agricultural practices. A Systematic Literature Review located and synthesized 50 articles out of an original sample of 567 relevant articles in six electronic databases. Analysis showed that the most commonly used features were temperature, rainfall, and soil type, with Artificial Neural Networks being the most used algorithm. Furthermore, another search found 30 studies that were dedicated to deep learning, and CNN were the most popular Deep Learning algorithms, as well as LSTM and DNN. [12] ML can be helpful in forecasting harvest production and informing agricultural choices. The ML techniques are beneficial as the farming system is a complicated system that entails different data points. This paper discusses various methods of soil and environmental-based methods of predicting yields. The aim is to create an

ML model that would help farmers to select crops and improve the yield, which would result in fewer losses and better prices. These models can be either descriptive or predictive, depending on the research objectives. [6] Agriculture plays a vital role in the Indian economy, with a significant percentage of the population, over 50 percent, depending on it. Climate change is a threat to the health of agriculture. ML is a decision support tool that supports CYP, which helps in the management and choice of crops. This study provides a systematic review of features applied in CYP and identifies various AI techniques for studying CYP. Neural Networks have their limitations, such as amplified errors in predictions, and supervised learning fails to work with the nonlinear relationship of the data. It aims at creating proper models to classify crops and estimate their yield, taking into account such factors as weather and crop diseases, to achieve higher accuracy in the estimation of crop yield using different methods of ML.

2.2. Deep Learning Approaches for Crop Yield Prediction

Recent studies have increasingly focused on DL to predict crop yields due to its ability to automatically extract spatial and temporal features from extensive and non-homogeneous data.

[13] DL is also being considered an important technique in crop yield forecasting, and it succeeds in learning with data sets and thus automatically discovers the important features. This literature review identifies gaps in the study of DL methods and assesses the impact of vegetation indices and environmental conditions on crop yields. A review of recent literature (2012-2022) indicates that the most commonly used DL methods are LSTM and CNN with satellite remote sensing technology. The most prevalent features in predictions are vegetation indices, but the effectiveness of vegetation indices differs among the methodologies. Some of the significant issues are to increase the accuracy of the model, its practical use by stakeholders, and the black box character of these models. [14] Remote sensing and the use of UAVs in smart farming are becoming popular in the detection of crops and weeds, biomass analysis, and the prediction of yields. This paper uses CNN to forecast crop yields using NDVI and RGB data. The methodological tests involving CNN parameters such as training algorithm, network depth, and hyperparameter optimization provided an average absolute error (MAE) of 484.3 kg/ha and an average absolute percentage error (MAPE) of 8.8 percent in early growth stages (June 2017). To allow subsequent growth (July and August 2017), the MAE was 624.3 kg/ha (MAPE: 12.6%). Interestingly, the CNN model worked well when using RGB data as opposed to using NDVI data. [15] The article presents a DL model that uses the CNN and the RNN to predict the yields of corn and soybean crops in the U.S. Corn Belt between 2016 and 2018. CNN-RNN model outperformed the other methods, such as RF and deep fully-connected neural networks, by a significant margin with an RMSE of 9 percent of average yields and an 8 percent of average yields. The main characteristics of the model are the possibility to model time dependencies related to the environment, predictive generalization to new environments, and the measurement of the influence of weather, soil conditions, and management practices on the changes in yield.

2.3. Hybrid and Comparative Models

Combination models that incorporate the advantages of both linear and nonlinear algorithms have also been considered.

[16] The genotype, environment, and interaction between genotype and environment affect crop yield. These relationships are essential for making accurate predictions, which require large datasets and sophisticated algorithms. Yield prediction in 2017 was performed using datasets of 2,267 maize hybrids across 2,247 locations in the 2018 Syngenta Crop Challenge. The DNN model of our winning team had a Root-Mean-Square-Error (RMSE) of 12% of the average yield using predicted weather data, and the perfect data had a Root-Mean-Square-Error (RMSE) of 11%. The input dimensions were reduced through feature selection without compromising accuracy. The model performed better than other techniques, such as Lasso, shallow neural networks, and regression trees. It was found that environmental factors are more influential on crop yield than the genotype. [17] Crop yield prediction is a highly complicated process that has been widely studied with the help of ML, especially ANN and Multiple LR. This paper discusses the connection between MLR and ANN and suggests a hybrid model of MLR-ANN, which may be used to provide better predictive accuracy. The model uses the coefficients of MLR to set the weights and biases of the input layer of the ANN instead of using random values. This hybrid model is compared to the conventional models, such as ANN, MLR, Support Vector Regression (SVR), KNN, and RF. It has been found that the MLR-ANN hybrid model is more accurate and takes into account computational time compared to traditional methods.

2.4. Application-Specific Studies

[2] Forecasting of crop production is essential in financial analyses of the agricultural sector, affecting import-export policies and the income of farmers. This paper provides a review of machine learning ML) algorithms in crop yield prediction, with a focus on palm oil. It discusses the current state of palm oil production, its popular characteristics, and forecasting algorithms. A critical analysis of the current machine learning ML) application in the palm oil industry, along with comparative research, is presented. The article highlights the benefits and challenges of machine learning ML) in predicting yields and proposes potential solutions for the future. It discusses remote sensing, plant growth, and disease detection, and suggests a future

architecture of palm oil yield prediction to improve the accuracy and minimize computation challenges. [18] Agriculture plays a significant role in India's economy and is the foundation of the country's civilization. Being an agrarian country, crop choice is crucial for economic development, as it depends on market prices, production rates, and government policies. To increase agricultural productivity, the application of ML methods can streamline crop choices, solve farmers' problems, and improve yield level, ultimately yielding positive results for the Indian economy.

2.5. Comparative Analysis and Research Gaps

In the literature that has been reviewed, there are several patterns and gaps in research:

Dominating Features and Data Sources: Temperature, rainfall, and soil features are the most dominant features. Spatial yield estimation is also commonly performed using satellite-based vegetation indices.

Algorithmic Trends: The classical models of ML, such as the RF and SVM, are not out of competition, yet DL architectures, particularly CNN, LSTM, and the hybrid CNN-RNN, demonstrate greater ability to process large volumes of unstructured agricultural data.

Hybrid Approaches: Predictive performance is better, and the training time of models is lower when linear and nonlinear methods are combined (e.g., MLR-ANN, CNN-RNN).

Difficulties: The typical limitations are an imbalance of data, absence of soil and management data, insufficient temporal coverage, and inability to interpret deep models.

Future Directions: It is evident that there has been a shift towards Explainable AI (XAI) to enhance the model transparency, IoT-based real-time predictive systems, and region-specific adaptive systems to meet local agricultural requirements.

Overall, the literature suggests that ML, primarily ensemble and DL techniques, can be used to provide practical solutions to the accurate forecasting of crop yields. The recent researches are dominated by the ANN, CNNs, LSTMs, and ensemble tree models (RF, Bagging) because of their ability to address nonlinear relationships and large datasets.

Nevertheless, there are still significant limitations, such as data availability, interpretability, and scalability, that future research should address with the help of hybrid modeling, IoT integration, and explainable frameworks. The knowledge of previous research works gives a powerful background to the present study, which further elaborates this area through conducting a comparative study of several ML models utilizing a worldwide agricultural dataset, assessing their precision, reliability, and calculating efficiency in crop output prediction.

Table 1. Overview of the dataset

	Unnamed: 0	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
0	0	Albania	Maize	1990	36613	1485.0	121.00	16.37
1	1	Albania	Potatoes	1990	66667	1485.0	121.00	16.37
2	2	Albania	Rice, paddy	1990	23333	1485.0	121.00	16.37
3	3	Albania	Sorghum	1990	12500	1485.0	121.00	16.37
4	4	Albania	Soybeans	1990	7000	1485.0	121.00	16.37
...	...	...	...	...	...	...	...	...
28237	28237	Zimbabwe	Rice, paddy	2013	22581	657.0	2550.07	19.76
28238	28238	Zimbabwe	Sorghum	2013	3066	657.0	2550.07	19.76
28239	28239	Zimbabwe	Soybeans	2013	13142	657.0	2550.07	19.76
28240	28240	Zimbabwe	Sweet potatoes	2013	22222	657.0	2550.07	19.76
28241	28241	Zimbabwe	Wheat	2013	22888	657.0	2550.07	19.76

3. Dataset Description

3.1. Dataset Overview

The data used in this study were obtained from the CYP Dataset on Kaggle [19]. It is an extensive set of agricultural, climatic, and environmental data on several countries and types of crops, making it suitable for creating and testing ML models to predict yields. The dataset comprises 8 key features and 28,242 records, encompassing both categorical and numerical variables. The records represent the agricultural output of a specific type of crop produced in a given country during a particular year, taking into account climatic and input-related factors such as rainfall, pesticide

use, and temperature. This dataset was selected due to its geographical diversity, broadness of climatic conditions, and various aspects of yield that influence it, which is critical in training ML models that can represent the complex and nonlinear interactions. It also shows the diversity of environmental conditions and agricultural methods in different countries, which is important in determining the generalization of models. The data is stored in CSV format, and preprocessed and explored data analysis was done with Python-based libraries like Pandas and NumPy. There were no missing or null values, which guarantees the consistency and reliability of data in training ML models.

3.2. Data Attributes and their Significance

The data set has eight columns, which are important determinants of crop yield. The variables and their importance are explained as follows:

1. Unnamed: 0 (Serial Number): The index of each record in the dataset. Though not a predictive feature, it helps in the identification of records and the management of data.
2. Area (Country): A categorical variable that states the country/region of cultivation of the crop. It is important because geography greatly influences agricultural output, as it determines the type of soil, climate, and farming methods. The sample comprises 101 different countries, representing a wide geographical range.
3. Item (Crop Type): Categorical variable that shows the type of crop grown (e.g., Maize, Wheat, Rice, Potatoes, Soybeans, Sorghum, etc.). Crop type is important because all crops have different biological characteristics, water requirements, fertilizer needs, and sensitivity to climate conditions. The dataset has 10 different types of crops.
4. Year: The Year of data collection, which is between 1990 and 2013. This time-varying aspect allows models to include the trends and changes in yield that vary with time due to technological advancement, changes in policies, or changes in climate.
5. hg/hayield (Target Variable): Represents the product of crops in hectograms/hectare (hg/ha). It is the target variable (dependent variable) in the research, and the leading indicator of agricultural productivity.
6. Average rainfall mm per Year: The mean amount of rainfall (in millimeters) in a country and Year. Rainfall is among the most important variables in crop yield, as it influences soil moisture, the need for irrigation, and plant growth in general.
7. Pesticides tonnes: The total pesticides used (in tonnes). The feature helps capture the effect of pest control practices on crop health and productivity. Nevertheless, its impact is nonlinear because it has environmental side effects when used excessively.
8. Avgtemp (Average Temperature): Means the average temperature per Year (in °C). Temperature affects the stages of crop development, including germination, flowering, and yield development. Both low and high extremes can negatively impact yield, making it a critical variable in prediction.

Data Types: The majority of the columns (6 out of 8) are of the int64 (or float64) data type. The only object data type columns are the Item and Area ones. Missing Values: There are no missing values within the dataset because every column contains 28242 non-null values.

3.3. Statistical Overview of the Dataset

In order to have a more comprehensive view of the structure and variation of the dataset, descriptive statistical analysis was performed on the numerical attributes. Table 2 shows the summary statistics of the key numerical variables.

Table 2. Descriptive summary of numerical attributes

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	28242.0	14120.500000	8152.907488	0.00	7060.2500	14120.50	21180.75	28241.00
Year	28242.0	2001.544296	7.051905	1990.00	1995.0000	2001.00	2008.00	2013.00
hg/ha_yield	28242.0	77053.332094	84956.612897	50.00	19919.2500	38295.00	104676.75	501412.00
average_rain_fall _mm_per_year	28242.0	1149.055980	709.812150	51.00	593.0000	1083.00	1668.00	3240.00
pesticides_tonnes	28242.0	37076.909344	59958.784665	0.04	1702.0000	17529.44	48687.88	367778.00
avg_temp	28242.0	20.542627	6.312051	1.30	16.7025	21.51	26.00	30.65

Table 2 is the summary of the descriptive statistics of the numerical variables in the crop yield prediction dataset of 28,242 observations. The data covers a variety of geographical areas and time periods. The statistical summary illustrates the variation and dispersion of the major agricultural indicators, including crop yield (kg/ha), annual rainfall (mm), pesticide application (tonnes), and average temperature (°C).

The average crop yield is nearly 77,053 hg/ha, and the standard deviation is very high, which means that there is a big difference in crop yield among different crops as well as among different locations. The annual rainfall ranges from 51 mm to 3240 mm, reflecting the diverse climatic conditions. Similarly, the use of pesticides is not evenly distributed, indicating that agricultural practices vary.

This dataset is particularly effective in training ML models that can generalize to other agricultural settings due to its variety in both geographic and climatic dimensions.

3.4. Key Insights and Observations

The initial data analysis indicates some crucial information regarding the model development and interpretation:

Variation in Climatic Conditions: The data set represents a broad range of environmental situations, from arid areas with low rainfall to tropical areas with substantial rainfall. This will enable ML models to be trained on heterogeneous data, improving their capability to be applied to various regions.

Great Diversity of Crop Yield: The standard deviation of the yield variable (~84,956) is very high, indicating that other factors, such as rainfall, temperature, and the type of pesticide used, vary differently according to the crop type and region.

Temporal Range and Technological Influence: The inclusion of data between 1990 and 2013 enables the analysis one trends over time, such as the advancement of agricultural technology, fertilizer use, and irrigation systems that could impact yield.

Good Representations of Particular Crops and Areas: Different nations have a very good number of records, which give a consistent data sample on the major types of crops. In the same way, a large percentage of crops such as Potatoes, Maize, and Wheat gives more reliability to the models of the category.

Balanced Data Quality: There is no imbalance in the data, as all attributes have 28,242 non-null entries. This feature enables the models to train uniformly without the need for data imputation.

Possible Correlation between Features: According to the preliminary correlation analysis, it is observed that rainfall, temperature, and pesticide use are likely to be correlated with yield to some degree. The feature importance analysis can further be used to explore these relationships.

In general, the dataset is highly suitable for predictive modeling, featuring a wide range of high-quality data, as well as abundant climatic, geographical, and agricultural data. It lays a concrete foundation for training, testing, and validating ML models to predict crop yields, both individually and at scale, accurately.

4. Methodology

The following section describes the general structure, pre-processing, ML models, configuration, training, validation plan, and evaluation metrics of this study. The methodology framework was developed to provide the right, stable, and repeatable crop yield forecasting based on ML algorithms, based on multiple regression.

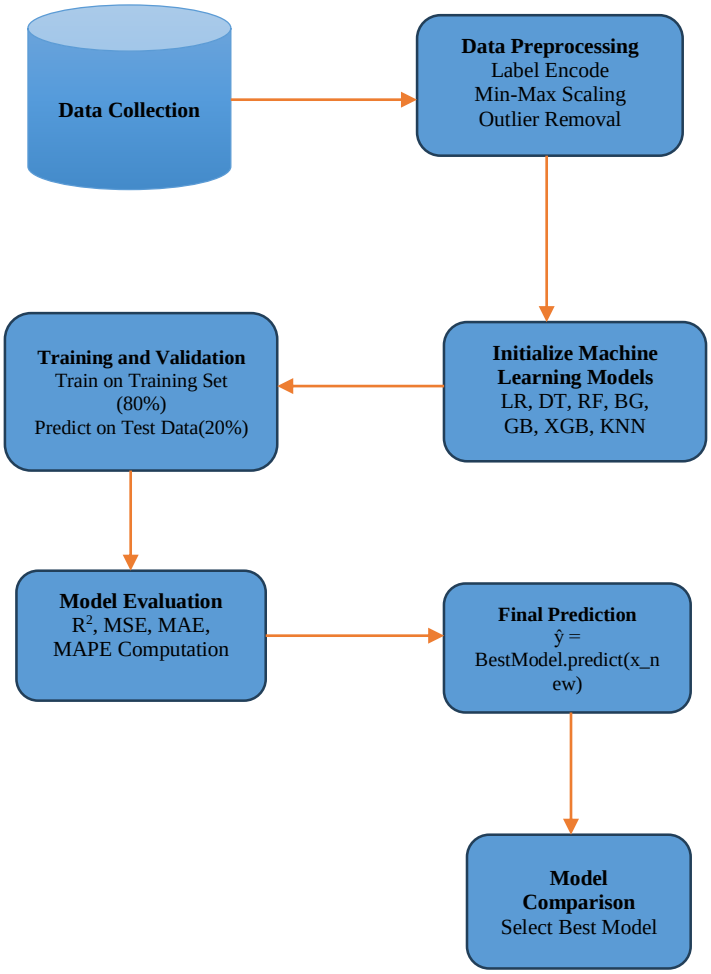

Fig. 1 Overall research framework of the CYP

Algorithm: CYP Framework

4.1. Research Framework

The research framework proposed for conducting the study on CYP is systematic and features a workflow that incorporates data acquisition, preprocessing, model training, evaluation, and prediction. The conceptual flow of the process is shown in Figure 1 below. Also, the research framework is given in Algorithm 1.

Algorithm:
Input:

Raw dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$,
where $x^{(i)}$ ={Area, Item, Year, Rainfall, Temperature, Pesticides}
and $y^{(i)} =$ yield (hg/ha)
ML Models:
$\mathcal{M}$ ={LR, DT, RF, BG, GB, XGB, KNN}

Output:

Best model $\mathcal{M}^*$
Performance metrics: R^2, MSE, MAE, MAPE
Predicted yield values $\hat{y}$

Algorithm

Dataset Collection: Raw dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$
Data Preprocessing

Encode categorical variables $\text{Area}_{\text{enc}}^{(i)} = \ell_A(\text{Area}^{(i)}), \text{Item}_{\text{enc}}^{(i)} = \ell_I(\text{Item}^{(i)})$
Normalize numerical features $X_{\text{scaled}}^{(i)} = \frac{X^{(i)} - X_{\min}}{X_{\max} - X_{\min}}$
Remove outliers $X^{(i)} \notin [Q1 - 1.5\,\text{IQR},\ Q3 + 1.5\,\text{IQR}] \Rightarrow$ discard

Split dataset $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test}, | \mathcal{D}_{test} | = 0.2N$
Train each model $\mathcal{M}_k = \arg\min_{\theta_k} \sum_i (y^{(i)} - \mathcal{M}_k(\tilde{x}^{(i)}))^2$
Predict on the test set. $\hat{y}_k^{(i)} = \mathcal{M}_k(\tilde{x}_{\text{test}}^{(i)})$
Evaluate metrics R_k^2, MSE_k, MAE_k, MAPE_k
Select the best model. $\mathcal{M}^* =$ $\arg\max_k R_k^2$ with minimal error
Final prediction $\hat{y}_{new} = \mathcal{M}^*(x_{\text{new}})$

4.2. Data Preprocessing

Preprocessing of data is crucial in preparing the data for use in machine learning ML). It assists in removing inconsistencies and ensures that the input features are properly formatted for ingestion by the model. There were the following preprocessing operations:

4.2.1. Handling of Categorical Variables (Area and Item)

The dataset has two categorical variables, i.e., Area (country name) and Item (crop type). These variables represent contextual information on geographical and biological diversity; however, most machine learning ML) algorithms cannot interpret these variables directly. As such, the categorical encoding was done through the process of Label Encoding, which assigns each category a unique numerical value. For example, other crops like "Maize," "Wheat and Potatoes were translated into numerical labels. The transformation enables the models to identify country and crop differences without increasing dimensionality, as would be the case with one-hot encoding.

Scaling of Numerical Attributes: Numerical attributes can be scaled to a specific range, resulting in a continuous numerical variable. The attributes *average_rainfall_mm_per_year*, *pesticides_tonnes*, *avg_temp*, and *Year* were normalized using Min-Max Scaling. The scaling operation converts all numerical variables to a normalized scale (0 to 1), thus not causing models such as K-Nearest Neighbors or GB to be biased towards variables that have larger numerical values. Mathematically:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

4.2.2. Outlier Detection and Removal

Since the dataset was broad in terms of geography and climate, it contained possible outliers, including extreme values of rainfall, pesticide use, or yield. To reduce their effect, simple statistical thresholding (using the interquartile range) was employed to identify and eliminate anomalies that exceeded the 99th percentile. This made the model's learning process stable, unaffected by unrealistic and erroneous data points.

4.2.3. Data Consistency and Validation

The final validation test ensured that all 28,242 records were complete and that none were missing or contained a null value. Following the preprocessing, the data was split into training and testing sets so as to continue with model development.

4.3. Machine Learning Models

The proposed study uses seven regression-based ML algorithms, which are supervised to predict crop yield using climatic and environmental factors. The algorithms capture various relationships and levels of data complexity. The mathematical expressions of the models are as follows.

4.3.1. Linear Regression

One of the most straightforward and most interpretable predictive models is the LR. It presupposes the linear connection between the dependent variable (crop yield) and the independent variables (rainfall, temperature, pesticide usage, etc.) [20, 21].

The model can be modeled mathematically as:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (2)$$

Where:

- $\hat{y}$= predicted crop yield (hg/ha),
- β_0= intercept term,
- β_i= coefficients of feature x_i,
- x_i= input features (e.g., rainfall, temperature, pesticide use),
- ϵ= random error term.

The model parameters β_iAre estimated using Ordinary Least Squares (OLS) by minimizing the residual sum of squares:

$$\text{minimize } \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \quad (3)$$

Despite its computing efficiency, LR is inefficient in the presence of nonlinear data such as agricultural data, whose yield and environmental relationships are nonlinear [22].

4.3.2. Decision Tree

DT Regressor predicts the crop yield by splitting the feature space recursively into regions in which the target value (Yield) is essentially the same [21]. It is a nonparametric, nonlinear model that can be used to identify intricate associations among rainfall, temperature, pesticide application, crop type, and yield [23, 24].

Given dataset:

$$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}, \quad (4)$$

Where

$$x(i) = [Area,\ Item,\ Year,\ r,\ p,\ t](i),$$
$$y(i) = Yield\ (hg/ha).$$

Goal: learn a function

$$\hat{y} = f(x) \quad (5)$$

That predicts yield based on environmental and crop features.

At each node, the tree chooses feature $j$and threshold $s$that best split the data into two child nodes:
Left child:

$$R_L(j,s) = \{(x,y): x_j \le s\} \quad (6)$$

Right child:

$$R_R(j,s) = \{(x,y): x_j > s\} \quad (7)$$

The optimal split minimizes the Sum of Squared Errors (SSE) or variance:

$$(j^*, s^*) = \arg\min_{j,s}[\tfrac{|R_L|}{|D|}\mathrm{Var}(R_L) + \tfrac{|R_R|}{|D|}\mathrm{Var}(R_R)] \quad (8)$$

Where

$$\mathrm{Var}(R) = \tfrac{1}{|R|}\sum_{i\in R}(y^{(i)} - \bar{y}_R)^2 \quad (9)$$

Once the tree assigns a set of samples to a leaf region R_m, the predicted crop yield for all points in that region is:

$$\hat{y}(x) = \bar{y}_{R_m} = \tfrac{1}{|R_m|}\sum_{i\in R_m} y^{(i)} \quad (10)$$

Thus, the prediction is the mean yield of all training samples in that leaf.

Formally, stop if:

$$\mathrm{Var}(R) < \epsilon \text{or } |\ R\ | < \text{min\_samples\_leaf} \quad (11)$$

The DT model is the sum of predictions over all leaf regions:

$$f(x) = \sum_{m=1}^{M} \bar{y}_{R_m}\ \mathbf{1}(x \in R_m) \quad (12)$$

4.3.3. Random Forest

RF is an ensemble regression model, which builds a series of DT through bootstrapped data and random selection of features at each split [25]. The trees are independent predictors of crop yield, and the ultimate prediction is achieved by averaging the results of all trees. This reduces the variance and improves accuracy compared to an individual DT [26]. The model helps capture nonlinearity between climatic and environmental factors that affect crop yield. If $T_1, T_2, \ldots, T_B$Represent Bindividual DT, the RF prediction is given by:

$$\hat{y} = \tfrac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (13)$$

Trees are trained on a random sample (bootstrap) of the data, and a random selection of features is employed at every split, bringing variety to the trees. This method is effective in minimizing overfitting and enhancing the capacity to generalize [27].

4.3.4. Bagging Regressor (Bootstrap Aggregating)

Another ensemble method that enhances the stability of the model is Bagging, which involves the combination of various estimators that are trained on various bootstrap samples of the dataset [28].

1. Let the dataset be
$$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}, y^{(i)} = \text{crop yield (hg/ha)}.$$
2. Choose the number of base models (trees) B.
3. For each model $b = 1, \ldots, B$, draw a bootstrap sample
$$\mathcal{D}_b \sim \text{Bootstrap}(\mathcal{D}).$$
4. Train a base regressor (DT)
$$T_b = \text{TrainTree}(\mathcal{D}_b).$$
5. Each tree recursively minimizes node variance:
$$(j^*, s^*) = \arg\min_{j,s}[\frac{|\ R_L\ |}{|\ R\ |}\mathrm{Var}(R_L) + \frac{|\ R_R\ |}{|\ R\ |}\mathrm{Var}(R_R)].$$
6. At each leaf region R_{bm}, tree prediction is
$$T_b(x) = \frac{1}{|\ R_{bm}\ |}\sum_{i\in R_{bm}} y^{(i)}.$$
7. Bagging prediction for any crop-yield input $x$is
$$\hat{y}_{Bag}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$
8. Since bootstrap samples differ, individual trees are decorrelated.
9. Variance reduction due to averaging:

$$\mathrm{Var}[\hat{y}_{Bag}] = \frac{1}{B}\mathrm{Var}[T] \text{(if independent)}.$$

10. Final predicted crop yield for a new sample x_{new}:
$$\hat{y}_{yield} = \hat{y}_{Bag}(x_{\text{new}}).$$

4.3.5. Gradient Boosting

Let $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ with $y^{(i)}$= crop yield (hg/ha).

1. Initialize model with a constant (stage 0):
$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y^{(i)}, \gamma).$$
2. For $m = 1, \ldots, M$(number of boosting rounds) compute pseudo-residuals:
$$r_{im} = -\frac{\partial L\left(y^{(i)}, F\left(x^{(i)}\right)\right)}{\partial F(x^{(i)})}\Big|_{F=F_{m-1}}, \qquad i = 1, \ldots, N.$$
3. Fit a weak learner $h_m(x)$(e.g., shallow regression tree) to $\{(x^{(i)}, r_{im})\}$by minimizing the squared error of residuals.
4. Optionally compute optimal step size. γ_mBy line search:
$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{N} L(y^{(i)}, F_{m-1}(x^{(i)}) + \gamma\, h_m(x^{(i)})).$$
5. Update model with learning rate $\nu \in (0,1]$:
$$F_m(x) = F_{m-1}(x) + \nu\, \gamma_m\, h_m(x).$$
6. Repeat steps 2–5 until $m = M$(or early stopping via validation loss).
7. Final ensemble predictor after Mrounds:
$$F_M(x) = F_0(x) + \nu \sum_{m=1}^{M} \gamma_m\, h_m(x).$$
8. For squared-error loss $L(y, F) = \frac{1}{2}(y - F)^2$, residuals simplify to $r_{im} = y^{(i)} - F_{m-1}(x^{(i)})$and γ_mIt is the least-squares fit coefficient.
9. Evaluate on test set with metrics (e.g., MSE, MAE, MAPE, R^2) using $\hat{y} = F_M(x)$(units: hg/ha).
10. Final crop-yield prediction for new input x_{new}:
$$\hat{y}_{\text{yield}} = F_M(x_{\text{new}})$$

4.3.6. XGBoost (Extreme Gradient Boosting)

XGBoost extends GB with additional regularization and optimization improvements [29]. It minimizes an objective function that balances accuracy and model complexity:

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where:

- $l(y_i, \hat{y}_i)$= differentiable convex loss function (e.g., MSE),
- $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$= regularization term,
- T= number of leaves in the tree,
- λ= L2 regularization parameter,
- w= leaf weights.

Each new tree $f_t(x)$is added to minimize the loss using the second-order Taylor expansion:

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

Where $g_i$and h_iThese are the first and second derivatives (gradients and Hessians) of the loss function. This second-order optimization and built-in regularization make XGBoost one of the most efficient and accurate boosting methods.

4.3.7. K-Nearest Neighbors (KNN)

KNN is a nonparametric, instance-based learning algorithm [30]. It predicts the target value for a new data point based on the average yield of its *k* nearest neighbors in the training dataset.

1. Choose an integer k(number of neighbors).
2. For a new input x, compute the Euclidean distance to all training points:
$$d(x, x^{(i)}) = \sqrt{\sum_{j=1}^{d} (x_j - x_j^{(i)})^2}.$$
3. Sort all distances $d(x, x^{(i)})$in ascending order.
4. Select the set $\mathcal{N}_k(x)$of the $k$nearest neighbors.
5. Retrieve the corresponding yield values:
$$\{y^{(i)} : x^{(i)} \in \mathcal{N}_k(x)\}.$$
6. Compute the KNN prediction as the mean yield of neighbors:
$$\hat{y}_{KNN}(x) = \frac{1}{k} \sum_{x^{(i)} \in \mathcal{N}_k(x)} y^{(i)}.$$
7. If using distance-weighted KNN, weight by inverse distance:

$$\hat{y}_w(x) = \frac{\sum_{i \in \mathcal{N}_k(x)} \frac{1}{d(x, x^{(i)})} y^{(i)}}{\sum_{i \in \mathcal{N}_k(x)} \frac{1}{d(x, x^{(i)})}}.$$

8. Normalize continuous features (rainfall, temperature, pesticides) to avoid scale bias.
9. Encode categorical features (Area, Item) using label or one-hot encoding.
10. The final crop-yield prediction for any new data point x_{new}:

$$\hat{y}_{\text{yield}} = \hat{y}_{KNN}(x_{\text{new}}).$$

4.4. Model Parameters and Configuration

The parameter configuration was standardized across all models to ensure fair comparison and reproducibility. Key configurations are summarized below in Table 3:

Table 3. Key parameters applied

Model	Key Parameters
Linear Regression	Default parameters
Random Forest	random_state = 42
Gradient Boosting	n_estimators = 100, learning_rate = 0.1, max_depth = 3, random_state = 42
XGBoost	random_state = 42
KNN	n_neighbors = 5
Decision Tree	random_state = 42
Bagging	n_estimators = 150, random_state = 42

To maintain consistency, the train-test split ratio was kept at 80/20 for training and testing. In addition, 5-fold cross-validation (k=5) was used to reduce bias and ensure that every model was tested on several subsets of the dataset. The method enhances the generalization of the models, since the performance is tested to be stable between different data partitions.

4.5. Model Training and Validation

The implementation of all the models was made in Python 3.10 and its data science ecosystem, which included Scikit-learn, XGBoost, Pandas, NumPy, and Matplotlib. The training and validation steps used included the following:

4.5.1. Training Phase

Each model was trained on 80% of the dataset using preprocessed features and the target variable (*hg/ha_yield*).

4.5.2. Validation Phase

Models were validated on the 20% test dataset. Cross-validation results were recorded for accuracy consistency. Predictions were generated and compared with actual yield values.

4.5.3. Performance Evaluation

Predicted and actual values were analyzed to compute performance metrics (R^2, MSE, MAE, and MAPE). Results were tabulated and visualized to facilitate model comparison.

4.6. Evaluation Metrics

To evaluate model performance comprehensively, four key statistical metrics were utilized [22, 31]:

4.6.1. R^2 Score

- Measures how well the model explains the variance in the dependent variable.
- Higher values (closer to 1) indicate better performance and stronger predictive capability.

4.6.2. Mean Squared Error (MSE)

- Quantifies the average squared difference between actual and predicted values.
- Lower MSE signifies higher model accuracy and fewer significant prediction errors.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (14)$$

4.6.3. Mean Absolute Error (MAE)

- Represents the average magnitude of absolute differences between predicted and actual yields.
- Provides an intuitive measure of average model deviation.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (15)$$

4.6.4. Mean Absolute Percentage Error (MAPE)

- Expresses prediction error as a percentage of actual yield values.
- Enables easy interpretability across scales.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (16)$$

The comparison of several metrics allows assessing the models reliably by measuring accuracy (R2) and error size (MSE, MAE, MAPE). The multi-metric design helps to avoid the over-dependence on one measure and to have a more balanced picture of the model functionality in various conditions.

5. Results and Analysis

In this section, the experimental results obtained through the use of seven machine learning ML) regression models to predict crop yields are presented and analyzed.

Every model was tested in a uniform experimental setup, an 80-20 train-test split, five-fold cross-validation, and four performance metrics were used, such as R^2 Score, MSE, MAE, and MAPE.

5.1. Model Performance Comparison

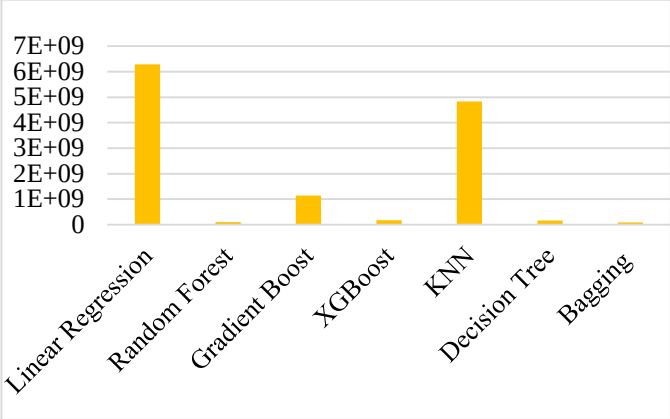

Fig. 2 MSE of different ML models in CYP

The comparison of the MSE of ML models to predict crop yield is presented in Figure 2. MSE is used to determine the extent to which the actual values of the predicted yield are close to the actual values, and a smaller MSE means a good prediction. The findings clearly indicate that the best EL models based on decision-tree architecture are better than the traditional regression models since they yield much fewer squared errors. Bagging, RF, and DT are the models that have a high ability to generalize and effectively reduce prediction errors. Conversely, less complex models, such as LR and KNN, have significantly larger error values, which confirms their inability to learn complicated nonlinear relationships in agricultural data effectively. GB and XGBoost are average in performance, yet they continue to produce more errors than the best-performing ensemble techniques. All in all, the MSE comparison confirms the fact that tree-based ensemble models are the most accurate and consistent in their yield estimation and thus are most effective in real-world agricultural forecasting.

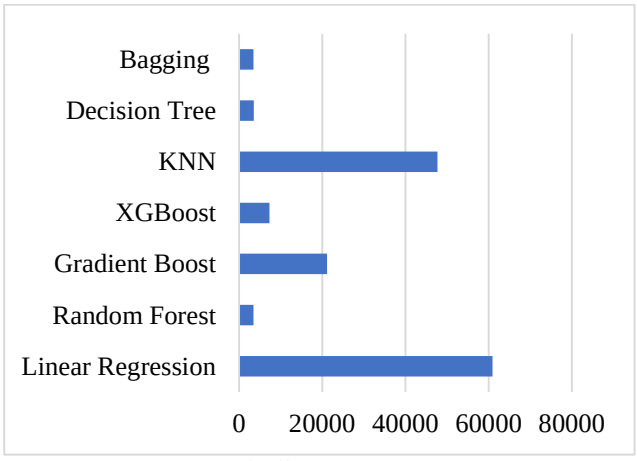

Fig. 3 MAE of different ML models in CYP

The values of the MAE of the various ML models are reported in Figure 3. MAE is the mean value of the errors in the estimated crop yield, and it is quantified in the same unit as the target variable. A more petite MAE indicates a higher predictive quality and less variation of the actual yields. The Bagging Regressor had the least MAE (3,450.50), followed closely by the RF (3,480.84) and DT (3,559.26), indicating that these models have high accuracy in estimating yields with low average error. XGBoost also exhibited a fair performance of an MAE of 7341.94, but not as accurate as the best ensemble models. On the contrary, the values of MAE in LR and KNN were much larger (60955.31 and 47716.35, respectively), which proves that these models are more likely to miss the nonlinear and complex relationships within the data. GB was fairly good but still demonstrated a relatively high error relative to the top ensemble methods. In general, the MAE comparison also confirms that tree-based ensemble models are more precise in yield prediction and can be more effectively applied to the agricultural forecasting case.

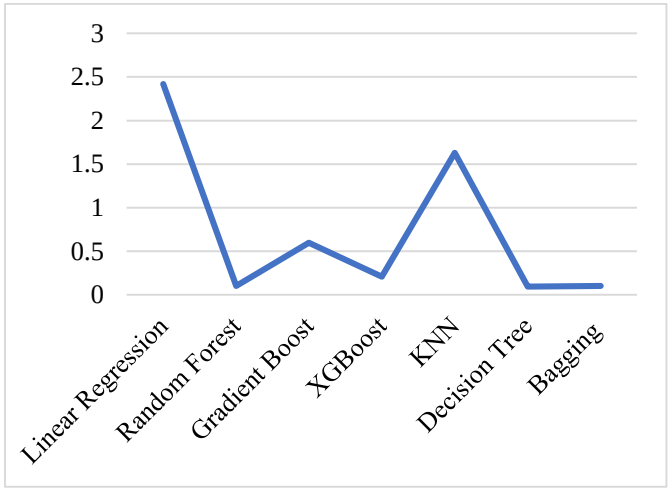

Fig. 4 MAPE of different ML models in CYP

Figure 4 illustrates the MAPE results for the tested models. MAPE is a decipherable measure that defines the error of prediction as a percentage, with lower scores indicating better results. The findings indicate that the DT model provided the lowest MAPE (0.096101), closely preceded by Bagging (0.101199) and RF (0.102571), which means that they have high predictive accuracy. These findings indicate that the prediction error of these models is lower than 0.11 percent, which is very tolerable in agricultural prediction. LR and KNN, on the other hand, had very large values of MAPE (2.419536 and 1.631186, respectively), which implies significant inaccuracy and proves that they are not able to model the complex interactions in the data. GB and XGBoost exhibited moderate performance, but were still not as accurate as the most successful ensemble tree-based models. Comprehensively, these findings from the MAPE do reinforce the fact that EL models provide more accurate and effective forecasts of crop yields and are therefore suitable for real-life applications in agricultural decision-support systems.

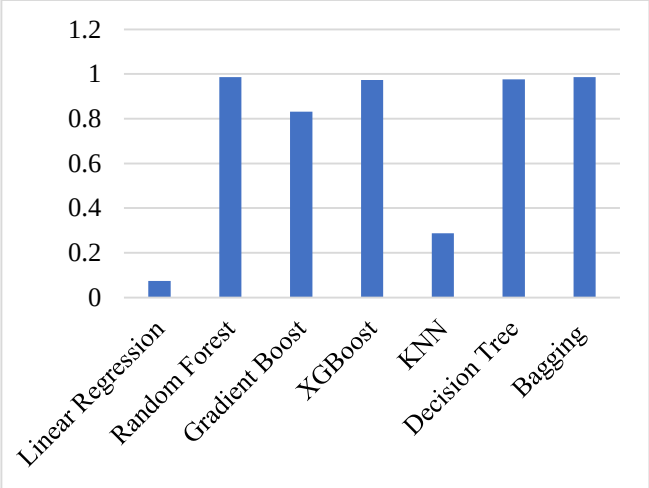

Fig. 5 R^2 of different ML models in CYP

Figure 5 presents the performance of the ML models in crop yield prediction in terms of R^2 score. R^2 value is the percentage of the change in the yield of crops that is attributed to the input features. The higher the value, the more accurate and reliable the predictions are. The highest R^2 value of 0.985881 was obtained by the Bagging Regressor, with a close second value of 0.985628, indicating the high predictive power and the high generalization of the model. The level of accuracy was also high (0.976174) in the DT model, which proved the usefulness of tree-based learning methods to model nonlinear agricultural data. LR and KNN models, on the contrary, had significantly lower values of R^2 (0.073724 and 0.288206, respectively), which means that they failed to represent the intricate associations between climatic variables and crop yield. XGBoost and GB had moderate performance and were still lower than the best ensemble models because they had a relatively high prediction variance. These findings support the fact that ensemble tree models are most appropriate in the prediction of crop yields because they are able to cope with environmental variability, nonlinearity, and feature interaction in agricultural data.

5.2. Key Observations

5.2.1. Best Performers

Bagging (R^2 = 0. 985881) and RF (R^2 = 0. 985628) had almost the same best accuracy. The two models help capture nonlinearities and multivariate dependencies.

5.2.2. Moderate Performers

DT (R^2 = 0.9761) and XGBoost (R^2 = 0.9732) were also good but not the best ensemble methods because of variance and tuning effects.

5.2.3. Weak Performers

KNN and gradient Boost had high variance and dimensionality issues. LR had the lowest adaptability as it assumes the existence of linear relationships.

5.2.4. Overall Observation

Ensemble models that used trees were shown to be most effective in predicting in different environmental conditions and species of crops.

5.3. Summary of Findings

Ensemble models work better than simple regression and distance-based approaches because they learn nonlinear and more complicated relationships between climate characteristics and crop yield. The Bagging Regressor had the minimum overall prediction error and maximum stability, and this proves its applicability in real-life agricultural forecasting. The mathematical analysis of error decomposition proves that ensemble averaging is an effective method of model variance, and it is optimal in the case of heterogeneous agricultural data.

6. Discussion

6.1. Key Findings

The findings of this paper are a clear indication that ML models built by an ensemble are a far better predictor of crop yield than the conventional and simple regression models. The Bagging Regressor and the RF were among the seven models assessed and had the highest predictive accuracy with the R^2 scores of 0.985881 and 0.985628, respectively.

The lowest error rates were also generated by these models, with values of MSE of 9.59 x 10^7 and 9.76 x 10^7, meaning that the models have an exceptional accuracy in estimating the yield. Tree-based ensemble models are used to effectively model a complex and nonlinear interaction between agricultural inputs like rainfall, temperature, and pesticide application. Their capacity to pool forecasts of numerous weak learners (DT) lowers variance and avoids overfitting, which is essential in heterogeneous agricultural data of varied climatic areas and crops.

Conversely, other models, including LR and KNN, did not work well, having the R^2 of 0.0737 and 0.2882, respectively. XGBoost (R^2 = 0.9732) and DT (R^2 = 0.9761) were also competitive in terms of results, but not better than ensemble bagging methods in terms of stability or generalization. All in all, the comparative analysis established that the EL methods, especially Bagging and RF, provide strong and stable yield predictions in diverse agricultural settings, which can be considered the best option to implement them in the real-life context of precision agriculture systems.

6.2. Practical Applications and Use Cases

The applied implications of these results are at various levels of the agricultural system, such as in policy making, making farm-level decisions, and in the agri-business process.

6.2.1. Government and Policy Planning

Correct ML-based yield prediction models can be used by governments to forecast food supply, regulate food imports and exports, and design agricultural policies. An ability to predict the performance of yields in regions will be helpful to authorities in predicting shortages or surpluses, efficient food distribution, and reducing drought or flood risks. Moreover, these models can be used to inform the allocation of subsidies, irrigation, and food security measures at the national level.

6.2.2. Agricultural Practitioners and Farmers

In the case of farmers, yield prediction models will enable them to have actionable intelligence on how to optimize agricultural practices.

ML models could be used. Choose the right species of crops that are adapted to the existing climatic and soil conditions. Optimize the use of fertilizers and pesticides, reducing the damage to the environment and maximizing the yield. Schedule irrigation schedules depending on the predictions of the rainfall and yield forecasts. Such insights are helpful in precision farming, enabling farmers to make informed and data-driven decisions to make farming more profitable and sustainable.

6.2.3. Agri-Tech Companies and Researchers.

Agri-tech companies will be able to incorporate the ML-based prediction models into the digital farming platform, mobile apps, and decision-support systems. These systems can provide real-time predictions on yields, weather, and cultivation advice specific to the regions or types of crops.

Moreover, ML advances agricultural research by making it possible to simulate crop performance in different climatic conditions to facilitate innovation in sustainable food production technologies. To conclude, ML-based yield prediction systems can revolutionize the agricultural sector by providing data-driven information to all stakeholders, including policymakers and farmers, promptly.

6.3. Addressing Challenges

Although this study produced good prediction results, a number of challenges and limitations were identified:

6.3.1. Data Imbalance and Regional Bias

The sample does not have the same number of records per country and crop (e.g., India and Potatoes take the leading positions in the sample). This may bias model learning towards commonly represented areas or types of crops and hinder generalization to less commonly represented data samples.

6.3.2. Missing Agronomic Variables

The dataset lacks important agronomic variables, such as soil fertility, irrigation, and type of fertilizer, which have been proven to have a significant impact on yield. Lack of these makes model predictions incomplete.

6.3.3. Model Interpretability

Climatic changes and extreme weather conditions (e.g., droughts, floods, heatwaves) add high temporal variability in yield patterns. These new changes may not be well predicted by the historical dataset (1990-2013), compromising long-term predictive performance.

6.3.4. Model Interpretability

Despite the high performance of ensemble models, they are complex black-box systems, and thus, it is challenging for non-technical stakeholders, such as farmers or policymakers, to interpret what individual predictions represent.

6.3.5. Data Granularity

The dataset is mainly run at the country level of aggregation, as opposed to field-level data. This lowers the spatial resolution and can hide the micro-level differences that can be used in localized decision-making. To solve these issues, more detailed datasets and more explainable AI methods will be needed, which can enhance the interpretability and reliability of models.

6.4. Model Complexity vs Interpretability

There is a trade-off between model performance and interpretability, which is critical.

Although ensemble models like Bagging and the RF had the best R^2, they are more complex and opaque. They are the product of a large number of FT joined together, so it is not easy to track the contribution of each of the input variables to the output. LR, on the contrary, is fully transparent and can be easily interpreted, but does not work well with nonlinear data in terms of prediction accuracy. This trade-off implies that high-accuracy models are preferable for use in operational deployment. However, explainable and interpretable AI systems are necessary for stakeholders to adopt.

To address these points, future work needs to consider Explainable AI (XAI) methods, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), to visualize and interpret feature contributions. Such practices can help stakeholders understand why certain predictions were achieved, thereby increasing their confidence in AI-guided agricultural recommendations.

6.5. Role of Evaluation Metrics in Model Selection

The assessment of regression models across multiple performance measures is critical for achieving a holistic understanding of the models' behavior. Although the R^2 Score is used to measure the extent to which the model can

account for the variation in crop yield, it by itself does not give the extent or magnitude of prediction errors.

For example, two models can be similar in terms of R-squared but differ significantly in terms of absolute errors. Therefore, the complementary error measures, i.e., MSE, MAE, and MAPE, were employed. MSE lends a quadratic penalty to large errors and rewards models that occasionally make extreme deviations.

MAE is an easily understood average of the magnitude of absolute errors, which is good in determining the overall reliability of the prediction. MAPE, which is in percent form, is interpretable, and therefore, the results of model errors can be easily compared across scales. The combination of these metrics also provided a balanced model assessment, indicating that the ensemble models had high explanatory power and low error deviation. Therefore, the multi-metric test confirmed the effectiveness of EL models and avoided over-dependence on a particular indicator.

6.6. Computational Efficiency and Scalability

The scale of ML models to large-scale or real-time agricultural applications directly depends on their computational complexity. LR and KNN are not only computationally inexpensive but also do not offer sufficient accuracy for complex data. DT and the RF tree-based models are moderate in terms of training time but exhibit significant predictive accuracy. Ensemble techniques (Bagging, GB, XGBoost) are computationally expensive because of model training and aggregation (repeated training and aggregation), particularly when using very large datasets.

In order to make it more scalable, subsequent studies may utilize cloud-based ML solutions and accelerate training models with the help of GPUs. Moreover, offline-trained lightweight models may be deployed to mobile or edge devices and used to make real-time predictions for farmers with limited computational capacity.

6.7. Data Availability Constraints

The limitation of the data diversity is one of the major limitations of this study. Despite having climatic and yield data for more than 100 countries, the dataset lacks several important agronomic factors, including soil type, irrigation level, fertilizer composition, and crop genetics.

The lack of these parameters restricts the ability of these models to explain variation in crop yield completely, especially in areas where non-climatic factors are the major ones. In order to overcome this, future research should focus on incorporating multisource agricultural data, such as:

IoT sensor data: in soil moisture, pH, and nutrient levels. Satellite and remote sensing data: to track vegetation indices and the land-use patterns.

Weather station networks: to measure real-time climatic variables such as humidity, speed of wind, and solar radiation.

By combining all these data streams with ML and AI, it is possible to produce more granular, real-time, and region-specific models of crop yield predictions.

7. Conclusion and Future Work

The paper examined the performance of different ML models to predict crop yield based on environmental and climatic factors, including rainfall, temperature, and pesticide use. A total of 28,242 records were used to test seven machine learning ML) models under a set of consistent experimental conditions. The findings indicate that EL models outperform traditional and distance-based methods. The Bagging had the best predictive performance with an R^2 equal to 0.985881, and the RF closely follows it with an R^2 of 0.985628. The models also yielded extremely low error values, indicating their strength and applicability in predicting yield in various agricultural settings.

The paper also adds to the field of agricultural analytics by conducting a comprehensive dataset characterization, comparing dozens of ML models, and finding ensemble-based techniques to be the most effective at estimating crop yields with reasonable accuracy. The variables observed to be the dominant predictors of yield variability were rainfall, temperature, and pesticides, underscoring the significance of environmental dynamics in agricultural modeling. The results highlight the power of ensemble models in capturing nonlinear relationships and reducing prediction variance. Therefore, they could be widely used in practice and decision support in agriculture. Nevertheless, several weaknesses were identified. The dataset lacked important agronomic factors, including soil fertility, irrigation rates, and crop genetics, which limits the integrity of the prediction framework. The imbalance on the regional level, particularly the prevailing status of such nations and the timeframe ending in 1990-2013, can also be reasons to question the model's applicability to current climatic conditions. Additionally, although ensemble models are highly accurate, they cannot be fully interpreted without the aid of additional explainability methods. Large-scale or real-time deployment is also problematic due to computational intensity.

Further studies are needed to examine hybrid structures that combine deep learning and ensemble methods, enabling models to leverage both hierarchical feature abstraction and variance reduction. Temporal responsiveness can be enhanced by integrating real-time sensor data from IoT, allowing crop yield predictions to become more responsive and field-specific. Techniques such as SHAP and LIME should be employed as explainable AI methods to enhance the transparency and trust of stakeholders. It is possible to

solve the imbalance in datasets and localized predictions by creating region-specific or crop-specific models. Also, the environmental signals can be extended using satellite and remote sensing data, including NDVI and soil moisture indices, to forecast yields at a higher resolution.

To conclude, ensemble learning, especially Bagging and RF, is a powerful and precise model to predict the crop yield. As data diversity, model interpretability, and real-time integration are improved, machine learning ML) can make a significant contribution to precision agriculture and help ensure food security worldwide.

Data Availability

Data is publicly accessible at https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset/data

References

[1] Sandeep Gupta et al., "Machine Learning-and Feature Selection-Enabled Framework for Accurate Crop Yield Prediction," *Journal of Food Quality*, vol. 2022, no. 1, pp. 1-7, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Mamunur Rashid et al., "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. [CrossRef] [Google Scholar] [Publisher Link]

[3] Alejandro Morales, and Francisco J. Villalobos, "Using Machine Learning for Crop Yield Prediction in the Past or the Future," *Frontiers in Plant Science*, vol. 14, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Janmejay Pant et al., "Analysis of Agricultural Crop Yield Prediction using Statistical Techniques of Machine Learning," *Materials Today: Proceedings*, vol. 46, no. 20, pp. 10922-10926, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Javad Ansarifar, Lizhi Wang, and Sotirios V. Archontoulis, "An Interaction Regression Model for Crop Yield Prediction," *Scientific Reports*, vol. 11. pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] D. Jayanarayana Reddy, and M. Rudra Kumar, "Crop Yield Prediction Using Machine Learning Algorithm," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 1466-1470, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Kavita Jhajharia et al., "Crop Yield Prediction using Machine Learning and Deep Learning Techniques," *Procedia Computer Science*, vol. 218, pp. 406-417, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Matin Kuradusenge et al., "Crop Yield Prediction using Machine Learning Models: Case of Irish Potato and Maize," *Agriculture*, vol. 13, no. 1, pp. 1-19, 2023 [CrossRef] [Google Scholar] [Publisher Link]

[9] Anikó Nyéki, and Miklós Neményi, "Crop Yield Prediction in Precision Agriculture," *Agronomy*, vol. 12, no. 10, pp. 1-4, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Hannah Burdett, and Christopher Wellen, "Statistical and Machine Learning Methods for Crop Yield Prediction in the Context of Precision Agriculture," *Precision Agriculture*, vol. 23, pp. 1553-1574, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Thomas van Klompenburg, Ayalew Kassahun, and Cagatay Catal, "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review," *Computers and Electronics in Agriculture*, vol. 177, pp. 1-18, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] S. Iniyan, V. Akhil Varma, and Ch Teja Naidu, "Crop Yield Prediction Using Machine Learning Techniques," *Advances in Engineering Software*, vol. 175, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Priyanga Muruganantham et al., "A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing," *Remote Sensing*, vol. 14, no. 9, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Petteri Nevavuori, Nathaniel Narra, and Tarmo Lipping, "Crop Yield Prediction with Deep Convolutional Neural Networks," *Computers and Electronics in Agriculture*, vol. 163, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[15] Saeed Khaki, Lizhi Wang, and Sotirios V. Archontoulis, "A CNN-RNN Framework for Crop Yield Prediction," *Frontiers in Plant Science*, vol. 10, pp. 1-14, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Saeed Khaki, and Lizhi Wang, "Crop Yield Prediction using Deep Neural Networks," *Frontiers in Plant Science*, vol. 10, pp. 1-10, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] P.S. Maya Gopal, and R. Bhargavi, "A Novel Approach for Efficient Crop Yield Prediction," *Computers and Electronics in Agriculture*, vol. 165, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ramesh Medar, Vijay S. Rajpurohit, and Shweta Shweta, "Crop Yield Prediction using Machine Learning Techniques," *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Bombay, India, pp. 1-5, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Crop Yield Prediction Dataset, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset/data

[20] Seema Sharma et al., "Enhancing Crop Yield Prediction Through Machine Learning Regression Analysis," *International Journal of Sustainable Agricultural Management and Informatics*, vol. 11, no. 1, pp. 29-47, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yueru Yan et al., "Crop Yield Time-Series Data Prediction based on Multiple Hybrid Machine Learning Models," *arXiv preprint*, pp. 1-7, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[22] El-Sayed M. El-Kenawy et al., "Predicting Potato Crop Yield with Machine Learning and Deep Learning for Sustainable Agriculture," *Potato Research*, vol. 68, pp. 759-792, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[23] Kalpesh S. Borse, and Prasit Agnihotri, "Application of Decision Tree (M5Tree) Algorithm for Multicrop Yield Prediction of the Semi-Arid Region of Maharashtra, India," *Journal of Soft Computing in Civil Engineering*, vol. 9, no. 1, pp. 64-90, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[24] R.N.V. Jagan Mohan, Pravallika Sree Rayanoothala, and R. Praneetha Sree, "Next-gen Agriculture: Integrating AI and XAI for Precision Crop Yield Predictions," *Frontiers in Plant Science*, vol. 15, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[25] Sudhakar Uppalapati et al., "Precision Biochar Yield Forecasting Employing Random Forest and XGBoost with Taylor Diagram Visualization," *Scientific Reports*, vol. 15, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[26] Juan Carlos Moreno Sánchez et al., "Improving Wheat Yield Prediction through Variable Selection Using Support Vector Regression, Random Forest, and Extreme Gradient Boosting," *Smart Agricultural Technology*, vol. 10, pp. 1-10, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[27] Yashraj Patil et al., "Comparative Analysis of Machine Learning Models for Crop Yield Prediction across Multiple Crop Types," *SN Computer Science*, vol. 6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] V. Ramesh, and P. Kumaresan, "Stacked Ensemble Model for Accurate Crop Yield Prediction Using Machine Learning Techniques," *Environmental Research Communications*, vol. 7, no. 3, pp. 1-17, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Mildred Virginia López Segura et al., "XGBoost Sequential System for the Prediction of Persian Lemon Crop Yield," *Crop Science*, vol. 65, no. 1, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[30] P. Phanindra Sai Kumar, and T. Rajendran, "Higher Accuracy of Crop Recommendation using Random Forest Algorithm over K-Nearest Neighbors Algorithm," *AIP Conference Proceedings*, vol. 3267, no. 1, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[31] Himanshu Pant et al., "Comparative Study of Crop Yield Prediction Using Explainable AI and Interpretable Machine Learning Techniques," *2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, Bhilai, India, pp. 1-7, 2025. [CrossRef] [Google Scholar] [Publisher Link]