THE SCIENTIFIC TEMPER (2026) Vol. 17 (4): 6166-6183
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2026.17.4.27
https://scientifictemper.com/

RESEARCH ARTICLE

IoT Aware Polynomial Regressive Ensemble Artificial Intelligence Model for Crop Yield Prediction in Cloud Computing Environment

P. Gayathri[1*], Dr. C. Jayanthi[2]

Abstract

Accurately estimating crop yields across large geographical regions is essential for ensuring food security and promoting sustainable development. The machine learning driven Artificial Intelligence (AI) based approaches have emerged for enhancing the precision of agricultural yield predictions. While many existing models have achieved higher accuracy by increasing their complexity, their ability to generalize remains limited due to variations in key features across different regions. To address this issue, IoT aware Polynomial Regressive Ensemble AI (IoT-PREAI) Model is introduced. IoT-PREAI technique is to perform several processes. In data harvesting process, the numbers of data samples are collected from the Crop Yield Prediction Dataset. In order to increase the size of input data samples, Adversarial scaling model is employed in IoT-PREAI technique. Data scrubbing is performed to clean the dataset by handling missing data using polynomial function and removing outliers. After that, feature selection is conducted using censored regression method by Roger and Tanimoto Distributed Feature Engineering. Following this, crop yield prediction is performed using the selected features through Adaptive Gradient weight preserved boosting technique. This method processes the data samples and generates the final crop yield prediction results while reducing the error using Nesterov Accelerated Gradient. Experimental evaluation considers several factors. The quantitatively analyzed results indicate that the proposed IoT-PREAI technique achieves higher crop yield prediction accuracy with minimal computation time, RMSE compared to conventional ensemble techniques.

Keywords: Crop yield prediction, Adversarial scaling model, Polynomial regression approach, Censored regression, Roger and tanimoto distributed feature engineering, Adaptive gradient weight preserved boosting, Nesterov accelerated gradient approach.

[1]Ph.D. Research Scholar (Full Time), PG and Research Department of Computer Science, Government Arts College (Autonomous) Karur-05, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

[2]Associate Professor and Research Supervisor, PG and Research Department of Computer Science, Government Arts College (Autonomous) Karur-05, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

***Corresponding Author:** P. Gayathri, Ph.D. Research Scholar (Full Time), PG and Research Department of Computer Science, Government Arts College (Autonomous) Karur-05, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India, E-Mail: msgayathriphd@gmail.com

How to cite this article: Gayathri, P., Jayanthi, C. (2026). IoT Aware Polynomial Regressive Ensemble Artificial Intelligence Model for Crop Yield Prediction in Cloud Computing Environment. The Scientific Temper, **17**(4):6166-6183.

Doi: 10.58414/SCIENTIFICTEMPER.2026.17.4.27

Source of support: Nil

Conflict of interest: None.

Introduction

The precision agriculture system support and helps the farmers in automating and updating their operations to enhance productivity in intelligent farming systems. Crop yield is a primary metric in precision agriculture that signifies the amount of yield per unit area of cultivated land. It provides a direct indicator of agricultural productivity and is necessary to estimate the effectiveness of farming practices, resource utilization. Recent advances in Artificial intelligence (AI) based data analytics and machine learning has released new ways for automatically predicting crop yields without human intervention.

An Ensemble regression using Extra Tree Regressor (ER-ETR) model was introduced with the intent of improving the crop yield prediction (Sudhamathi & Perumal, 2024). Although the model effectively minimizes the error rate, it encounters significant challenges in terms of time consumption during crop yield prediction. The AdaBoost algorithm combined with the Gray Level Co-occurrence Matrix (AdaBoost GLCM) was developed to enhance the

Received: 08/04/2026 **Accepted:** 22/04/2026 **Published:** 25/04/2026

accuracy of crop yield prediction through effective feature selection (Nagesh et al. 2024). However, the approach achieved the highest level of accuracy.

Feature selection is significant task in Internet-of-Things (IoT)–based crop-yield prediction. Study was conducted on five feature-selection families namely filter, wrapper, embedded, bio-inspired and deep learning (Parthasarathy & Manikandasaran, 2026) with higher accuracy. But, it failed to select relevant features with lesser time. To address this issue, Whale Optimization Algorithm (WOA) was designed for predicting the crop yield (Menaha & Lavanya, 2024). However, the accurate prediction in the shortest possible time is demanding issue. Stochastic kernelized discriminant extreme learning machine classifier (SKDELMC) was investigated for crop yield forecast with vast soil and weather big data (Anita & Shakila, 2024). But, the accuracy was not improved. Study was focused on drones and important attention in agriculture (Gomathi et al. 2023). Different applications of modern agriculture, such as resource optimization, crop monitoring, disease detection, and yield prediction were employed.

A hybrid deep learning technique was developed for cloud-based smart agriculture system (Sharma & Rathore, 2024). However, the data collection and management systems were not well established for accurate crop yield prediction. Crop Yield Prediction Algorithm (CYPA) was introduced to support precision agriculture and enhance the effectiveness and accuracy of crop yield forecasting (Talaat, 2023). However, the algorithm meets considerable difficulties when dealing with high-dimensional datasets. An IoT-based solution for smart farming was developed to accurately predict crop productivity (Mathi et al. 2023). However, it failed to achieve the better accuracy and accurate decision-making. A large-scale crop yield prediction method was introduced by extracting the temporal features of crop growth (Xiang et al. 2025). However, the error rate in crop yield prediction remained high.

Different machine learning models were developed with the aim of ensuring robust and accurate predictive capabilities (Razavi et al. 2024). However, it failed to explore advanced architectures to enhance prediction accuracy and model generalization. Machine learning approaches were developed by selecting the relevant features to optimize prediction accuracy and resource management (Vhatkar et al. 2024). However, the integration and AI-enabled systems was not employed to further enhance crop prediction accuracy. A weight-tuned deep convolutional neural network was developed with the intent of predicting the high crop yield (Subramaniam & Marimuthu, 2024). However, the precision agriculture technologies were not employed to provide accurate crop yield prediction and decision-making support. A transfer-learning-based framework was developed for improving yields prediction (Li et al. 2024). However, the model increases the accuracy but the time complexity was not minimized. Machine Learning model was developed with the aim of improving the wheat yield prediction (Ashfaq et al. 2024). However, it failed to consider the impacts of soil features to achieve accurate yield prediction. In order to identify the impact of the weather features, a novel LASSO regression was developed for accurate district-level yield prediction (Heilemann et al. 2024). However, it failed to enhance the accuracy of yield loss projections related to extreme weather events. An explainable AI (XAI)-based smart agriculture system was designed] for precision farming (Martin et al. 2024). However, explainable models were less accurate than complex models. Deep learning models were developed for small-field crop yield prediction (Stiller et al. 2024). However, the fine-tuning process remained unaddressed. A Random Forest regression model was introduced to enhance the crop yield prediction by reducing the mean absolute error (Panigrahi et al. 2023). However, many Artificial Intelligence (AI) techniques were not employed on large data samples to predict the crop yield.

Contributions

The most significant contributions of IoT-PREAI are listed below.

- To enhance crop yield prediction agriculture domain in cloud computing, a novel method called the IoT-PREAI has been developed by incorporating different processes namely data augmentation, data scrubbing, feature engineering and prediction.
- To reduce the crop yield prediction time, the IoT-PREAI employs data scrubbing and feature engineering process. The data scrubbing includes polynomial regressive based missing data handling and outlier or noisy data removal using Davies–Gather Outlier method. This process ensuring that only high-quality inputs are used. Feature engineering then extracts and selects the most significant attributes using censored regression with Roger and Tanimoto similarity, thereby enhancing the efficiency and accuracy of the predictive model.
- To enhance the accuracy of crop yield prediction, Ensemble AI technique is employed in IoT-PREAI by applying Adaptive Gradient weight preserving boosting technique. The boosting ensemble method increases the accuracy by minimizing the error in crop yield prediction through integrating the Nesterov Accelerated Gradient method. This in turn reduces the incorrect prediction of crop yield, thereby minimizing the RMSE.
- Finally, a comprehensive assessment is performed using various performance metrics to show the improved effectiveness of the IoT-PREAI technique in crop yield prediction when compared to conventional ensemble learning techniques.

Organizations

The rest of the paper is structured as follows: Section 2 discusses related works. Section 3 introduces the proposed of IoT-PREAI technique and its main components. Section 4 describes the experimental settings and the dataset used. Section 5 reports and evaluates the results of various methods based on different performance metrics. Lastly, Section 6 offers the overall conclusion of the paper.

Related Works

Stacking ensemble learning model was designed for predicting crop yields within Indian agriculture sector (Renju & Brunda, 2024). However, the performance of precision and recall analysis remained unaddressed. A Machine Learning (ML) models were developed to predict future crop production and yields with minimal RMSE (Nikhil et al. 2024). However, the application of IoT aware crop growth models was not incorporated to improve the prediction accuracy of the models. Random forest machine learning algorithm was designed with the aim of predicting agriculture yields and minimizing the model error (Sharma et al. 2023). However, this algorithm failed to enhance accuracy with minimal complexity compared to more sophisticated AI techniques. In order to improve the crop yield prediction, machine learning and deep learning techniques were introduced (Jhajharia et al. 2023). However, it failed to forecast the factors that influence crop yield.

An improved optimizer function (IOF) integrated with long short-term memory (LSTM) model was developed (Bhimavarapu et al. 2023). However, it failed to focus on applying the advanced tools for upgrading the performance of the prediction model. A novel deep neural networks model was developed to predict crop yields in based on climate change, fertilizer use, and crop area (Demirhan, 2025). However, it failed to predict crop yields in by considering the socioeconomic factors for crop yields prediction. The machine learning (ML) method was designed to predict yield by considering climatic variation, soil diversity, and fertilizer applications (Mitra et al. 2024). However, it failed to develop AI-based crop models and reduce the gap between advanced technologies and the agricultural domain.

Extreme gradient Boost model was developed with the intent of forecast the crop yield (Gharakhanlou & Perez, 2024). However, it failed to improve model accuracy and applicability by considering soil data collection. A random forest algorithm was designed to accurately predicting the soybean yield (Santos et al. 2024). However, the model has high computational complexity. Various advanced modeling techniques were designed to forecast wheat-yield prediction based on weather conditions and yield data. However, it did not incorporate an advanced AI model to enhance yield predictions (Iqbal et al. 2024). Hybrid deep learning-based models were developed for crop yield prediction with minimal time consumption (Oikonomidis et al. 2022). However, the fine-tuning process was major issue. A hybrid feature selection and optimized machine learning model was developed for accurate crop yield prediction (Abdel-Salam et al. 2024). However, it failed to include additional features to improve the accuracy of the prediction model.

Multimodal ensemble (MME) method was introduced using a particle filtering (PF) algorithm to increase the performance of season crop yield prediction (Zare et al. 2024). However, it failed to examine the effects of regional model calibration and model weighting scheme on the performance of data analysis. An integration of Convolutional Neural Network and Recurrent Neural Network (CNN-RNN) were developed for cocoa yield prediction using climatic dataset (Olofintuyi et al. 2023). However, it failed to improve the accuracy by considering more advanced network architecture. A Random Forest Extreme Gradient model was designed with the intent of predicting cotton yield based on recorded weather data and also efficiently handling large-scale datasets (Haider et al. 2024). However, it did not account for the impact of vital meteorological factors such as soil temperature and humidity to further improve the accuracy of the predictions.

Materials and Methods

Agriculture plays a vital role in sustaining the national economy and meeting the food demands in rising population. However, the increasing impact of climate change poses significant challenges to maintain a consistent and secure food supply. To solve these issues, modern agricultural practices are increasingly by incorporating scientific and technological solutions aimed to balance food production or yield. The unpredictability of weather conditions has made it more difficult for farmers to adopt sustainable and flexible farming methods. As a result, the advanced technologies and data-driven approaches have been widely applied for forecasting crop yields accurately. Reliable estimation of agricultural output is essential for early detection of risks to food security. This paper introduces a novel approach called IoT based Polynomial Regressive Ensemble AI (IoT-PREAI) for crop yield prediction in cloud computing using an ensemble learning model. The detailed description of the complete framework of IoT-PREAI is described in Figure 1.

Figure 1 above depicts the architecture diagram of the proposed IoT-PREAI technique for accurate prediction of crop yield or production. The model incorporates IoT sensors and devices installed on the agricultural field to collect the soil parameters, weather conditions etc. and it stored into the cloud server for further processing. After that, the proposed IoT-PREAI technique includes four essential processes namely data harvesting, data scrubbing,

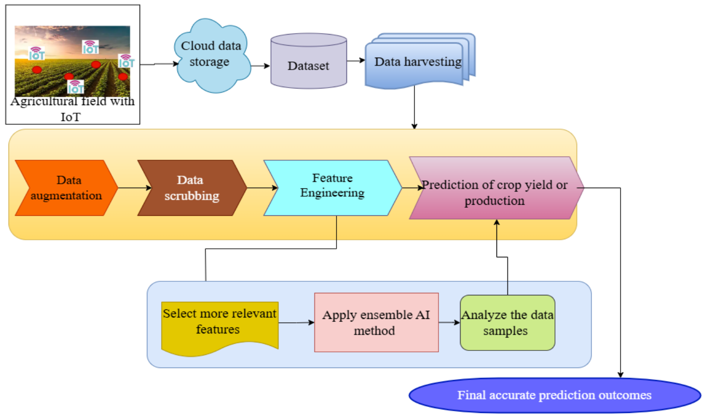

Figure 1: Architecture of proposed IoT-PREAI technique

feature engineering, classification and hyperparameter tuning. At first, numbers of data samples are gathered by using Adversarial scaling model at data harvesting process from the Crop Yield Prediction Dataset. To execute data scrubbing, polynomial regression and Davies–Gather Outlier Identification are applied for deal with missing data and eliminating outliers. Then, the feature engineering is carried out for selecting relevant features with Roger and Tanimoto Distributed Feature Engineering. After, the classification is performed by Adaptive Gradient weight preserved boosting technique for prediction of crop yield. Finally, hyperparameter tuning is performed with Nesterov Accelerated Gradient approach to reduce error for obtaining final crop yield prediction results. These five fundamental processes are contributed into the IoT-PREAI technique to further enhance the accuracy of crop yield prediction in agriculture sector with minimal time consumption. Therefore, the different process of proposed IoT-PREAI technique is explained briefly in the following subsections.

Data Harvesting

In the model, data harvesting refers to the process of collecting the data to assist in the early prediction of Smart Farming Sensor Data for Yield Prediction dataset https://www.kaggle.com/datasets/atharvasoundankar/smart-farming sensor-data-for-yield-prediction. This dataset represents simulated smart farming activities using IoT sensor readings. It includes key environmental and operational factors influencing crop yields across 500 farms situated in diverse regions such as India, the United States, and Africa. By using these datasets, effectively acquire high-quality data for developing and validating predictive models to perform the yield prediction assessment. The dataset consists of 22 attributes and 500 instances for accurate prediction of crop yield.

Let us consider the dataset 'DS' and data samples 'D' as well as features $\{F_1, F_2, \dots, F_m\}$ are arranged in the form of matrix. Therefore, the input matrix is formulated as given below,

$$IM = \begin{bmatrix} F_1 & F_2 & \dots & F_m \\ D_{11} & D_{12} & \dots & D_{1n} \\ D_{21} & D_{22} & \dots & D_{2n} \\ \vdots & \vdots & \dots & \vdots \\ D_{m1} & D_{m2} & \dots & D_{mn} \end{bmatrix} \tag{1}$$

Where, IM denotes an input matrix where each column indicates a number of features $F = \{F_1, F_2, \dots, F_m\}$, each row represents a number of data samples or instances or records '$D = \{D_1, D_2, \dots, D_k\}$' respectively.

Before initiating the machine learning process, the proposed IoT-PREAI technique applies a data augmentation strategy to expand the size of dataset, particularly in cases where the original data is limited. This process utilizes Adversarial scaling model is applied to generate the new data samples that reflect the uniqueness of the existing dataset, thereby increasing both its volume and variability. The model comprises of generator and discriminator, to create realistic data instances by minimizing the loss function.

The adversarial scaling model considers the number of input data collected from the dataset is given to the generator part. This unit is used to transform an input data samples from the low-dimensional into high-dimensional space.

$$D_{new} = g\left(D_{input}\right) \tag{2}$$

Where 'D_{input} represents a data samples from input domain, D_{new} denotes the newly generated data in another domain through the generator 'g'. The data generation process involves applying scaling to modify the distribution of the dataset.

$$D_{new} = \sum_{i=1}^{k} D_i * S \tag{3}$$

Where,

Table 1: Attribute Description

S: No.	*Attributes or column name*	*Description*
1	farm_id	Unique ID for each smart farm
2	Region	Geographic region (e.g., North India, South USA)
3	crop_type	Types of crop grown such as Wheat, Rice, Maize, Cotton, Soybean
4	soil_moisture_%	Soil moisture content in percentage
5	soil_pH	Soil pH level (5.5–7.5 typical range)
6	temperature_C	Average temperature during crop cycle (in °C)
7	rainfall_mm	Total rainfall received in mm
8	humidity_%	Average humidity level in percentage
9	sunlight_hours	Average sunlight hours received per day
10	irrigation_type	Type of irrigation: Drip, Sprinkler, Manual, None
11	fertilizer_type	Fertilizer used: Organic, Inorganic, Mixed
12	pesticide_usage_ml	Daily pesticide usage in milliliters
13	sowing_date	Date when crop was sown
14	harvest_date	Date when crop was harvested
15	total_days	Crop growth duration (harvest - sowing)
16	yield_kg_per_hectare	Target variable: Crop yield in kilograms per hectare
17	sensor_id	ID of the IoT sensor reporting the data
18	Timestamp	Random in-cycle timestamp when the data snapshot was recorded
19	Latitude	Farm location latitude (10.0 - 35.0 range)
20	Longitude	Farm location longitude (70.0 - 90.0 range)
21	NDVI_index	Normalized Difference Vegetation Index (0.3 - 0.9)
22	crop_disease_status	Crop disease status: None, Mild, Moderate, Severe

$$S = \frac{|D_i - \mu_D|}{v^2} \tag{4}$$

Where, D_{new} represents a newly generated data, D_i indicates an already available data in the dataset, 'S' indicates a scaling, μ_D denotes a mean of all data 'D', v indicates a standard deviation. In this way, new data samples are created based on the existing data. These generated samples are then passed to the discriminator unit which distinguishes the original and synthesized data with minimal loss.

Data Scrubbing

Data scrubbing, also known as data cleansing, involves preprocessing the input dataset but identifying inconsistencies in the dataset to improve its quality. This process includes handling missing values, removing outliers, and obtaining standardizing data formats. Effective data scrubbing helps to enhance the performance of accurate crop yield prediction. In data scrubbing, missing data refers to the absence of values in a dataset that affects the performance of crop yield prediction. In order to solve the issues, a novel weighted average non-negative factorization model is employed to replace the missing values with substituted values to allow for complete data analysis. Weighted average non-negative factorization is used to measure the multivariate analysis while minimizing its cost function, rather than treating these missing data as zeros.

The polynomial regression of the known value is computed within the particular features.

$$D_{ms} = \beta_0 + \beta_1 D_1 + \beta_2 D_2^{\,2} + \beta_3 D_3^{\,2} \ldots \beta_n D_n^{\,2} \tag{5}$$

Where, D_{ms} denotes a missing data value, $\beta_0, \beta_1, \beta_2 \ldots \beta_n$ indicates a polynomial coefficient, which are determined based on available data samples, $D_1, D_2, \ldots, D_n$ indicates known data samples in the dataset. The non-negative factorization minimizing its cost function as given below,

$$H = \arg min |D_i - D_{ms}|^2 \tag{6}$$

Where, H indicates outcomes of non-negative matrix factorization model, $|D_i - D_{ms}|$ denotes a cost function, $\arg min$ denotes an argument of minimal function. As a result, the missing values in the input data are estimated by minimizing the difference between the observed data and the data produced by polynomial function. This reconstruction aims to best approximate the original matrix, and the imputed missing values that minimize the reconstruction error while maintaining non-negativity.

Followed by, the outlier's detection is performedfor identifying data points that deviate significantly from the rest of the data within the dataset. These data are known as outliers. Davies–Gather Outlier Identification method is employed for detecting the multiple outliers within the dataset. First, the method finds the median value of the dataset, and then identifying data points that deviate significantly from this median.

Let us consider the number of data or samples '$D = \{D_1, D_2, \ldots, D_n\}$'. The median value is computed as follows,

$$M = med\{D_1, D_2, \ldots, D_n\} \tag{7}$$

The compute the maximum absolute deviation 'MAD' using Manhattan distance as follows,

$$MAD = \sum_{i=1}^{n} |M - D_i| \tag{8}$$

The data points whose distance exceeds a threshold is said to be an outlier. Otherwise the data is said to be a normal. The detected outlier data are removed to obtain the cleaned dataset for further processing.

// Algorithm 1: data scrubbing
Input: **Dataset 'DS', features** $F_1, F_2, \ldots, F_m$ and augmented data samples $D_1, D_2, \ldots, D_n$ **Output:** Preprocessed dataset
Begin **Step 1: Collect n**umber of data samples $D_1, D_2, \ldots, D_n$ from the dataset **Step 2: For each** sample 'D' **Step 3** Perform missing data imputation using (5) **Step 4:** Compute median test using (7) **Step 5:** if ($MAD > T$) **then** **Step 6:** Data is said to be outlier **Step 7:** **else** **Step 8:** Data is said to be normal **Step 9: End if** **Step 10**: Remove the outlier data **Step 11: End for** **Step 12: Return (Preprocessed dataset)** **End**

Algorithm 1 outlines the detailed steps involved in data scrubbing aimed at enhancing the accuracy and reducing the computational time for predicting crop yield. The process begins by gathering multiple augmented data samples from the dataset. Next, missing values are addressed using a weighted average technique. Subsequently, outliers are detected using maximum absolute distance. If the distance exceeds a predetermined threshold, the data point is determined as an outlier, otherwise it is considered normal. Finally, the cleaned and preprocessed dataset is obtained for accurate crop yield prediction.

Feature Engineering

Following scrubbing step, the proposed IoT-PREAI technique performs the feature engineering refers to the process of selecting the essential features to minimize the dataset's dimensionality. This model employs a censored regression method technique to identify and retain important features while discarding irrelevant ones. Censored regression is a machine learning method by analyzing the features and retains the more significant features aims to decrease the overall prediction time by applying Roger and Tanimoto similarity index.

In regression analysis, there are two outcomes are observed such as left-censoring and right-censoring. Left-censoring occurs when it is known that certain observed feature values fall below a predefined threshold. Conversely, right-censoring occurs when outcomes exceed a specified threshold. These censoring methods are particularly useful in modeling situations where part of features is only observed. In the feature selection process, censored regression models are often applied to measure the relationship between the independent variables (i.e. target variable) and dependent variable (i.e. features)

Let us consider the N features $F_1, F_2, \ldots, F_m$ x1,…,xN in high-dimensional space collected from the dataset. The relationship between the target variable and features are measured as follows,

$$RT(F_j, T_k) = \frac{|F_j \cap T_k|}{n + |F_j \Delta T_k|} \tag{9}$$

Where $RT(F_j, T_k)$ denotes a Roger and Tanimoto similarity index, $F_j \cap T_k$ denotes an exact match between the feature and target vectors, $F_j \Delta T_k$ indicates a difference between the features and target variable. The similarity index output provides output results from 0 to 1$[0 \le RT(F_j, T_k) \le 1]$. Based on the similarity outcomes, the regression method distinguishes the features into left and right censoring.

Figure 2 demonstrates censored regression that manages two observations of data samples such as left-censoring and right censoring in datasets. In left-censoring, Roger and Tanimoto similarity index values below a certain threshold are recorded. In contrast, right-censoring occurs when similarity index exceeds a specified threshold. Therefore, the features in the right censoring are more significant for accurate crop yield prediction than the left censored.

$$Q = \begin{cases} RT(F_j, T_k) > T, & features\ in\ RC \\ RT(F_j, T_k) < T, & features\ in\ LC \end{cases} \tag{10}$$

Where, Q denotes a regression outcome, T indicates a threshold, RC and LC denotes a right and left censoring points respectively. As a result, the features with right censoring points are selected for crop yield prediction and others are removed from the dataset. The algorithm for feature engineering is given below,

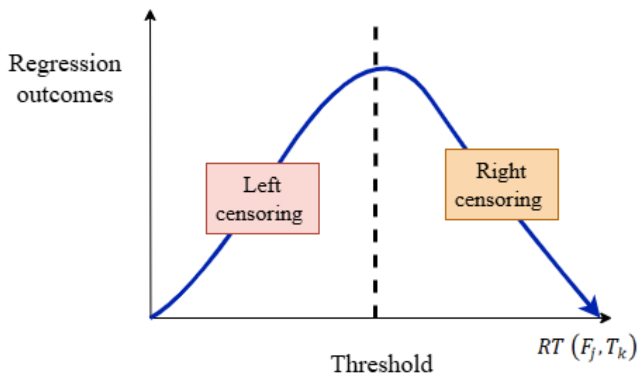

Figure 2: Censored regression analysis

Algorithm 2: Feature engineering
Input: **Dataset** ' DS ', **features** $F_1, F_2, \ldots, F_m$ and data samples $D_1, D_2, \ldots, D_n$
Output: select relevant features
Begin **Step 1:** Collect the scrubbing dataset as input **Step 2:** **For** each feature ' F_j ' in dataset ' DS ' **Step 3:** Measure the similarity using (9) **Step 4:** Apply regression to analyze similarity with threshold using (10) **Step 5.** $if(RT(F_j, T_k) > T)$ **then** **Step 6:** Q **returns** relevant features **Step 7:** **else** **Step 8:** Q **returns** irrelevant features **Step 9:** **End if** **Step 10:** Select the relevant features subsets and remove others **Step 11:** **End for** **End**

Algorithm 2 presents a feature engineering method to identify the most relevant features for crop yield prediction while reducing time consumption. The process starts with a scrubbing dataset consisting of various input features and data samples. It then evaluates the similarity among these features and target value. By applying a predefined threshold, the regression algorithm distinguishes relevant features subsets and irrelevant features subsets. Finally, only the most relevant features are selected, enhancing crop yield prediction accuracy and ensuring efficient processing time.

Ensemble AI Technique Based Crop Yield Prediction

The final process of IoT-PREAI technique is to perform the crop yield prediction with the set of the relevant features selected from the dataset. The proposed IoT-PREAI technique utilizes the Ensemble AI techniques for accurate crop yield prediction by employing Adaptive Gradient weight preserving boosting technique (AGWPBoost). It aims to enhance the generalization ability of the model by preserving the weight distribution during training phase. Contrary to traditional boosting methods, AGWPBoost accurately adjusts sample weights by applying an adaptive Nesterov Accelerated Gradient method. This approach helps in reducing overfitting, especially when working with noisy and unlabeled datasets, and ensures more reliable crop yield predictions in agriculture domain by minimizing the root mean square error.

AGWPBoost is an ensemble artificial intelligence approach that improves the accuracy and robustness of predictive models by integrating multiple weak learners to form more effective strong prediction outcomes. The main aim is to iteratively reduce the errors made by weak learners and increases the prediction performance. In contrast, a strong learner demonstrates high accuracy and reliable prediction outcomes by effectively minimizing classification errors.

Figure 3 illustrates the process of the Ensemble AI techniques. This proposed boosting ensemble technique first constructs 'b' weak learners, which are base classifiers using a artificial neural network (ANN). The algorithm uses the selected significant features from the training dataset $\{D_i, Y\}$ as input for the weak learners. In this training set, D_i represents the input training data, and Y represents the prediction output for the ensemble classification methods.

Figure 4 illustrates the schematic structure of an artificial neural network (ANN). It consists of three major layers such as an input layer, hidden (intermediate) layers, and an output layer arranged in a hierarchical manner. Each neuron in the input layer is assigned a distinct data as input. These hidden layers are designed to carry out computational operations, including analyzes the selected features and the data. The outcomes of these computations are forwarded to the output layer, which generates the final crop yield predictions for particular crop in the farm. In this layer, each neuron corresponds to a particular classification category, representing a distinct activity type.

The ANN model consists of training set { D , Y } where D denotes an input data matrix collected from the dataset with the selected features and a target variable output ' Y '. The input is given to the neurons of input layer structure. The neurons in the input layer did not perform any computational process.

$$X = \sum_{i=1}^{n} D_i * W_{ih} + B \tag{11}$$

Where, X represents a weighted sum output, D_i indicates an input training data, W_{ih} represents a weight between neuron in input layer and hidden layer, and B represents the bias. Then the input training data is transferred into hidden layers where the data analysis is carried out by analyzing the training and testing data samples using Soergel coefficient. The Soergel coefficient is applied for

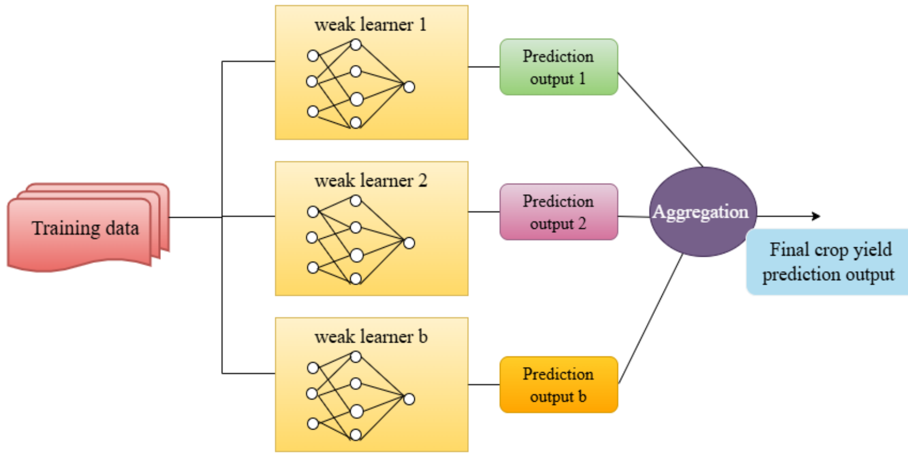

Figure 3: Structure of Ensemble AI techniques

measuring the similarity analysis between the features in the given dataset.

$$SC = \frac{\sum D_i D_{ts}}{\left(\sum D_i^2 + \sum D_{ts}^2 - \sum D_i D_{ts}\right)^2} \quad (12)$$

Where, SC denotes Soergel similarity coefficient, D_i denotes a training data and D_{ts} indicates a testing data. The coefficient (TC) provides the output ranges between 0 and 1. The softmax activation function provides the final prediction results at the output layer.

$$Z = \sigma\left[h_t * W_{ho}\right] \quad (13)$$

Where, Z indicates a predicted crop yield output of the weak learner, σ indicates a softmax activation function, which provides an outcome based on the similarity coefficient 'SC'. In this way, the weak learner provides the crop yield prediction outcomes.

The weak learner has some training errors and it hard to obtain accurate prediction results. Therefore, ensemble AI technique integrates all weak learners' results to obtain the strong prediction outcomes.

$$Y = \sum_{l=1}^{b} Z_l \quad (14)$$

Where, Y denotes the final prediction output of ensemble AI technique, Z_i represents a prediction output of weak learner. Iteratively, the ensemble AI technique initializes the weights to each weak learner.

$$Y = \sum_{l=1}^{b} Z_l * \vartheta \quad (15)$$

Where ϑ indicates the weights assigned to weak learner. The weights are initially assigned as random integers. Once these weights are assigned, the training error for each weak learner is evaluated to enhance the overall crop yield prediction accuracy. The error rate is determined by computing the squared difference between the actual and the predicted output of the weak learner. The error is calculated using the following expression:

$$ER = \left[Z_{l\,Act} - Z_{iobs}\right]^2 \quad (16)$$

Where, ER indicates an error, $Z_{l\,Act}$ represents actual prediction results, Z_{iobs} indicates observed prediction results of the weak learner. Followed by, the initial weights get updated based on the error. In order to minimize the error, the adaptive Nesterov Accelerated Gradient method is applied.

$$\vartheta_{new} = \vartheta_{old} - \eta\varphi_t \quad (17)$$

Where,

$$\varphi_t = \tau\varphi_{t-1} + (1-\tau)\frac{\partial ER}{\partial \vartheta_{old}} \quad (18)$$

Where, ϑ_{new} represents an updated weight, ϑ_{old} denotes a previous weight of weak leaner, η denotes a learning rate $(\eta < 1)$, '$\frac{\partial ER}{\partial \vartheta_{old}}$' represents a partial derivative of the error rate with respect to current weight 'ϑ_{old}', τ default value 0.9, φ is initialized 0. After updating the weights, an optimal weight is determined to minimize the error in the crop

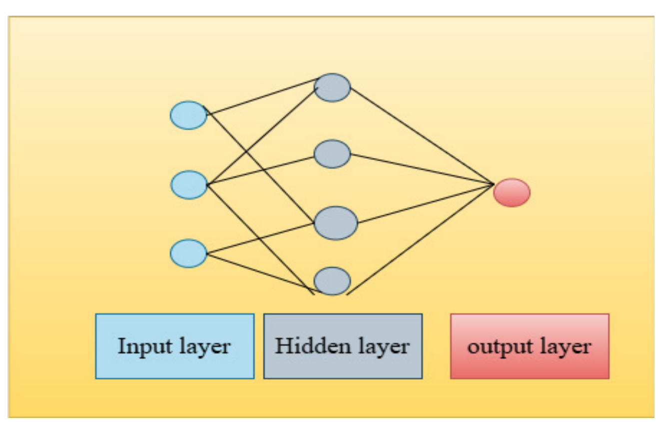

Figure 4: Process of ANN

yield prediction process. In this way, accurate crop yield prediction results are obtained. The algorithm of Ensemble AI techniques for accurate crop yield prediction is given below.

/ Algorithm 3: Ensemble AI techniques for accurate crop yield prediction

Input: **Dataset '** DS **', selected features** $F_1, F_2, \ldots, F_r$ and data samples $D_1, D_2, \ldots, D_n$
Output: Increase prediction accuracy

Begin

Step 1: For each data samples **'** D_i **' do**
Step 2: Construct 'b' set of weak learners
Step 3: Training data samples is given as input layer
Step 4: Measure the weighted sum using (11)
Step 5: **For each** training and testing data samples---**hidden layer**
Step 6: **Measure** Soergel coefficient using (12)
Step 7: **End for**
Step 8: Apply softmax activation function using (13) and obtain prediction results
Step 9: Combine all the weak learner's results

$$Y = \sum_{l=1}^{b} Z_l$$

Step 10: **For each** Z_l
Step 11: Assigns the similar weight 'ϑ'
Step 12: Calculate the error ER
Step 13: Update the weight based on the error using (17) (18)
Step 14: Find a prediction results with minimum error
Step 15: Return (crop yield prediction)
Step 16: End for
End

Algorithm 3 outlines the procedure of ensemble AI-based prediction aimed at improving the accuracy of crop yield while reducing error. The method begins by constructing 'b' weak learners as ANN for each training data. The ANN receives the set of training data in the input layer. The Soergel coefficient is applied to analyze the training data samples and the testing data samples. Based on analysis results, crop yield prediction is obtained at output layer with softmax activation function. After that, weak learner's results are combined. For each weak learner, the training error is measured to evaluate its contribution. Then, an adaptive gradient descent is applied to update weight and find the accurate ensemble's prediction output with minimal error. This adaptive method aims to minimize the overall training error, thereby enhancing prediction accuracy and reducing the error rate. This adaptive approach is designed to reduce the total training error, which in turn improves accuracy of crop yield forecasting.

Experimental Settup

Experimental assessment of the proposed IoT-PREAI technique and existing methods namely ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024) are implemented using python language. In order to conduct the experiment, Smart Farming Sensor Data for Yield Prediction dataset is consider and it taken from the https://www.kaggle.com/datasets/atharvasoundankar/smart-farming sensor-data-for-yield-prediction. For the experimental analysis, the Yield Prediction dataset is utilized to predict the particular crop yield in terms of kilogram per hectare. This dataset includes CSV files across 500 farms situated in diverse regions such as India, the United States, and Africa. The dataset includes 22 attributes and 500 instances for accurate prediction of crop yield. Feature descriptions are listed in table 1.

Implementation results analysis using IoT-PREAI technique

The implementation of the IoT-PREAI technique is thoroughly analyzed to assess its effectiveness in crop yield prediction. The analysis focused on several key processes including data harvesting, data augmentation, data scrubbing, feature engineering and prediction using Smart Farming Sensor Data for Yield Prediction dataset.

Step 1: Data harvesting

Data harvesting is the process of systematic collection of relevant agricultural data from various sources through an IoT-enabled sensors and smart devices deployed across agricultural fields. The harvested data is transmitted to a centralized cloud server and it stored in database. This ensures the input data is both accurate and consistent for further analysis. During the harvesting step, relevant agricultural data are collected from the dataset Smart Farming Sensor Data for Yield Prediction dataset. The sample data harvesting is shown in Figure 5.

Step 2: Data augmentation

Data augmentation in IoT-PREAI technique refers to the process of artificially increasing the size and diversity of the training dataset to improve the model's robustness. The given dataset comprises of 22 attributes and 500 instances and it may not be sufficient to train complex machine learning models effectively. To address these challenges, Adversarial scaling model is applied in IoT-PREAI technique to enrich the dataset before model training. After data augmentation, the dataset is expanded to 1,000 instances, with a total size of 234 KB. The data augmentation outcome is illustrated in Figure 6.

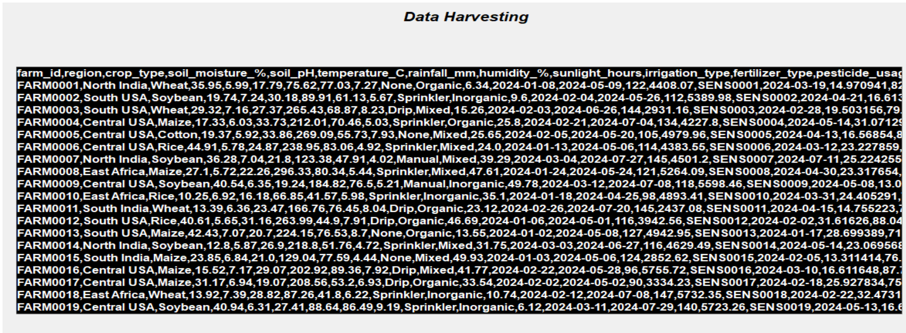

Figure 5: Data harvesting

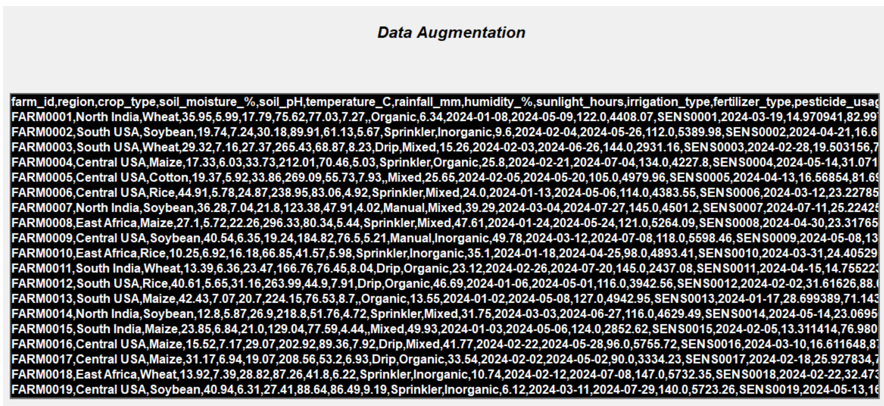

Figure 6: Data augmentation

Step 3: Data scrubbing

Data scrubbing is a preprocessing step in the IoT-PREAI framework for crop yield prediction. It involves handing missing data and removing outlier from the dataset collected via IoT sensors and other sources. The size of the data before data scrubbing was 234 KB, and after data scrubbing, it was reduced to 220 KB. Missing data imputation and outlier data removal are demonstrated in Figure 8 and 9.

Step 4: Feature engineering

Feature engineering in the IoT-PREAI involves selecting the meaningful input features that improve the performance of crop yield prediction models. In IoT-PREAI, the censored regression is employed for selecting 13 significant features from total 22 attributes in the dataset by means of Roger and Tanimoto similarity coefficient.

Figure 9 illustrates the Roger and Tanimoto similarity analysis which evaluates the degree of similarity between input features and crop yield outcomes. This analysis helps identify the 13 relevant attributes such as soil moisture, soil Ph, temperature, rainfall, humidity, sunlight hour, pesticide usage, total days, yield kg per hectare, latitude, longitude, NDVI index, and crop disease status that contribute significantly to yield prediction by measuring binary feature correlations.

Step 5: Crop yield prediction

Finally, the prediction of crop yield is carried out using an Ensemble AI by applying Adaptive Gradient weight preserving boosting technique. The method is trained using 13 selected relevant features to maximize crop yield prediction in in terms of kilogram per hectare and reduce error rate. The predicted crop yield for each crop results are shown in Figure 10.

Evaluation Metrics

This section presents a detailed explanation of several evaluation metrics used to assess crop yield prediction performance. These include crop yield prediction accuracy, precision, recall, F1-score, root mean square error (RMSE), and the crop yield prediction time, along with their respective mathematical formulations.

Crop yield prediction accuracy

It is calculated by dividing the number of correctly predicted crop yields by the total number of data instances. The mathematical expression for crop yield prediction accuracy is given below:

$$CPA = \left(\frac{TP + TN}{TP + TN + FP + FN} \right) * 100 \tag{19}$$

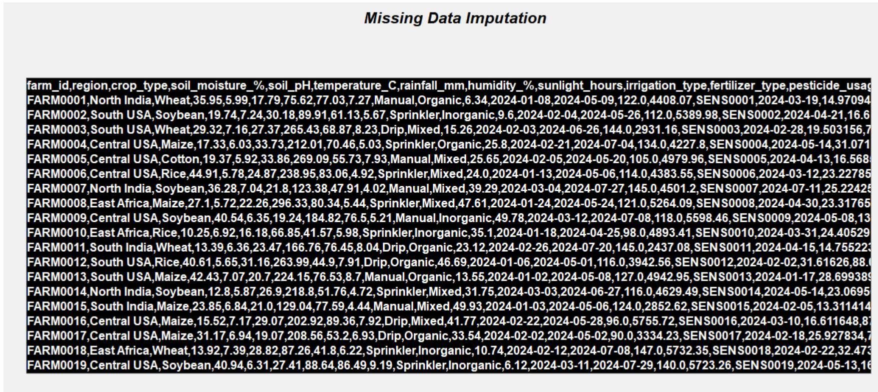

Figure 7: Missing data imputation

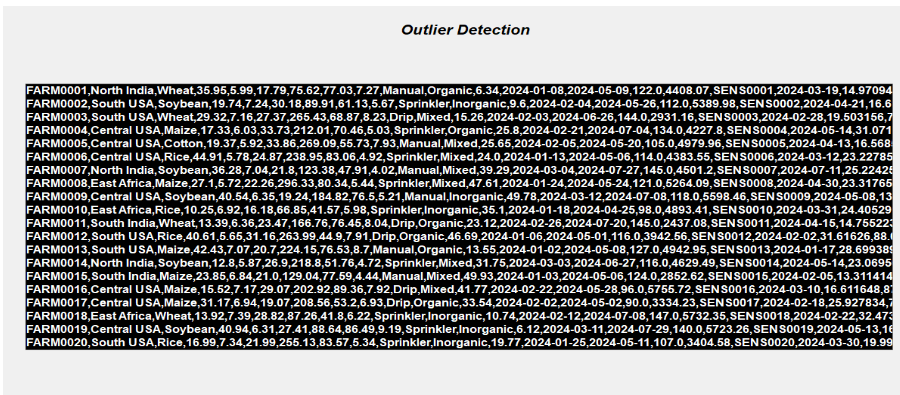

Figure 8: Outlier data removal

Where, CPA indicates a crop yield prediction accuracy, TP denotes the true positive, TN indicates the true negative, FP indicates the false positive, FN denotes the false negative. It is measured in percentage (%).

Precision

it is determined by calculating the ratio of true positive predictions to the total number of predicted positive cases, which includes both true positives as well as false positives. It reflects the accuracy of the model in correctly identifying the crop yield among the predicted positive outcomes. The formula for precision is expressed as follows,

$$\boldsymbol{Prec} = \left(\frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FP}}\right)*100 \tag{20}$$

Where, $\boldsymbol{Prec}$ denotes a precision, TP denotes the true positive, FP represents the false positive.

Recall

it also known as sensitivity refers to the model's capability to correctly identify all relevant instances within the dataset. It is calculated as the ratio of true positive predictions to the total of true positives and false negatives. The mathematical formula of recall is expressed as follows,

$$Recall = \left(\frac{TP}{TP+FN}\right)*100 \tag{21}$$

Where, TP indicates the true positive, FN represents the false negative.

F1 score

it also known as F-measure serves as the harmonic mean of precision as well as recall, offering an integrated measure that balances both metrics. It is mathematically determined using the following expression.

$$F1-Score = 2*\left(\frac{\boldsymbol{Prec}*Recall}{\boldsymbol{Prec}*Recall}\right) \tag{22}$$

Where, PRS denotes a precision, RCl indicates a recall.

Root Mean Square Error (RMSE)

it is used to evaluate the crop yield prediction accuracy of a model by measuring the differences between predicted and actual values. It is defined as the square root of the mean of the squared deviations between predicted outputs and true values across all data samples. The RMSE is calculated as follows,

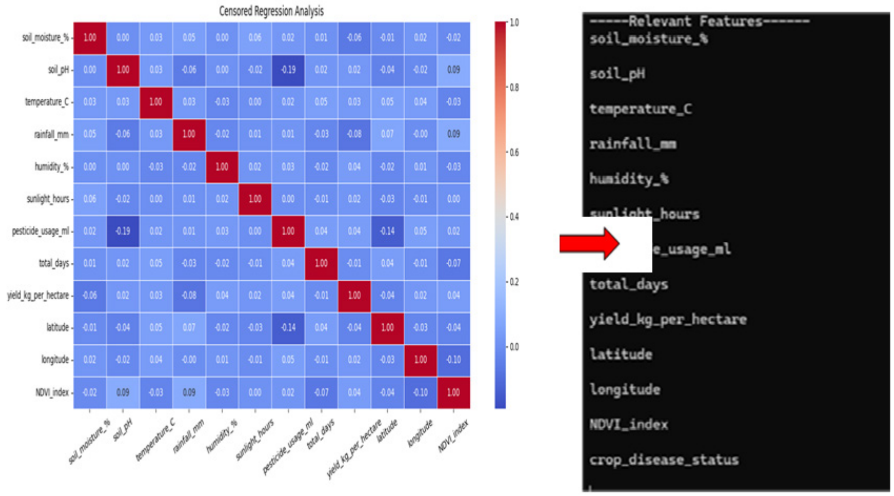

Figure 9: Similarity analyses of features and crop yield

$$RMSE = \left[\sqrt{\frac{\left(Y_{Actual} - Y_{predicted}\right)^2}{n}} \right] \tag{23}$$

Where, $RMSE$ indicates a root mean square error, Y_{Actual} indicates the actual crop yield prediction output, $Y_{predicted}$ denotes a predicted crop yield results, n indicates a number of data.

Crop yield prediction time

It refers to the duration required by the algorithm to complete the crop yield prediction process. The prediction time is represented as follows,

$$CPT = \sum_{i=1}^{n} D_i * time(CP) \tag{24}$$

Where, CPT indicates a crop yield prediction time based on the data 'D_i' and the actual time consumed in crop yield prediction denoted by '$time(CP)$. It is measured in terms of milliseconds (ms).

Performance Metric Analysis

In this section, performance of the IoT-PREAI technique and existing methods namely ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024) are evaluated with various metrics, including crop yield prediction accuracy, precision, recall, F1 score, root mean square error, crop yield prediction time with different number of data.

Table 2 illustrates the result analysis of crop yield prediction accuracy for proposed IoT-PREAI and existing ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024).

Figure 11 given above shows the accuracy of crop yield prediction in relation to the number of data samples collected by varying from 100 to 1000. As shown in two-dimensional graph, the x-axis indicates the number of data samples, while the y-axis represents the corresponding

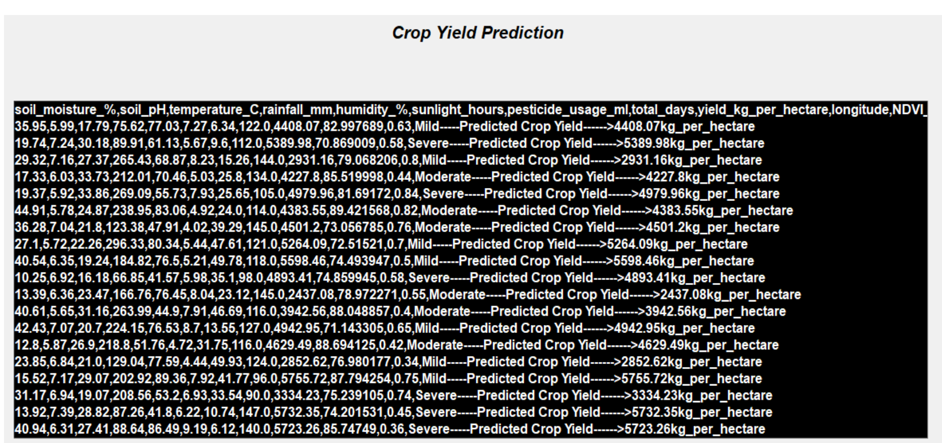

Figure 10: Crop yield prediction results

Table 2: Comparative analysis of crop yield prediction accuracy

Number of data samples	*Crop yield prediction accuracy (%)*		
	Proposed IoT-PREAI	*Existing ER-ETR*	*Existing AdaBoost GLCM*
100	94	90	88
200	94.88	90.05	87.56
300	94.23	90.74	88.05
400	94.89	91.23	88.45
500	95.36	91.36	88.95
600	95.03	90.78	87.47
700	94.87	91.46	88.63
800	95.02	90.78	87.45
900	95.78	91.85	88.69
1000	94.32	90.65	87.56

crop yield prediction accuracy. The above graphical analysis illustrates that the proposed IoT-PREAI technique constantly provide the improved performance than ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024). In the initial run with 100 data samples, the IoT-PREAI technique achieved an 94% of crop yield prediction accuracy, whereas accuracies of ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024) observed to be 90% and 88% respectively. In conclusion, the proposed IoT-PREAI technique improved crop yield prediction by approximately 4% and 8% when compared to ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024). The enhanced performance was achieved through the implementation of a precise crop yield prediction system called Ensemble AI techniques for accurate crop yield prediction by using Adaptive Gradient weight preserving boost technique. It helps to analyze the data samples by constructing the number of weak learners and predict the crop yield based on soil and environment characteristics using Soergel similarity coefficient. Based on the coefficient results, accurate crop yields are correctly predicted for the specific crop. This in turn increases the true positive and true negative. in order to minimize the false positive and false negative, adaptive Nesterov Accelerated Gradient method is employed to minimize the error in crop yield prediction. This ensures more reliable and accurate crop yield predictions in agriculture domain. Result analysis of precision for proposed IoT-PREAI and existing ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024) are demonstrated in Table 3.

Figure 12 depicts a performance analysis of precision versus number of data samples collected from the dataset. As shown in graph, horizontal axis indicates the number of data samples, while the vertical axis reflects the corresponding precision values in crop yield prediction. Among the three available methods, the IoT-PREAI technique frequently provides improved precision results compared to the other ensemble approaches. For example, when 100 data samples were considered, the IoT-PREAI technique achieved a precision of 95.45%. By applying the existing methods ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024), the performance of precision was observed to be 93.18% and 92.04% respectively. Overall, the averaged results indicate that the IoT-PREAI technique improves precision by approximately 4% compared to ER-ETR (Sudhamathi & Perumal 2024) and 2% compared to AdaBoost GLCM (Nagesh et al. 2024). This main aim is primarily due to the use of extreme learning, which enables the deployment of adaptive Nesterov Accelerated Gradient to guide the optimization process, and model weights are iteratively updated to refine performance. This approach effectively increases the true positive rate while minimizing the false positives, resulting in higher overall precision. Result analysis of recall for proposed IoT-PREAI and existing ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024) are demonstrated in Table 4.

Figure 13 portrays the performance comparison of recall versus the number of data samples collected from the dataset. The IoT-PREAI technique illustrates a noticeable

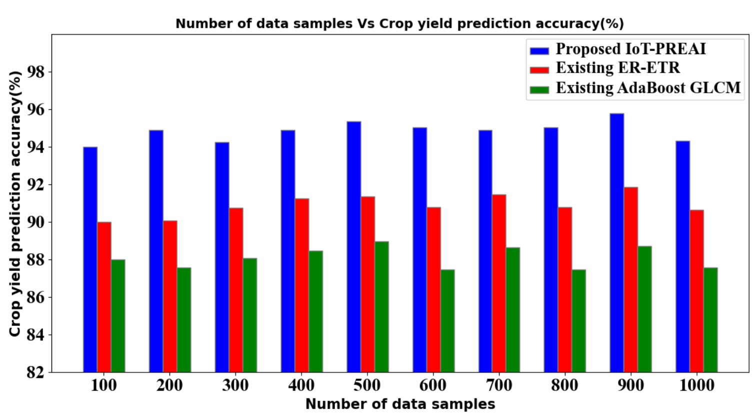

Figure 11: Result comparison graph of crop yield prediction accuracy

Table 3 Comparative analysis of precision

Number of data samples	Precision (%)		
	Proposed IoT-PREAI	*Existing ER-ETR*	*Existing AdaBoost GLCM*
100	95.45	93.18	92.04
200	95.36	93.56	92.05
300	94.66	93.05	91.56
400	95.87	93.45	91.78
500	94.78	92.89	90.85
600	95.96	93.45	91.62
700	95.86	93.75	91.74
800	95.56	93.06	91.69
900	94.74	93.56	91.62
1000	94.62	93.06	91.06

Table 4: Comparative analysis of recall

Number of data samples	Recall (%)		
	Proposed IoT-PREAI	*Existing ER-ETR*	*Existing AdaBoost GLCM*
100	97.67	95.34	94.18
200	97.36	95.12	94.11
300	97.56	95.36	93.65
400	97.05	95.12	93.78
500	96.89	94.74	93.74
600	96.78	94.52	93.26
700	97.25	94.63	92.74
800	97.63	95.45	93.63
900	97.23	95.63	93.45
1000	97.05	95.12	93.12

improvement in performance of recall when compared to the existing methods ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024). In the initial training process using 100 data samples, the performance of recall rate was observed by the IoT-PREAI technique was 97.67%. By applying ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024) was recorded 95.34% and 94.18%, respectively. Average of ten experimental results indicates the performance of recall using IoT-PREAI was improved by 4% and 2% when compared to existing ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024). This improvement is achieved owing to the integration of the Ensemble AI technique, which contributes to provide more accurate crop yield predication in terms of kilogram per hectare. Consequently, the model achieves a higher rate of true positives and effectively reduced the false negatives during the prediction process. Result analysis of F1 score for proposed IoT-PREAI and existing ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024) are demonstrated in Table 5.

Figure 14 illustrates the graphical analysis of F1 score comparison across different data samples, ranging from 100 to 1000 samples. The findings illustrate that the proposed IoT-PREAI consistently outperforms conventional ensemble learning approaches in terms of F1 score. Each method was evaluated over ten independent runs. Form the analysis, IoT-PREAI technique demonstrating superior performance in the evaluation of F1 score. The average value of ten runs indicates that the IoT-PREAI technique increased the F1 score by 2% than ER-ETR (Sudhamathi & Perumal 2024) and a 4% improvement over AdaBoost GLCM (Nagesh et al. 2024). This enhanced performance was achieved by effectively handling of both precision and recall for accurately predicting the crop yield in the network. Table 6 explains result analysis of F1 score for proposed IoT-PREAI and existing ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024).

Figure 15 displays a graphical illustration of the Root Mean Square Error (RMSE) in relation to varying data sample sizes, ranging from 100 to 1000. The comparison involves three methods namely the proposed IoT-PREAI technique, ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024). The horizontal axis indicates the number

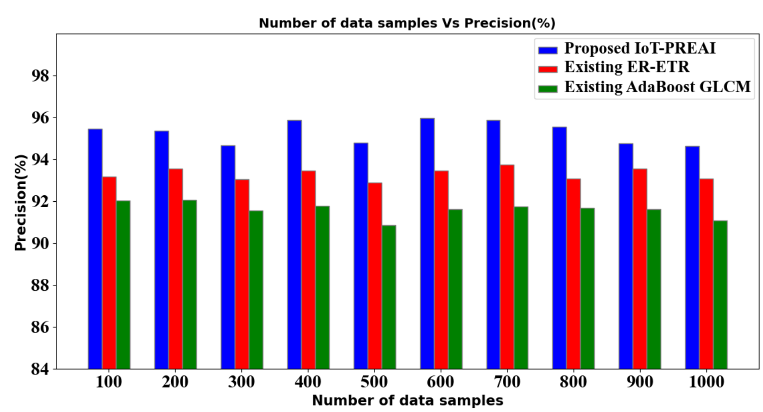

Figure 12: Result comparison graph of precision

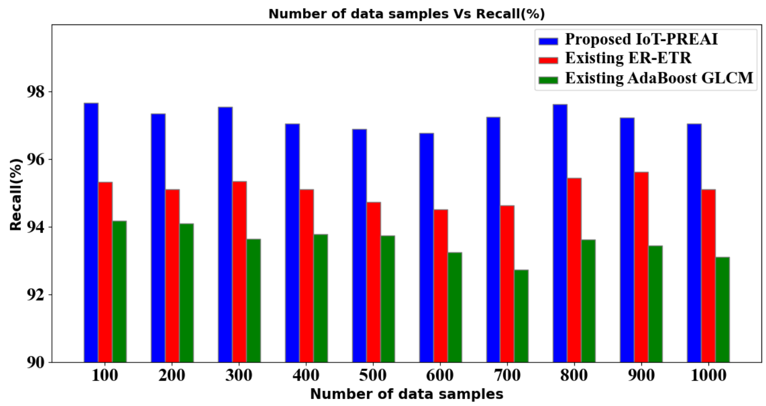

Figure 13: Result comparison graph of recall

of data, while the vertical-axis indicates the corresponding performance of RMSE values. The results illustrate that the IoT-PREAI technique considerably minimizes the RMSE values than the two baseline ensemble methods. For instance, with 100 data samples, the IoT-PREAI technique achieved an RMSE of 0.6, whereas method ER-ETR recorded 1 and method AdaBoost GLCM observed1.2. Similar trends were observed across different data samples. Overall, the IoT-PREAI technique reduced RMSE by approximately 44% compared to ER-ETR (Sudhamathi & Perumal 2024) and 57% compared to AdaBoost GLCM (Nagesh et al. 2024) in predicting crop yield. This improvement is achieved due

Table 5: Comparative analysis of F1 score

Number of data samples	*F1 score (%)*		
	Proposed IoT-PREAI	*Existing ER-ETR*	*Existing AdaBoost GLCM*
100	96.54	94.24	93.09
200	96.34	94.33	93.06
300	96.08	94.19	92.59
400	96.45	94.27	92.76
500	95.82	93.80	92.27
600	96.36	93.98	92.43
700	96.54	94.18	92.23
800	96.58	94.23	92.64
900	95.96	94.58	92.52
1000	95.81	94.07	92.07

Table 6: Comparative analysis of RMSE

Number of data samples	*RMSE*		
	Proposed IoT-PREAI	*Existing ER-ETR*	*Existing AdaBoost GLCM*
100	0.6	1	1.2
200	0.36	0.70	0.87
300	0.33	0.53	0.68
400	0.25	0.43	0.57
500	0.20	0.38	0.49
600	0.20	0.37	0.51
700	0.19	0.32	0.42
800	0.17	0.32	0.44
900	0.14	0.27	0.37
1000	0.17	0.29	0.39

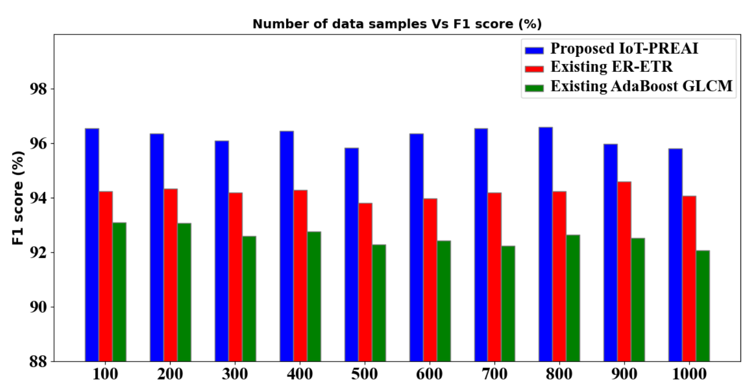

Figure 14: Result comparison graphs of F1 score

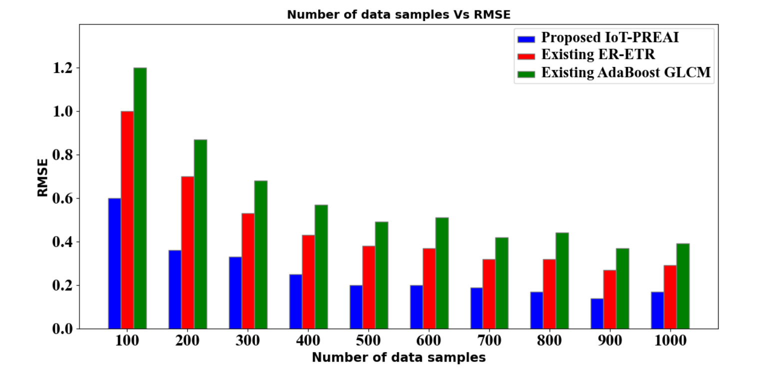

Figure 15: Result comparison graphs of RMSE

Table 7: Comparative analysis of crop yield prediction time

Number of data samples	*Crop yield prediction time (ms)*		
	Proposed IoT-PREAI	*Existing ER-ETR*	*Existing AdaBoost GLCM*
100	26	28	32
200	30.6	32.8	35.6
300	32.3	35.6	38.8
400	34.5	36.9	40.5
500	35.8	40.5	43.8
600	37.9	42.6	46.7
700	39.4	43.8	48.5
800	41.6	46.5	50.6
900	43.5	48.7	52.8
1000	46.8	50.4	54.6

to the application of the Nesterov Accelerated Gradient descent strategy, which refines the output of weak learners by adjusting weights, thereby reducing the overall prediction error and improving RMSE performance. Table 7 presents result analysis of crop yield prediction time for proposed IoT-PREAI and existing ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024).

Figure 16 presents a performance analysis of crop yield prediction time among the proposed IoT-PREAI technique, ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024). The graph plots crop yield prediction time against varying input data, ranging from 100 to 1000 samples. As the dataset size increases, all three methods exhibit a prediction time in increasing trend. In the case of 1000 input samples, IoT-PREAI technique achieved a crop yield prediction time of 26ms, while ER-ETR (Sudhamathi & Perumal 2024) and AdaBoost GLCM (Nagesh et al. 2024) recorded 28ms and 32ms, respectively. These results indicate that IoT-PREAI technique reduces crop yield prediction time by approximately 9% compared to ER-ETR (Sudhamathi & Perumal 2024) and 17% compared to AdaBoost GLCM (Nagesh et al. 2024). The efficiency is achieved by applying a Data scrubbing and Feature Engineering. In the scrubbing process, inaccurate and outlier data are removed from a dataset. The main aim is to ensure that the data is accurate, consistent, and usable for analysis or processing. During the Feature Engineering, the censored regression is applied to analyze the features with the target samples. This process effectively identifies and retains relevant features while discarding less significant ones, thereby enhancing the crop yield prediction process and reducing computational time.

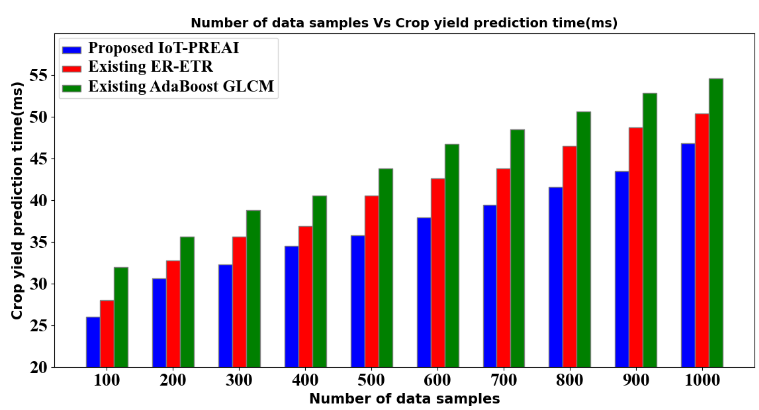

Figure 16: Result comparison graphs of crop yield prediction time

Conclusion

Crop yield prediction is a major concern to transform the agriculture industry for food security and sustainable development. This paper introduces a cloud-enabled smart agriculture model called IoT-PREAI technique with the intent to enhance crop yield prediction through the use of ensemble AI learning techniques. The main aim is to increase the accuracy of yield forecasting and decision-making for farmers. The IoT-PREAI technique utilizes data scrubbing and feature selection process in order to reduce the computation time of crop yield prediction in smart farming. The system also used Ensemble AI architecture that employs weak learners and to analyze multi-source data, such as weather conditions, soil moisture levels and providing final crop yield prediction results. A thorough experimental assessment is carried out using essential performance indicators such as crop yield prediction accuracy, precision, recall, F1 score, RMSE, and the crop yield prediction time across different data sample. The findings indicate that the IoT-PREAI technique consistently increases the better performance compared to conventional ensemble learning models, providing higher accuracy, improved precision and recall, better F1 scores, and reduced time as well as RMSE.

Acknowledgement

None.

Funding

The author declare that no funds, grants or other support were received during the preparation of this manuscript.

Availability of Data and Material

Smart Farming Sensor Data for Yield Prediction dataset is used and publically available within the article. It is taken from the https://www.kaggle.com/datasets/atharvasoundankar/smart-farming-sensor-data-for-yield-prediction.

References

Abdel-Salam, M., Kumar, N., & Mahajan, S. (2024). A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. *Neural Computing and Applications*, *36*(33), 20723-20750. https://doi.org/10.1007/s00521-024-10226-x

Anita, M., & Shakila, S (2024). Stochastic kernelized discriminant extreme learning machine classifier for big data predictive analytics. *The Scientific Temper, 15,* 395-404. Doi: 10.58414/SCIENTIFICTEMPER.2024.15.spl.46

Ashfaq, M., Khan, I., Alzahrani, A., Tariq, M. U., Khan, H., & Ghani, A. (2024). Accurate wheat yield prediction using machine learning and climate-NDVI data fusion. *IEEe Access*, *12*, 40947 – 40961. DOI: 10.1109/ACCESS.2024.3376735

Bhimavarapu, U., Battineni, G., & Chintalapudi, N. (2023). Improved optimization algorithm in LSTM to predict crop yield. *Computers*, *12*(1), 10, 1-19. https://doi.org/10.3390/computers12010010

Demirhan, H. (2025). A deep learning framework for prediction of crop yield in Australia under the impact of climate change. *Information Processing in Agriculture*, *12*(1), 125-138. https://doi.org/10.1016/j.inpa.2024.04.004

Gomathi, P., D Rose, D.D., Kumar, R.S., Priya, M.S., Dinesh, S. Ramarao, M. (2023). Computer vision for unmanned aerial vehicles in agriculture: applications, challenges, and opportunities. *The Scientific Temper, 14 (3),* 957-962 Doi: 10.58414/SCIENTIFICTEMPER.2023.14.3.61

Gharakhanlou, N. M., & Perez, L. (2024). From data to harvest: Leveraging ensemble machine learning for enhanced crop yield predictions across Canada amidst climate change. *Science of The Total Environment*, *951*, 1-15.https://doi.org/10.1016/j.scitotenv.2024.175764

Haider, S. T., Ge, W., Li, J., Rehman, S. U., Imran, A., Sharaf, M. A. F., & Haider, S. M. (2024). An ensemble machine learning framework for cotton crop yield prediction using weather parameters: a case study of Pakistan. *IEEE Access*, *12*, 124045-124061. DOI: 10.1109/ACCESS.2024.3454511

Heilemann, J., Klassert, C., Samaniego, L., Thober, S., Marx, A., Boeing, F., ... & Gawel, E. (2024). Projecting impacts of extreme weather events on crop yields using LASSO regression. *Weather and Climate Extremes*, *46*, 1-15.https://doi.org/10.1016/j.wace.2024.100738

Iqbal, N., Shahzad, M. U., Sherif, E. S. M., Tariq, M. U., Rashid, J., Le, T. V., & Ghani, A. (2024). Analysis of wheat-yield prediction using machine learning models under climate change scenarios. *Sustainability*, *16*(16), 1-26.https://doi.org/10.3390/su16166976

Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *218*, 406-417. https://doi.org/10.1016/j.procs.2023.01.023

Li, J., Zhang, D., Yang, F., Zhang, Q., Pan, S., Zhao, X., ... & Zhao, C. (2024). TrG2P: A transfer-learning-based tool integrating multi-trait data for accurate prediction of crop yield. *Plant Communications*, *5*(7), 1-12. https://doi.org/10.1016/j.xplc.2024.100975

Martin, R. J., Mittal, R., Malik, V., Jeribi, F., Siddiqui, S. T., Hossain, M. A., & Swapna, S. L. (2024). XAI-powered smart agriculture framework for enhancing food productivity and sustainability. *IEEE access*, *12*, 168412 – 168427. DOI: 10.1109/ACCESS.2024.3492973

Mathi, S., Akshaya, R., & Sreejith, K. (2023). An internet of things-based efficient solution for smart farming. *Procedia Computer Science*, *218*, 2806-2819. https://doi.org/10.1016/j.procs.2023.01.252

Menaha, M., & Lavanya, J. (2024). Crop yield prediction in diverse environmental conditions using ensemble learning. *The Scientific Temper, 15 (3),* 2740-2747. Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.43

Mitra, A., Beegum, S., Fleisher, D., Reddy, V. R., Sun, W., Ray, C., ... & Malakar, A. (2024). Cotton yield prediction: a machine learning approach with field and synthetic data. *IEEE Access*, *12*, 101273 – 101288. DOI: 10.1109/ACCESS.2024.3418139

Nagesh, O. S., Budaraju, R. R., Kulkarni, S. S., Vinay, M., Ajibade, S. S. M., Chopra, M., ... & Kaliyaperumal, K. (2024). Boosting enabled efficient machine learning technique for accurate prediction of crop yield towards precision agriculture. *Discover Sustainability*, *5*(1), 1-9. https://doi.org/10.1007/s43621-024-00254-x

Nikhil, U. V., Pandiyan, A. M., Raja, S. P., & Stamenkovic, Z. (2024).

Machine learning-based crop yield prediction in south india: performance analysis of various models. *Computers*, *13*(6), 1-27. https://doi.org/10.3390/computers13060137

Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied artificial intelligence*, *36*(1), 1-18. https://doi.org/10.1080/08839514.2022.2031823

Olofintuyi, S. S., Olajubu, E. A., & Olanike, D. (2023). An ensemble deep learning approach for predicting cocoa yield. *Heliyon*, *9*(4). 1-20.https://doi.org/10.1016/j.heliyon.2023.e15245

Panigrahi, B., Kathala, K. C. R., & Sujatha, M. (2023). A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. *Procedia Computer Science*, *218*, 2684-2693. https://doi.org/10.1016/j.procs.2023.01.241

Parthasarathy, V., & Manikandasaran S.S. (2026). Feature Selection Techniques for IOT Crop Yield Prediction Using Smart Farming Sensor Data. *The Scientific Temper, 17 (1),* 5477-5490. Doi: 10.58414/SCIENTIFICTEMPER.2026.17.1.12

Razavi, M. A., Nejadhashemi, A. P., Majidi, B., Razavi, H. S., Kpodo, J., Eeswaran, R., ... & Prasad, P. V. (2024). Enhancing crop yield prediction in Senegal using advanced machine learning techniques and synthetic data. *Artificial Intelligence in Agriculture*, *14*, 99-114. https://doi.org/10.1016/j.aiia.2024.11.005

Renju, K., & Brunda, V. (2024). Optimizing crop yield prediction through multiple models: an ensemble stacking approach. *International Journal of Data Informatics and Intelligent Computing*, *3*(2), 52-58. DOI:10.59461/ijdiic.v3i2.120

Santos, L. B., Gentry, D., Tryforos, A., Fultz, L., Beasley, J., & Gentimis, T. (2024). Soybean yield prediction using machine learning algorithms under a cover crop management system. *Smart Agricultural Technology*, *8*, 100442. 1-8.https://doi.org/10.1016/j.atech.2024.100442

Sharma, A. K., & Rathore, A. S. (2024). Design and Implementation of a Cloud Based Smart Agriculture System for Crop Yield Prediction using a Hybrid Deep Learning Algorithm. *Current Agriculture Research Journal*, *12*(2), 714-725. http://dx.doi.org/10.12944/CARJ.12.2.17

Sharma, P., Dadheech, P., Aneja, N., & Aneja, S. (2023). Predicting agriculture yields based on machine learning using regression and deep learning. *IEEe Access*, *11*, 111255-111264. DOI: 10.1109/ACCESS.2023.3321861

Stiller, S., Grahmann, K., Ghazaryan, G., & Ryo, M. (2024). Improving spatial transferability of deep learning models for small-field crop yield prediction. *ISPRS open journal of photogrammetry and remote sensing*, *12*, 1-11.https://doi.org/10.1016/j.ophoto.2024.100064

Subramaniam, L. K., & Marimuthu, R. (2024). Crop yield prediction using effective deep learning and dimensionality reduction approaches for Indian regional crops. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, *8*, 1-9.https://doi.org/10.1016/j.prime.2024.100611

Sudhamathi, T., & Perumal, K. (2024). Ensemble regression based Extra Tree Regressor for hybrid crop yield prediction system. *Measurement: Sensors*, *35*, 1-9. https://doi.org/10.1016/j.measen.2024.101277

Talaat, F. M. (2023). Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. *Neural Computing and Applications*, *35*(23), 17281-17292. https://doi.org/10.1007/s00521-023-08619-5

Vhatkar, K. N., Koparde, S. A., Kothari, S., Sarwade, J., & Sakur, K. (2025). Enhancing prediction of crop yield and soil health assessment for sustainable agriculture using machine learning approach. *MethodsX*, *14*, 1-15.https://doi.org/10.1016/j.mex.2025.103418

Xiang, W., Long, L., Liu, Z., Dai, F., Zhang, Y., Li, H., & Cheng, L. (2025). A crop model based on dual attention mechanism for large area adaptive yield prediction. *Smart Agricultural Technology*, *11*, 1-9.https://doi.org/10.1016/j.atech.2025.100957

Zare, H., Weber, T. K., Ingwersen, J., Nowak, W., Gayler, S., & Streck, T. (2024). Within-season crop yield prediction by a multi-model ensemble with integrated data assimilation. *Field Crops Research*, *308*, 1-16. https://doi.org/10.1016/j.fcr.2024.109293