Agris on-line Papers in Economics and Informatics

Volume XVIII Number 3, 2026

Evaluation of AI Predictability on Crop Yield using XAI - An Experimental Approach

Ramesh Varathan[1], P. Kumaresan[2]

[1] School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Tamil Nadu, India

[2] Centre For Artificial Intelligence Research (C-FAIR), Vellore Institute of Technology, Tamil Nadu, India

Abstract

Accurate crop yield prediction is essential for improving agricultural planning and sustainability. This study explores the integration of explainable artificial intelligence (XAI) with machine learning models to enhance transparency in agricultural decision-support systems. A comparative evaluation of multiple regression algorithms - Linear, Ridge, Lasso, Elastic Net, Decision Tree, Random Forest, AdaBoost, XGBoost, CatBoost, LightGBM, Gradient Boosting, and K-Nearest Neighbour Regression (KNR) is performed for crop yield prediction. Categorical attributes are transformed using One-Hot Encoding, while model reliability is ensured through 5-Fold Cross-Validation and hyperparameter optimization using Grid Search. Among the evaluated approaches, KNR demonstrated the best predictive capability with an R^2 value of 0.9827 and reduced errors (MAE:9.03, RMSE:117.58, MSE:13824.73). To improve interpretability, SHAP, LIME, and ELI5 techniques are employed to explain feature contributions and model decision. The proposed framework provides transparent and reliable insights, supporting data-driven agricultural management and sustainable crop production strategies.

Keywords

XAI, SHAP, LIME, CYP, Agriculture, ML, KNR.

Varathan, R. and Kumaresan P. (2026) "Evaluation of AI Predictability on Crop Yield using XAI - An Experimental Approach", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 3, pp. 79-101. ISSN 1804-1930. DOI 10.7160/aol.2026.180307.

Introduction

Farming is essential to human society, supplying food and serving as a foundation for employment and economic stability. Although humans have consumed grains and plants for over 10,000 years, the organized cultivation of crops and the management of arable land are comparatively recent advancements. Active land and vegetation management started about approximately 11,000 years ago, during the Neolithic period. Agriculture plays a crucial role in India's economy, providing the majority of the country's food needs and greatly enhancing economic stability. India's rapidly expanding population and significant climate change have made it more important than ever to maintain the agri-food value chain.

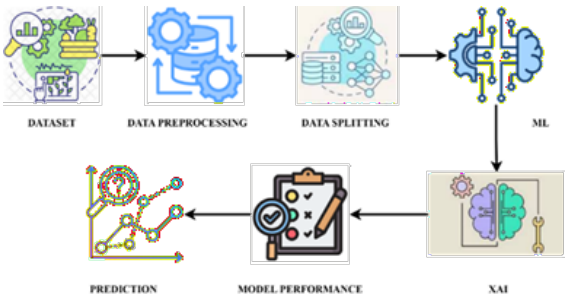
Climate change is altering temperature patterns and weather conditions, which significantly influences agricultural productivity. Variations in rainfall, rising temperatures, and unexpected climatic events can disrupt crop growth and reduce overall yield quality. These environmental changes also affect soil health, nutrient availability, and farming practices, making food production more uncertain. Addressing these challenges requires sustainable agricultural strategies and adaptive technologies to maintain stable food supplies and support long-term food security.

Many scientific methods have been applied in agriculture to balance food supply and demand. However, significant climate variability creates challenges for farmers in adopting more flexible and sustainable practices (Sharma et al., 2023). With the use of modern technology and innovative farming methods, agriculture needs to boost production while minimizing resource use. Therefore, precise crop production estimation is crucial to tackling food security issues. In India, farming is responsible for 70 percent of the global water consumption (Khan and Noor, 2019).

ML, a branch of AI, enables more accurate outcome prediction without the need for explicit

programming. ML algorithms analyze historical data to forecast predicted outcomes. DL, a specialized subset of machine learning, utilizes algorithms modelled after biological neurons. With the advent of powerful processors, massive datasets, and flexible software libraries, deep learning has become one of the most popular machine learning methods available today. DL is characterized by the use of models with several processing layers that acquire data representations at different abstraction levels (LeCun et al., 2015).

Modern AI-based systems offer sophisticated solutions to a numerous intricate challenge in sector such as agriculture, healthcare, nutrition, energy, and transportation, all of which significantly contribute to enhancing people's quality of life. In this model operate with minimal humanlike involvement and permit only a small degree of inaccuracy. Recent advancements in DL have shown impressive precision in tackling intricate tasks including facial recognition, image segmentation, and object identification. Despite their high precision, it often remains challenging for users to understand how these algorithms arrive at their decisions. Consequently, AI researchers have devoted considerable effort to developing tools, methodologies, and approaches that facilitate user understanding and reliance on the output generated by machine learning models. In this field of study, known as XAI, the goal is to elucidate the rationale and underlying processes responsible for biases in AI-driven models (Arrieta et al., 2020).



Source: Authors´ own work

Figure 1: Workflow of Machine Learning and Explainable AI System for Forecasting agricultural yield.

The use of AI in agriculture is bringing a revolution in predicting agricultural outputs. Through machine learning algorithms, these predictions provide farmers with important insights into anticipated yields and consumption trends. The limited transparency of AI models, often described as black boxes, leads to doubts among agricultural stakeholders about the reliability and credibility of the predictions they generate.

To tackle these issues, XAI has emerged as a key research focus, aiming to enhance the transparency and interpretability of AI decisions by making the reasoning behind the model outputs more understandable to users. The research investigates the interaction of forecasting agricultural output and XAI, emphasizing practical relevance and interpretability in agricultural contexts (Akkem et al., 2024). The outlines the integration of ML and explainable AI in a system developed for forecasting agricultural output (Figure 1). Using XAI for crop yield prediction involves training AI models on agricultural data and then employing XAI techniques to understand which factors most influence the model's yield forecasts. This offers clear insights into the predictions, enabling farmers and researchers to pinpoint the main factors influencing crop yield. In turn, this understanding aids in making more informed decisions to enhance agricultural practices.

This work aims to underscore the critical role of interpretability in the crop yield forecasting and to guide its seamless integration into current predictive models. In pursuit of this goal, we investigate the key issues faced when integrating AI into crop yield forecasting systems. By evaluating sophisticated XAI tools such as SHAP, LIME, and ELI5, the paper outlines a systematic approach for increasing the clarity and trustworthiness of AI-powered crop yield prediction models. As agriculture continues to embrace digital transformation, the integration of AI and XAI empowers farmers to make predictions that are not only accurate but also interpretable. By enhancing decision-making and promoting sustainable farming, this progress also aligns with the larger goal of securing global food supplies. The subsequent sections of this paper delve into the evolving domain of crop yield prediction, its foundations in AI, and the growing role of XAI in enhancing predictive systems.

Gap analysis and contribution

1. Many crop yield prediction studies primarily focus on improving prediction accuracy using machine learning and deep learning models. The proposed framework integrates multiple XAI techniques (SHAP, LIME, and ELI5) to explain model predictions.
2. Several studies rely on a simple train-test split, which may lead to overfitting and unreliable performance estimates. The framework employs 5-Fold Cross-Validation to ensure robust model evaluation. Hyperparameter tuning using Grid

Search further improves model reliability and generalization.

3. Categorical agricultural variables are sometimes improperly handled, reducing model effectiveness. One-Hot Encoding is applied to transform categorical features into machine-readable representations. This improves learning efficiency and predictive performance.
4. Current research predominantly focuses on ensemble methods such as Random Forest, XGBoost, and LightGBM. K-Nearest Neighbour Regression (KNR) is less explored despite its potential effectiveness. The proposed framework combines high predictive performance with explainability. The integration of KNR and XAI techniques provides accurate predictions while maintaining transparency and accountability.

The existing literature on crop yield prediction predominantly emphasizes predictive accuracy while overlooking model interpretability, comprehensive algorithm comparison, and robust validation mechanisms. To address these limitations, this study proposes an explainable machine learning framework that integrates extensive regression model evaluation, hyperparameter optimization, 5-fold cross-validation, and multiple XAI techniques (SHAP, LIME, and ELI5). The framework achieves superior prediction performance using KNR while simultaneously providing transparent explanations of model decisions, thereby supporting trustworthy and sustainable agricultural decision-making.

The contribution of this work for crop yield prediction is specified below:

1. Comprehensive comparative evaluation of twelve ML regression algorithms for crop yield prediction.
2. One-Hot Encoding for Nominal agricultural features without Ordinal Bias.
3. Robust Validation through 5-fold cross-validation combined with GridSearchCV.
4. Identification of KNR as the best performing model with better metrics.
5. XAI framework deployment of SHAP, LIME, and ELI5.
6. Actionable decision support insights mapped to three stakeholder groups.
7. Comparative Performance Analysis of Models for Production Prediction
8. SOTA comparison demonstrating competitive superiority.

The manuscript is further organized in the following manner. Section 2 presents a comprehensive review of related work and discusses the foundational concepts of ML and XAI in agricultural domain. Section 3 delineates the methodological framework, including a detailed description of the dataset and feature engineering processes, and preprocessing techniques. Section 4 reports the experimental results and provides an in-depth analysis and interpretation of the findings. Section 5 discusses outcome of the implementation and the analysis of various ML models and XAI interpretation. Finally, Section 6 concludes the key contributions and outlining potential for future research. The major contributions are finally summarized in Section 6, which also outlines areas that could be explored further.

Related work

In the relevant article, the importance of yield predictions from several key ML and XAI papers is addressed. The relevant work focuses at the importance of ML and XAI in several fundamental research studies pertaining to crop yield forecasts.

The combined ML with agronomic crop modeling principles to develop a foundational approach for large scale agricultural productivity forecasting. Their methodology followed a system focused on reusability, accuracy, and modularity. Feature extraction was performed using crop simulation results in combination with meteorological, remote sensing, and soil datasets from the MCYFS database (Abasha et al. 2023; Dilli et al., 2023). However, this work focused on large-scale pan-European forecasting and did not incorporate XAI interpretability methods to explain individual model predictions. Khaki and Wang (2019) introduced a deep neural network framework for predicting corn hybrid yield, verifying yield outputs, and evaluating yield variability, utilizing a combination of genotype data and environmental variables, including soil and weather conditions. Incorporating weather data showed that the model performed well, with an RMSE of 12 percent for average yield predictions and 50 percent for the standard deviation. However, the study was limited to a single crop type and did not provide any explainability of feature contributions. Abbas et al. (2020) investigated the effectiveness of four machine learning algorithms: Elastic net, linear regression, k-nearest neighbors, and support vector regression for predicting potato tuber yield across six field datasets in Atlantic Canada. The study focused on data collected during the 2017 and 2018 growing

seasons from fields located in Prince Edward Island and New Brunswick, labelled as PE2017, PE2018, NB2017, and NB2018. Among the algorithms tested, SVR consistently outperformed the others across all datasets, achieving RMSE values of 5.97, 4.62, 6.60, and 6.17 t/ha for PE2017, PE2018, NB2017, and NB2018, respectively. A limitation of this study is that only four algorithms were compared and no XAI techniques were applied to explain the predictions. Vijayan and Chowdhary (2025) introduced a hybrid optimization algorithm, WOA_APSO (Whale Optimization Algorithm combined with Adaptive Particle Swarm Optimization), to enhance disease classification in rice crops. This approach integrates bio-inspired optimization techniques with deep learning models, specifically utilizing a Convolutional Neural Network (CNN) for classification tasks. The hybrid algorithm focuses on optimizing feature selection, thereby improving the CNN's performance in identifying various rice crop diseases. The proposed model achieved a remarkable accuracy of 97.5% on the benchmark dataset, outperforming models such as Support Vector Machines and Artificial Neural Networks. However, this work addressed disease classification rather than yield quantity prediction, and model decisions were not made interpretable through XAI. Khosla et al. (2020) concentrated on predicting kharif crop yields in the Visakhapatnam region of Andhra Pradesh. They employed a Modular Neural Network (MNN) to forecast monsoon rainfall, a key factor affecting kharif crop productivity. Following this, Support Vector Regression (SVR) was used to estimate actual yields using regional rainfall data. This integrated MNN-SVR approach supported the implementation of effective agricultural practices, resulting in enhanced crop yields and outperforming other prediction methods. While effective, the study was region-specific and lacked a systematic comparison across multiple ML algorithms or any XAI integration. Fagade and Pawar (2020) applied SVM and ANN models to classify crops based on environmental and soil-related parameters such as rainfall, temperature, soil type, pH, and humidity, utilizing data sourced from Maharashtra's agricultural website. The dataset encompassed nine agricultural zones, and a user-friendly interface was created to allow farmers to input relevant information. The ANN model achieved a high prediction accuracy of 86.80%, demonstrating its effectiveness.

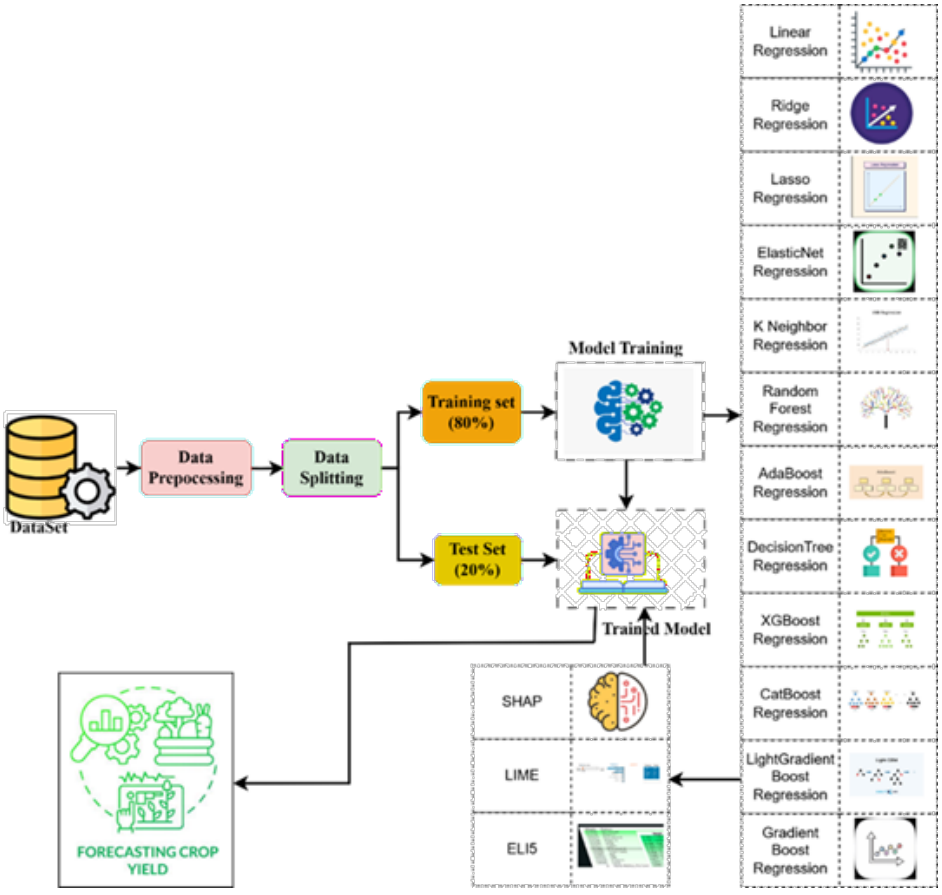
Abekoon et al. (2025) explored the predictive capabilities of soil nitrogen, phosphorus, and potassium concentrations an 85-day cabbage growth cycle using explainable machine learning methods such as SHAP and LIME. Their analysis indicated that plant height and leaf count positively influenced nutrient level predictions, while the number of growth days and average leaf area exhibited a negative impact. According to their analysis, the number of growth days and average leaf area had a negative effect on nutrient level predictions, whereas plant height and leaf count had a positive effect. The model's interpretability and practical utility were validated through experiments with greenhouse-grown cabbage, culminating in the development of a user-friendly application aimed at supporting agricultural decision-making and optimization. However, this study focused on soil nutrient prediction rather than crop yield and was restricted to a single controlled crop environment. In their study, Agarwal et al., (2025) introduced the Successive Integration Hybrid Forecasting Model (SIHFM), a hybrid forecasting framework that leverages both statistical and dynamic techniques to estimate cotton crop yields using historical datasets from 1964 to 2020. The ARIMA model was employed for initial statistical forecasting, followed by LSTM and GRU models for dynamic predictions using the outputs from ARIMA. This hybrid framework successfully captured temporal trends and nonlinear patterns, resulting in improved yield prediction accuracy. A key limitation is that the study was restricted to cotton yield in India did not include XAI techniques for model transparency. Abdel-salam et al., (2024) proposed a novel crop yield prediction framework integrating hybrid feature selection with an optimized Support Vector Regression (SVR) model. The framework uses a new FMIG-RFE hybrid feature selection method after preprocessing with K-means clustering and CFS. SVR hyperparameters were adjusted using an improved Crayfish Optimization Algorithm (ICOA), which increased prediction accuracy. However, the study relied on a single SVR model and did not benchmark against a diverse set of regression algorithms, nor did it provide XAI based interpretability. Joshi et al., (2024) associates Long Short-Term Memory (LSTM), one-dimensional Convolutional Neural Network (1D-CNN), and Bidirectional LSTM (Bi-LSTM) models to create an explainable DL framework for agricultural output forecasting. They used methods like SHAP, Integrated Gradients (IG) (Abasha et al., 2025), and LIME to interpret model decisions. Bi-LSTM performed the best out of all the models, with an R2 score as high as 0.88. According to their analysis, the most important factors influencing winter

wheat yield were temperature, precipitation, and enhanced vegetation index (EVI) during later stages of crop growth. While this study applied SHAP and LIME, it exclusively used deep learning models which are inherently less interpretable than ensemble regression methods. The present study addresses all of these limitations by applying 12 diverse regression algorithms with three XAI methods (SHAP, LIME, ELI5) in a leakage free, unified framework on Indian crop data.

Materials and methods

Proposed system

The proposed system architecture (Figure 2) presents an agricultural forecasting framework like ML regression and XAI. The process begins with the acquisition of a historical agricultural dataset that includes features such as crop, crop year, state, season, area, annual rainfall, fertilizer, pesticide, and yield. Data preprocessing is the first step, and it includes managing outliers, scaling numerical features, encoding categorical variables, and handling missing values. Following preprocessing, the dataset is split into two subsets: 20% for testing and 80% for training. LR, RR, Lasso Regression, ENR, KNR, RFR, ABR, DTR, XGBR, CBR, LGBMR, and GBR are among the machine learning regression models that are subsequently created using the training data. Following the training phase, model performance is quantitatively assessed on the test dataset using standard evaluation metrics, including MAE, MSE, RMSE, and R2_Score. To improve model transparency and interpretability, Explainable AI techniques like SHAP, LIME, and ELI5 are applied. These methods allow for a clear comprehension of the decision-making process by providing insights into the significance of features and the logic underlying model predictions. Ultimately, the most accurate and interpretable model is chosen for crop yield forecasting, aiding in informed agricultural planning and management.



Source: Authors´ own work

Figure 2: Predictive System Architecture for Agricultural Yield Forecasting using ML and XAI.

Pseudo Code: Agricultural Yield Forecasting using Supervised Regression Models Integrated with Explainable Artificial Intelligence

Input:

- Dataset (D):
- Features (X): crop, crop year, season, state, annual rainfall, area, fertilizer, pesticides.
- Target (y): yield
- Regression Models (M): Any regression model in ML.
- Explainability Methods: SHAP, LIME, ELI5.
- Evaluate the model with suitable performance metrics.

Algorithm steps:

Step 1: Load and Preprocess the Dataset

- Load the dataset D and handle missing values.
- Encode categorical features (Crop, Season, State) using one-hot encoding.
- Normalize numerical features (Annual_ Rainfall, Area, Fertilizer, Pesticides, Crop_Year) using MinMaxScaler or StandardScaler.
- Partition the dataset into training (80%) and testing (20%).

Step 2: Cross-Validation & Hyperparameter Tuning

- Initialize k-fold CV with k splits
- Define parameter grid for model M
- Apply GridSearchCV with pipeline, param_grid, and k-fold CV
- Fit GridSearchCV on {X_train, y_train}
- best_pipeline = BRST_ESTIMATOR from GridSearchCV.

Step 3: Train the Regression Model (M)

- Select an appropriate supervised regression algorithm M for modelling the relationship between input features and the target variable. The relationship between the target variable and the input features is modeled by algorithm M.
- Fit the selected regression model M using the training dataset to learn the underlying mapping from features to target output.

Step 4: Model Evaluation

- Generate predicted target values best_ pipeline by applying the trained model M to the test feature set (X_test)
- Compute performance measures:
  - Root Mean Squared Error (RMSE)
  - Mean Squared Error (MSE)
  - Mean Absolute Error (MAE)
  - R2_Score

Step 5: Explainability Using XAI Methods

Step 5.1: Global Explainability Using SHAP

- To measure the global contribution of each input feature to the model's predictions, compute SHAP values.
- Rank input features according to their relative importance in influencing the model's crop yield predictions, as determined by the computed SHAP values.

Step 5.2: Local Explainability Using LIME

- Select a test instance and generate perturbations.
- Train a simple local model to approximate the predictions of M.
- Identify the most important features influencing this specific prediction.

Step 5.3: Model Debugging & Explanation Using ELI5

- Use ELI5 to get feature importance rankings and visual explanations for how predictions are made.
- Generate weight-based feature importance for regression models.

Step 6: Decision Support and Interpretation

- If SHAP highlights "State" and "Season" globally → Useful for regional planning.
- If LIME shows high importance for "Annual_Rainfall" and "Fertilizer" → Farmers should focus on these factors.
- If ELI5 confirms feature importance rankings → Model's decisions are interpretable.

OUTPUT:

- Predicted Crop Yield (y') for new inputs.
- Best_Params = best_pipeline parameters.
- Performance evaluation metrics.
- Feature importance ranking from SHAP, LIME, and ELI5.
- Human-readable explanations for decisions using ELI5.

Dataset

Mendeley Data's historical agricultural production crop dataset (http://doi.org/10.17632/ncw2vbcgnk.2), encompasses records from 1997 to 2020, spanning 27 Indian states and 3 Union Territories. It comprises 19,689 instances, each described by ten distinct attributes as outline (Table 1), with detailed attribute definitions provide (Table 2). The dataset captures six agricultural seasons and includes data on 55 unique crop types cultivated across diverse agro-climatic regions of India.

Dataset	Total number of attributes	Number of instances
Crop Yield Dataset	9	19689

Source: Authors´ own analysis

Table 1: Dataset description.

No of Attributes	Description of the attributes
crop	Specifies the crop species or variety cultivated during the recorded agricultural cycle.
crop year	Denotes the specific calendar year corresponding to the cultivation period of the crop.
season	The period of the agricultural cycle during which the crop was planted and harvested
state	The Indian state where the agricultural activity took place.
area	The size of the land (in hectares) used for cultivating the crop.
annual rainfall	The annual rainfall (in millimeters) record in the crop cultivation area.
fertilizer	The crop's fertilizer dosage, expressed in kilograms.
pesticide	Pesticide quantity applied during crop cultivation, measured in kilograms.
yield	The production of the crop, calculated as yield per unit of cultivated area.

Source: Authors´ own analysis

Table 2: Description of the attributes.

Data preprocessing

Data preprocessing is essential to the effectiveness of AI and ML. Given that ML models are highly dependent on data quality, proper preprocessing is necessary to enable efficient learning before the data is fed into the model.

One hot encoding

A preprocessing method is used to translate categorical variables to a number that can be used in a machine learning model is one-hot encoding (Bolikulov et al., 2024). It generates individual binary columns for every category and makes it easy for models to deal with nominal data without causing artificial ranking among the categories. This technique helps make sure that each category is treated individually and avoids bias that comes from the assignment of numerical values to categorical attributes. In case of agricultural datasets, one-hot encoding helps to represent the difference across states, seasons, and crop types which can effectively help the prediction models to capture the regional and seasonal influence on the crop yield.

Training and testing data

The dataset is separated into two sections: the test dataset and the training dataset. Our model trains and tests using the dataset in an 80:20 random state of 42.

Machine learning models

Linear regression

Linear regression seeks to find a straight-line relationship between input features and the output feature. It assumes that the residuals are normally distributed with constant variance and that there is no correlation among the predictor variables (Samira et al., 2022).

Formula for predicting crop yield with nine input features and one output feature can be represented as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \\ + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon \quad (1)$$

where,

Y: The output feature, representing the predicted crop yield.

X_1, X_2, ……, X_8 = The nine input features, which could include factors such as Crop, Crop_Year, State, Season, Area, Annual_Rainfall, Fertilizer, Pesticide.

β_0 = The intercept term, representing the baseline crop yield when all input features are zero.

β_1, β_2, ……, β_8 = The coefficients foe the respective input features, indicating their impact on the crop yield.

ε = The error term, accounting for any variations not explained by the model.

The coefficients are determined using methods like Ordinary Least Squares. By interpreting the coefficients, one can understand the impact of each factor on crop yield. For instance, a positive coefficient for rainfall indicates an increase in yield with more rainfall.

Ridge Regression

Ridge Regression (RR) is a regularized regression technique that improves the stability of prediction models when strong correlations exist among independent variables. It introduces an L2 penalty term that shrinks the regression coefficients, helping to control model complexity and reduce overfitting. This approach stabilizes parameter estimation and improves prediction reliability while retaining all input variables in the model. As a result, ridge regression is widely used in agricultural data analysis and decision-support systems to handle multicollinearity and enhance model robustness (Sudha et al., 2022).

$$J(\beta) = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j{}^2 \quad (2)$$

Where:

$J(\beta)$ = Cost function to minimize

Y_i = Actual crop yield for the ith observation

β_0 = Intercept term

β_j = Coefficients of the j^{th} input feature

X_{ij} = Value of the jth feature for the ith observation

λ = Regularization parameter

n = Number of observations

p = Number of input features

Lasso regression

A linear regression method that combines regularization and variable selection is called the Least Absolute Shrinkage and Selection Operator (LASSO). LASSO is frequently used in agricultural settings. It operates by limiting the sum of the absolute values of the model coefficients by α and minimizing the traditional sum of squared errors. By reducing the coefficients of redundant or irrelevant features to precisely zero, the hyperparameter α enables the model to choose only the most significant predictors. LASSO regression's primary goal is to find a feature subset that reduces the prediction error for a continuous response variable. It is regarded as a simple yet effective linear model that highlights the most important variables in large datasets and has built-in regularization. However, for LASSO to function well, feature scaling and regularization parameter tuning must be done carefully (Sarr and Sultan, 2023).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \lambda\sum_{i=1}^{n}|\beta_i| \quad (3)$$

Where,

Y = Predicted crop yield

$X_1, X_2, \ldots, X_n$ = The independent input features

β_0 = Intercept term

$\beta_1, \beta_2, \ldots, \beta_n$ = The feature coefficients

λ = The regularization parameter that controls the strength of the penalty.

The coefficients are estimated by minimizing the lasso loss function. Its ultimately delivers interpretable results by selecting the most influential features affecting crop yield, making it particularly suitable for high-dimensional agriculture datasets.

Elastic Net regression

Elastic Net is a regression technique designed to handle situations where there is a high correlation among the features of a Agri dataset. It addresses the limitations of both Lasso Regression and Ridge Regression, providing a balanced approach to selecting the most relevant predictors while minimizing prediction errors. Unlike Lasso and Ridge individually, Elastic Net applies both L1 (Lasso) and L2 (Ridge) penalties simultaneously, resulting in a more effective and accurate prediction model (Kuruguntu Mohan et al., 2022).

$$Y = \beta_0 + \sum_{i=1}^{n}\beta_i X_i + \lambda_1\sum_{i=1}^{n}|\beta_i| + \lambda_2\sum_{i=1}^{n}\beta_i{}^2 \quad (4)$$

Where:

Y = Agricultural forecasting output

β_0 = Intercept term

β_i = Coefficients of the input features

X_i = Input features

n = Number of features

$\lambda_1\sum_{i=1}^{n}|\beta_i|$ = L1 regularization (Lasso),

$\lambda_2\sum_{i=1}^{n}\beta_i{}^2$ = L2 regularization (Ridge)

λ_1 and λ_2 = Hyperparameters controlling the strength of L1 and L2 regularization

K-Neighbor regression

The K-Nearest Neighbor algorithm is a distribution-free technique used for forecasting of agricultural output. In KNN, the parameter k plays a critical role in determining prediction accuracy. For a given test sample, the algorithm identifies k nearest neighbors from the training dataset and estimates the yield by averaging the response values of these k-neighbors (Maya Gopal and Bhargavi, 2019).

$$Y = \frac{1}{K}\sum_{i=1}^{k} Y_i \quad (5)$$

Where,

Y = Predicted crop yield

K = Number of nearest neighbors

Y_i = Actual crop yield of the ith nearest neighbor

Random Forest Regression

Random Forest Regression (RFR) is a robust ensemble learning technique widely used in crop yield prediction, along with resist the overfitting, capability to non-linear relationships, and efficiency in managing high-dimensional datasets. This method aggregates multiple decision trees to enhance predictive accuracy and stability, making it suitable for complex agricultural datasets that include factors like rainfall, fertilizer usage, Irrigation, Season, Area, and climatic conditions (Liaw and Wiener, 2002). RFR has been successfully applied in crop yield prediction across different regions and crop types. It leverages historical numerical datasets that include environmental, agronomic, and meteorological features.

$$Y = \frac{1}{M}\sum_{m=1}^{M} T_m(X) \quad (6)$$

Where,

Y = Predicted crop yield

M = The total number of estimators (decision trees) in the random forest model

$T_m(X)$ = Output of the m-th decision tree for the input X.

AdaBoost Regression

Adaptive Boosting Regression (ABR) is a ML technique that combines multiple weak learners in a sequential manner to form a strong predictor. The process starts by initializing weights for all training samples. In each iteration, a base estimator is fitted, and greater weight is given to the misclassified or poorly predicted observations, allowing the next learner to focus more on these difficult cases. This process is repeated, with each new learner aiming to correct the errors of its predecessor, ultimately improving the overall prediction performance (Nguyen and Ngo, 2025).

Decision Tree Regression

A decision tree addresses both classification and regression challenges through two primary functions. First, it selects the most significant features that influence decision-making. Second, it establishes the best possible outcome based on these selected features, often yielding a probability distribution over the potential results. Within a decision tree structure, each internal node represents a feature, each branch corresponds to a decision or split, and each leaf node indicates a final prediction or outcome. To initiate the tree construction, a specific feature serves as the root node. The tree-building process is then completed by recursively splitting the data (Burdett and Wellen, 2022).

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad (7)$$

Where:

$\hat{Y}$ = represents the estimated agricultural output

N = denotes the total number of observations within a terminal (leaf) node

y_i = represents the actual crop yield of the i^{th} sample in the terminal node

XGBoost regression

XGBoost is a well-known and effective open-source library that uses decision trees to implement the gradient-boosting algorithm. Regression trees are used by XGBoost as weak learners in regression tasks. A continuous score connects each input data point to a leaf. The model incorporates a regularization term (L1 and L2) to control complexity while optimizing a convex loss function. In order to predict the residuals (errors) of earlier trees, new trees are iteratively added during training. The final prediction is then improved by integrating these new trees with the ones that already exist. This method is frequently called "gradient boosting" because it continuously reduces loss with each additional model (Pavithra and Soundrapandiyan, 2024).

$$Y = F_M(X) = \sum_{m=1}^{M} f_m(X) + F_0 \quad (8)$$

Where:

Y = Predicted crop yield

$F_M(X)$ = Final XGBoost model prediction after M boosting iterations.

$F_m(X)$ = The prediction generated by the m^{th} weak learner.

F_0 = Initial forecast

M = Total number of estimators (iterations)

Cat Boost regression

Cat Boost is a gradient boosting-based machine learning specifically optimized for handling categorical features efficiently, where all nodes are split in the same way at each depth level. This structure helps to prevent overfitting and enhances computational efficiency (de Villiers et al., 2024).

Light Gradient Boost Machine regression

Light Gradient Boosting Machine (LGBM) is an ensemble algorithm based on gradient boosting decision trees. Unlike traditional decision trees that expand level wise, LightGBM grows leaf-wise, leading to more efficient and accurate splits. It uses histogram-based, cost-effective techniques where the time complexity depends on the number of bins rather than the size of the dataset (Mitra et al., 2024). The model generates predictions using an ensemble of decision trees. The forecast for crop yield Y' is obtained by summing the outputs of all trees in the ensemble.

$$Y' = \sum_{m=1}^{M} f_m(X) \quad (9)$$

Where,

Y' = Agricultural forecasting output

M = The overall number of decision trees included in the ensemble model

$f_m(X)$ = The output of the mth tree for the input feature set X

Gradient Boosting Regression

Gradient Boosting Regression (GBR) is a ML technique that builds predictive models by combining multiple weak learners, typically decision trees, to improve prediction accuracy. The method sequentially reduces prediction errors by learning from previous model residuals and refining the model in successive iterations. GBR is effective in agricultural forecasting because it can handle complex datasets and capture non-linear relationships among climatic variables, soil factors, and historical yield records. As a result, it provides reliable insights for crop yield prediction and supports informed decision-making in agricultural management (Mishra et al., 2020).

$$Y = F_M(X) = F_0(X) + \sum_{m=1}^{M} \gamma_m T_m(X) \quad (10)$$

Where,

Y = Predicted crop yield

M = Number of boosting iteration (trees).

$F_M(X)$ = Final boosted model prediction

$F_0(X)$ = Initial prediction.

$T_m(X)$ = Prediction from the m-th weak learner (decision tree)

γ_m = Learning rate (step size) controlling the contribution of each tree.

Explainable Artificial Intelligence for crop yield prediction

An emerging field known as XAI provides various methods to make the opaque nature of machine learning (ML) and deep learning (DL) models more transparent, allowing for explanations that are comprehensible to humans. Due to significant advancements in the application of ML and DL techniques, researchers in AI and ML are increasingly focusing on XAI.

SHapley Additive exPlanations

The Shapley Value, derived from game theory, quantifies the incremental impact of each player to the overall outcome (Ponce-Bobadilla et al., 2024). The explainable AI tool SHAP utilizes Shapley values to assess feature contributions to predictions.

$$g(z') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j z'_j \quad (11)$$

Where,

g = represents the interpretation model.

$z' \in \{0,1\}^M$ = indicates whether a specific feature is observable (1) or not (0).

Each feature is assigned a Shapley value (ϕ_i), a real number representing its contribution. Essentially, Shapley values work by calculating a feature's average contribution across every possible combination of features (Molnar, 2020). SHAP provides both global and local insights into model behavior and offers a model-agnostic way to estimate these Shapley values.

Local Interpretable Model-Agnostic Explanations

LIME is a strategy that provides a locally interpretable model by roughly modeling any black box machine learning model to clarify each prediction (Ribeiro et al., 2016). This technique works by first altering the original data points and then using these altered points to get predictions from the main model. Next, it trains an interpretable model using these new data points. LIME weights these new points based on how close they are to the original ones. This way, the resulting "explanation model" can explain each of the original data points. In this process, $\Omega(g)$ measures how complex the interpretable model g. The π_x indicates the size of the neighborhood around the original data point x, and f is

the complex model we're trying to understand. The explanation model, *g*, is designed to be as accurate as possible in mimicking the original model's (*f*) predictions, which is measured by minimizing *L(f, g, π_x)*.

$$explanation(x) = argmin_{g \in G}\ \mathcal{L}(\mathrm{f}, \mathrm{g}, \pi_x) + \Omega(\mathrm{g}) \quad (12)$$

Explain Like I'm 5

ELI5 intended for the interpretation of black-box machine learning models, providing both global and local explanations. It calculates feature importance for tree-based models by assessing how much each feature influences the prediction score, based on the change in score from a parent node to a child node within the decision tree (Bhattacharya, 2022).

$$\hat{Y} = Base\ Value + \sum_{i=1}^{n} Contribution\ (x_i) \quad (13)$$

Performance metrics

Root Mean Square Error

Root Mean Squared Error is a standard metric in regression analysis used to quantify the difference between predicted and actual values. It's calculated by taking the square root of the average of the squared differences between these values. A smaller RMSE indicates that the model fits the data more accurately (Uribeetxebarria et al., 2023). This metric is frequently used to evaluate the precision and effectiveness of regression models.

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

Where, *n* = no. of crop yield records

y_i = actual crop yield for the ith observation

$\hat{y}_i$ = predicted crop yield for the ith observation

$y_i - \hat{y}_i$ = prediction error for the ith observation.

Coefficient of determination

The coefficient of determination, more commonly known as the R2_Score, is a statistical measure used in predictive modeling. It assesses how well the independent variable(s) explain the variation in the dependent variable. Essentially, it tells us the proportion of the dependent variable's total variation that the model successfully accounts for. The R2_Score is calculated by dividing the explained variance by the total variance (Saadio et al., 2022).

$$\mathrm{R2\_Score} = 1 - \frac{\sigma_r^2}{\sigma^2} \quad (15)$$

In this context, σ_r^2 signifies the sum of the squared difference between the expected and the observed values, while σ^2 represents the sum of the squared difference between the observed values and their mean.

Mean Squared Error

Mean Squared Error is a standard evaluation metric that computes the average of the squared discrepancies between the actual and predicted crop yield values (Joshi et al., 2023). By squaring each error term, the MSE places a greater penalty on larger errors, which helps in identifying models that make significant prediction mistakes. As *r* result, MSE is highly effective for measuring the precision and consistency of crop yield forecasting models.

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i}\right)^2 \quad (16)$$

Where:

n = Total number of observations

Y_i = Actual crop yield for the i-th observation

$\hat{Y}_i$ = Predicted crop yield for the i-th observation.

Mean Absolute Error

MAE is a statistical metric used in regression analysis that measures the average magnitude of error by calculating the absolute differences between the forecast and observed yield values (Oikonomidis et al., 2022). Mathematically, MAE is often represented as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \quad (17)$$

Where:

Y_i represents the observed value for the i^{th} data value

$\hat{Y}_i$ represents the estimated value for ti^{th} data value

n represents the total number of data values.

Results and discussion

This section details the experimental results of the proposed ML model and XAI applied to crop yield prediction. The study employed systematically developed ML and XAI models. Model performance was assessed using standard regression metrics and compared against other ML models to identify the most suitable candidate for providing input to the XAI framework. The experiments were conducted in Python on a system equipped with an Intel Core i9-12900 processor, 32 GB of RAM, a 1 TB SSD and an NVIDIA RTX graphics card.

The comparison of various machine learning regression models used for crop yield prediction, evaluated using four key performance metrics: MAE, MSE, RMSE, and R2_Score (Table 3 below). Together, these metrics offer a comprehensive assessment of each model's accuracy and reliability. Among all the models, the KNR achieved the best overall performance, with the lowest MAE (9.03), MSE (13824.73), RMSE (117.58), along with a high R2_Score of 0.9827, indicating excellent predictive capability. Other high-performing models include RFR, which also yield low errors (MAE: 9.47, MSE: 15682.08, RMSE: 125.23) and a strong R2_Score of 0.9804. GBR, CBR, and DTR also show strong performance with R2_Scores above 0.97, reflecting their effectiveness in capturing complex patterns in the data. In contrast, the linear models – LR, RR, and Lasso Reg. – have relatively high MAE (ranging from 60.97 to 64.12) and RMSE values (389.88), indicating that they are less effective for this task due to their limitations in handling non-linear relationships. Elastic Net Regression shows the worst performance across all metrics, with the highest MAE (420.82), MSE (177091.92), RMSE (79.47), and a low R2_Score of 0.7790, suggesting that it is not suitable for accurate crop yield prediction in this context. ABR shows the mid performance of MAE (26.63), MSE (84741.90), RMSE (291.10), and R2_Score (0.8942). LGBM and XGBR fall in the middle tier. LGBM has decent MAE (18.0) and a respectable R2_Score (0.93), while XGBR is slightly less accurate with higher errors but still better than linear models.

Cross Validation (K-Fold)

The Cross Validation (K-Fold) technique is used to forecast how well models will perform. The dataset is divided into folds, which are groups of k identical dimensions. The model uses k-1 folds for training, and this process is repeated k times. We obtain a reliable evaluation of the algorithm's adaptability by computing the average performance across all iterations. This technique is essential for choosing models, assessing overall performance, and avoiding overfitting in machine learning models. Overconfidence in the model's performance may result from training it on the complete dataset, despite the fact that it may produce amazing results. A training set and a testing set are the two subsets into which the dataset is usually separated. While the independent test set provides an objective assessment of the chosen model, the training set is utilized in the cross-validation procedure to adjust model parameters. The cross-validation process uses one of the five equal segments of a dataset as the test set and trains the model on the other four segments. An accurate assessment of the model's performance is ensured by rotating the test set across the folds.

1st Fold Training set: 15751 Testing set: 3938 split 1

2nd Fold Training set: 15751 Testing set: 3938 split 2

3rd Fold Training set: 15751 Testing set: 3938 split 3

4th Fold Training set: 15751 Testing set: 3938 split 4

5th Fold Training set: 15752 Testing set: 3937 split 5

Grid Search Cross Validation

Grid Search CV is used to determine suitable hyperparameters for ML models in CYP. It systematically tests various parameter combinations and evaluates them using cross-validation to prevent overfitting. Selecting the parameter values that provide the best results

Cross Validation	ML Models	Performance Metrics			
		MAE	MSE	RMSE	R2_Score
K-Fold = 5, GridSearch CV	K-Neighbor Regression	9.03	13824.73	117.58	0.9827
	Random Forest Regression	9.47	15682.08	125.23	0.9804
	Gradient Boosting Regression	11.78	18727.51	136.85	0.9766
	CatBoost Regression	12.25	19124.43	138.29	0.9761
	Decision Tree Regression	9.38	23184.71	152.27	0.9711
	LGBM Regression	18.00	50112.14	223.86	0.9375
	XGBoost Regression	14.85	61515.74	248.02	0.9232
	AdaBoost Regression	26.63	84741.90	291.10	0.8942
	Linear Regression	63.88	151851.10	389.68	0.8105
	Ridge Regression	64.12	152006.70	389.88	0.8103
	Lasso Regression	60.97	151998.06	389.87	0.8103
	ElasticNet Regression	79.47	177091.92	420.82	0.7790

Source: Authors´ own analysis

Table 3: Comparison of ML Model Performance in agricultural forecasting.

increase the model's reliability and accuracy. This approach supports agricultural yield forecasting by considering factors such as rainfall, area, pesticides, and fertilizer.

Root Mean Square Error Comparison

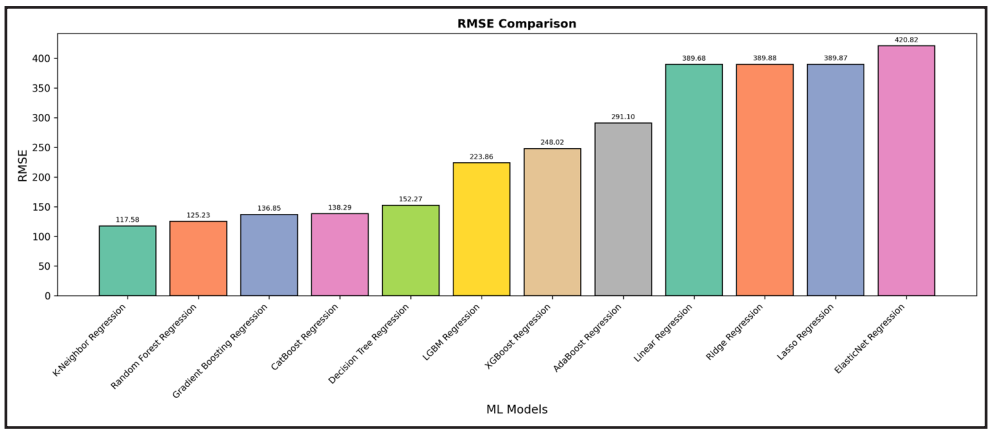
The RMSE comparison in the predictive performance of various ML models applied to CYP (Figure 3). It is evident that linear models such as LR, RR, and Las. Reg. yielded moderated results with RMSE values around 389 and ENR (420.82) performed poorly, highlighting their inability to capture the complex non-linear relationships in agricultural data. In contrast, non-linear and ensemble-based models demonstrated superior performance. The K-Neighbors Regressor achieved the lowest RMSE (117.58), followed closely by RFR (125.23), GBR (136.85), and CBR (138.29) confirming their effectiveness in handling the variability of agricultural features. Other ensemble methods such as DTR (152.27) also performed better than linear models, whereas LGBMR (223.86), XGBR (248.02), and ABR (291.10) showed relatively higher error rates. Overall, the results indicate that non-linear ensemble techniques are more suitable for accurate CYP compared to traditional linear regression methods.

Mean Absolute Error comparison

The MAE comparison in the performance of different ML models for CYP based on their MAE values (Figure 4). Lower MAE values indicate higher predictive accuracy, as they represent smaller average deviations between actual and predicted yields. Among the models, KNR achieved the best performance with the lowest MAE of 9.03, followed closely by DTR (9.38), RFR (9.47), GBR (11.78), CBR (12.25), XGBR (14.85), all of which showed strong predictive ability. In contrast, linear models such as LR (63.88), RR (64.12), and ENR (79.47) exhibited much higher error rates, reflecting their limitations in handling the non-linear and complex relationships present in agricultural datasets. Lasso Reg. (60.97) performed slightly better than other linear approaches but still fell behind the ensemble and tree-based models. ABR (26.63) and LGBMR (18.00) achieved moderate accuracy compared to the best performing models. Overall, the results highlight that non-linear ensemble and neighbor-based methods are significantly more effective than linear models in accurately predicting crop yield.

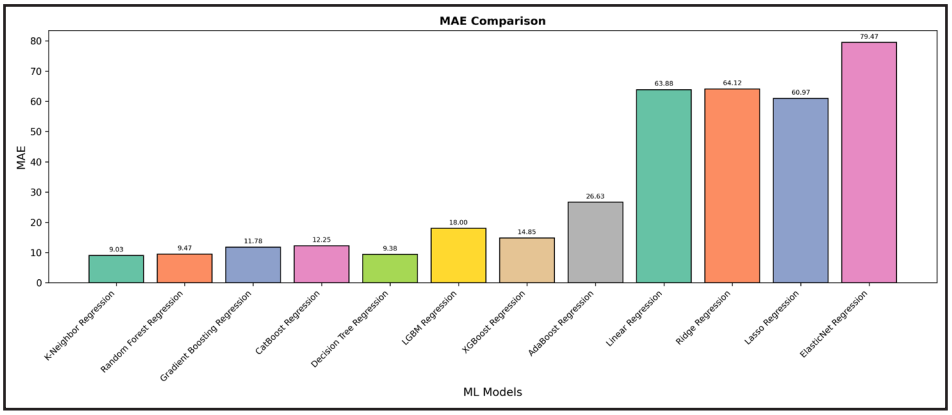
Mean Squared Error comparison

The MSE comparison in presents the performance of different ML models for CYP using MSE as the evaluation metric (Figure 5). Since MSE penalizes larger errors more heavily, lower values indicate higher prediction accuracy. From the Figure 5, it is evident that linear models such as LR (151851.10), RR (152006.70), Lasso Reg. (151998.06), and ENR (177091.92) performed poorly, producing very high error values and failing to capture the complex non-linear relationships in the dataset. In contrast, non-linear models and ensemble-based methods achieved much lower MSE values. The best-performing models were KNR (13824.73), RFR (15682.08), GBR (18727.51), CBR (19124.43), all of which demonstrated strong predictive capability. DTR (23184.71), LGBMR (50112.14), and XGBR (61515.74) also performed reasonably well, though with slightly higher error values compared to the top models. ABR (84741.90) showed moderate accuracy but was less effective than other ensemble approaches. Overall, the results



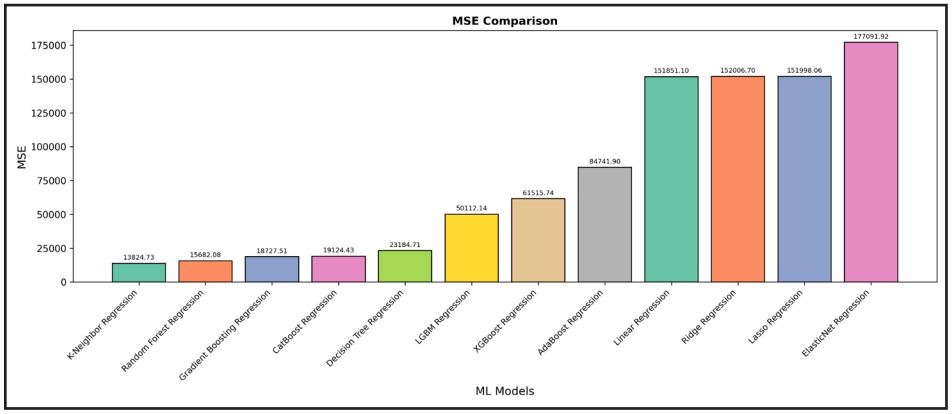
Source: Authors´ own analysis

Figure 3: RMSE-based performance comparison of ML Regression Models for agriculture.



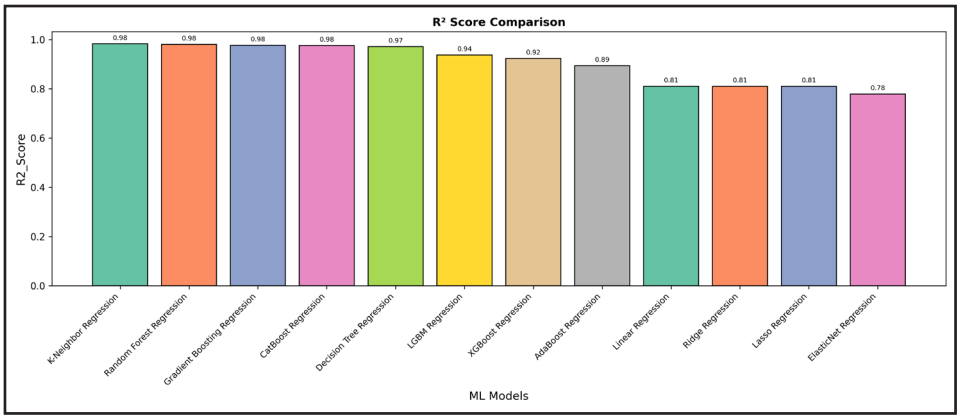
Source: Authors´ own analysis

Figure 4: MAE-based performance comparison of ML Regression Models for agriculture.



Source: Authors´ own analysis

Figure 5: MSE-based performance comparison of ML Regression Models for agriculture.



Source: Authors´ own analysis

Figure 6: R2_Score-based performance comparison of ML Regression Models for agriculture.

highlight that ensemble-based boosting methods and neighbor-based techniques significantly outperform traditional linear regression models, making them more suitable for accurate and reliable CYP.

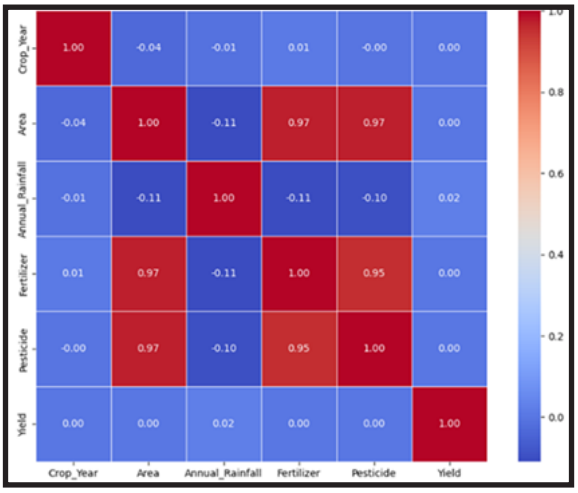
R2_Score comparison

The R2_SCORE comparison in demonstrates the goodness of fit of different ML models used for CYP (Figure 6 above). The R2_SCORE implies the model performance. From the results, linear

models such as Lasso Reg. (0.8103), RR (0.8103), and LR (0.8105) achieved moderate performance, while ENR (0.7790) performed poorly, showing limited ability to capture the complex patterns in agricultural data. In contrast, non-linear and ensemble-based models performed significantly better. The KNR (0.9827) achieved the highest R2_SCORE, indicating excellent predictive accuracy. Similarly, RFR (0.9804), GBR (0.9766), CBR (0.9761), DTR (0.9711), LGBMR (0.9375), and XGBR (0.9232) all produced strong results, highlighting their effectiveness in modelling the non-linear interactions among agricultural features. ABR (0.8942) also performed well, though slightly below other ensemble approaches. Overall, the emphasizes that non-linear and ensemble-based models provide superior performance compared to traditional linear methods, making them more suitable for robust and accurate CYP (Figure 6).

Correlation heat map

The correlation matrix, which illustrates the relationships among various factors within the dataset (Figure 7).



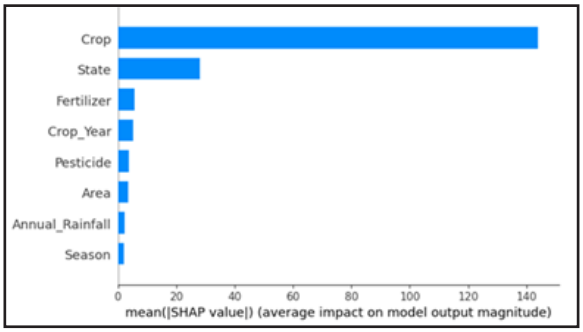
Source: Authors´ own analysis

Figure 7: Correlation heat map.

The analysis of the correlation heat map provides significant insights into how different features within the dataset are related. A key finding is the existence of strong positive correlations, where some features (Area, Pesticide, and Fertilizer) have correlation scores close to 1. If any climate condition is one feature increases, the other features also increase, then similarly any one feature also decreases, the other feature also decrease. The strong positive correlations are recommended for direct and regularly predictable relationship between these features. On the contrary, almost all features in the dataset have very low and negative correlations; Most features are around zero or below the correlation value. Weak correlations indicate that features are not related in a linear fashion. Changes in one aspect do not always imply changes in another aspect.

SHAP, LIME, ELI5

The SHAP feature importance for a regression model (Figure 8).



Source: Authors´ own analysis

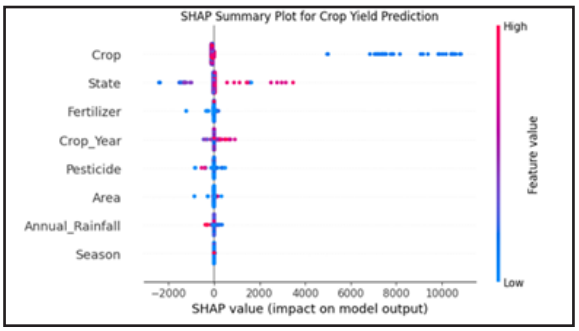
Figure 8: SHAP feature importance plot for crop yield prediction model.

We're looking at a plot where the average impact of each feature on the model's output is on the x-axis, and the input features themselves are listed along the y-axis. From the chart, it is evident that crop has the highest impact on the model's predictions by a significant margin, with a SHAP value well above 140, signifying that is the key factor in ascertaining the target variable. The next most impactful features are State, Fertilizer, and Crop_Year, all of which have moderate contributions to the model's output. Other features such as State, Season, Annual_ Rainfall, and Area exhibit minimal influence on the model's prediction, as indicated by their comparatively low SHAP values. This indicates that although they are included in the dataset, their individual impact on the model's predictions is relatively limited. Overall, the chart highlights that the model heavily relies on Crop related attributes, with land State and input factors like Fertilizer and Crop_Year also playing notable roles in prediction accuracy, whereas temporal or categorical variables have a lesser effect.

The SHAP summary plot for crop yield prediction (Figure 9). This graph illustrates the influence of various features on the model's results, positioning the features on the vertical axis and their SHAP values on the horizontal axis. The features are ranked from top to bottom based on their importance, with Crop showing the highest impact, followed by Area, Fertilizer, Pesticide,

State, Season, Annual_Rainfall, and Crop_Year.

Each point on the plot corresponds to an observation in the dataset, and its color shows the feature's value: red for high values and blue for low values.

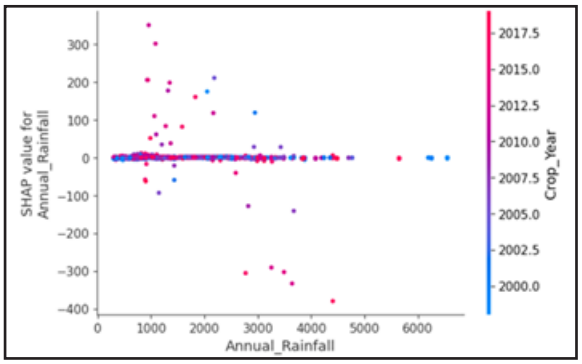


Source: Authors´ own analysis

Figure 9: SHAP summary plot for Crop Yield Prediction.

The position along the x-axis shows how much that particular observation's feature affected the prediction. Points further to the right indicate stronger positive impact on yield prediction, while points to the left show negative impact. For production, we can see many high-value (red) points extending far to the right, suggesting that higher production levels strongly correlate with higher predicted yields. The Area feature shows a more mixed pattern with both positive and negative effects. Features toward the bottom of the chart, like Season and Annual_Rainfall, have relatively smaller impacts on the model's predictions compared to the top features.

The SHAP dependence plot for Annual_Rainfall (Figure 10) reveals a non-linear, threshold-based relationship between rainfall and crop yield prediction.

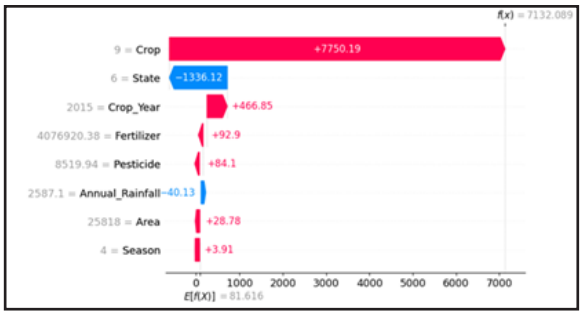


Source: Authors´ own analysis

Figure 10: SHAP dependence plot for Annual_Rainfall vs Crop_Year.

Moderate rainfall (500–1500 mm) exerts the strongest positive influence on predicted yield, with SHAP values reaching up to +350, consistent with the optimal water requirements of major Indian food crops. Beyond 1500 mm, the positive effect diminishes progressively, transitioning to strongly negative contributions (as low as −400) for rainfall exceeding 3000 mm — reflecting the yield-suppressing effects of waterlogging, flooding, and nutrient leaching under excessive precipitation. The Crop_Year colour interaction further reveals that recent years (2015–2017) exhibit more consistently positive rainfall effects, suggesting that improvements in agricultural technology and infrastructure have enhanced the crop system's capacity to convert adequate rainfall into higher yields. These findings underscore Annual_Rainfall as the most influential environmental driver of crop yield in the proposed KNR-based framework, while simultaneously highlighting the importance of managing excess water through drainage and irrigation control in high-rainfall Indian agricultural regions.

The detailed SHAP waterfall plot explaining the contribution of each feature to a specific crop yield prediction (Figure 11). The model starts with a base expected value E[f(x)] of 81.616 and arrives at a final prediction f(x) of 7132.089. Each feature's contribution is visualized as a bar with its specific impact value.



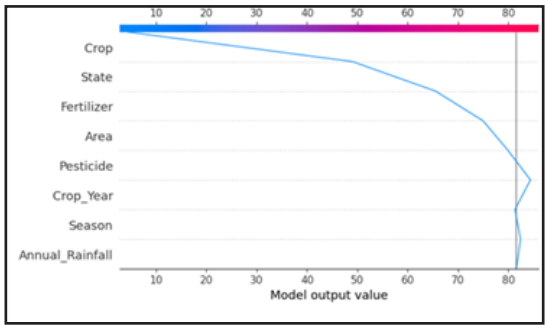
Source: Authors´ own analysis

Figure 11: Waterfall model for SHAP.

Crop (9) has the largest positive impact, contributing +7750.19 to the prediction. Other significant positive contributors including Crop_Year (2015) adding +466.85, Fertilizer (4076920) adding +92.9, and Pesticide (8519.94) adding +84.1. State (6) has a negative impact of -1336.12 and Annual_Rainfall (2587.1) has a negative impact of -40.13 shown in blue. The remaining features have smaller positive impacts: Area (25818) adds +28.78, and Season (4) contributes +3.91. This visualization effectively demonstrates how each agricultural and environmental factor contributes to the final yield prediction, with Production being overwhelmingly the most influential positive

factor, while State characteristics slightly reduce the predicted yield. The model transforms a base expectation of about 82 units to a much higher prediction of over 7,000 units based primarily on these feature values.

The SHAP decision plot (Figure 12) traces the cumulative feature contribution pathway for a selected high-yield test instance.

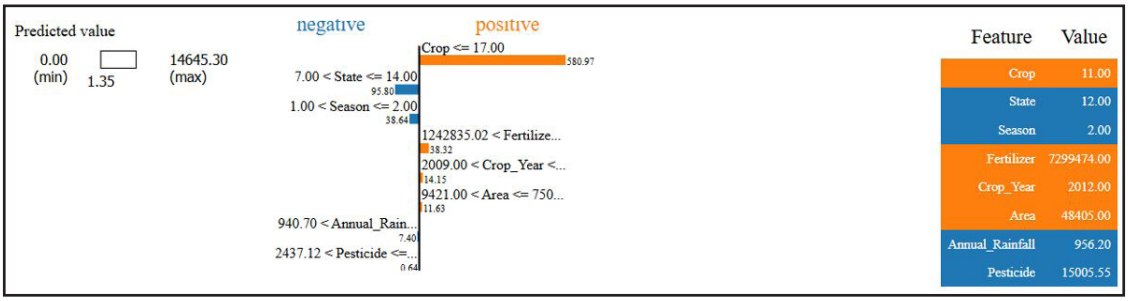


Source: Authors´ own analysis

Figure 12: SHAP summary plot – global feature importance for crop yield prediction.

It illustrates how the KNR model progresses from a baseline mean output of approximately 10 kg/ha to a final prediction of ~83–84 kg/ha through sequential, interpretable feature adjustments. Crop type emerges as the dominant driver, contributing approximately +55 units the largest single shift confirming that crop variety selection is the primary determinant of yield for this instance. State and Fertilizer provide secondary positive contributions of +8 and +5–6 units respectively, reflecting the joint influence of geographic agro-climatic context and nutrient management on productivity. Season and Annual_Rainfall contribute minimally for this instance, indicating their values fall close to the dataset mean for these features. The smooth, predominantly rightward trajectory of the decision path transitioning from blue to pink across the colour gradient visually confirms that this is a high-yield scenario where all primary features align favourably, and no significant negative corrections are introduced. This instance-level explanation validates the agronomic coherence of the KNR model and demonstrates the practical utility of SHAP decision plots in providing transparent, step-by-step justification of individual crop yield predictions for farmer-level and policy-level decision support.

The LIME local explanation for the selected test instance (Figure 13) reveals that the KNR model's prediction of 1.35 kg/ha a near-minimum yield esults from a net balance of strongly competing feature influences. Crop type exerts the dominant positive influence (+580.97), confirming that crop variety selection is the most critical agronomic determinant at the instance level. However, this positive force is substantially counteracted by the geographic state (−95.80) and seasonal context (−38.64), which together impose structural yield penalties that reflect regional productivity constraints and seasonal growing condition limitations. Fertilizer application contributes a moderate positive effect (+38.32), while Crop_Year (+14.15) and Area (+11.63) provide smaller supportive contributions. Annual_Rainfall and Pesticide contribute marginally (−7.40 and −0.64 respectively), indicating that their values for this instance are near the model's learned neutral threshold. The LIME explanation confirms that the model's local decision logic is agronomically coherent and practically interpretable: to improve yield for instances of this type, interventions should prioritise crop variety optimisation and targeted state-level agricultural support — as these two factors collectively determine the largest share of the predicted yield outcome. The absence of 'Production' from all feature contributions further validates the data leakage correction, confirming that the model's explanations are grounded exclusively in genuine agronomic and environmental relationships.



Source: Authors´ own analysis

Figure 13: LIME explanation for local feature importance in crop yield prediction.

ELI5 presents the feature importance weights and their interpretations in the context of a crop yield prediction model (Table 4). The values represent the average contribution of each feature, with associated uncertainties, indicating how influential each variable is in determining the yield outcome. Crop is by far the strongest predictor of yield, with a weight nearly 5.5 times greater than the next feature, reflecting the dominant role of intrinsic biological differences between crops. Geographic state ranks second, capturing the combined influence of climate, soil, irrigation infrastructure, and regional farming practices. Among agronomic inputs, fertilizer contributes the most stable and meaningful signal, followed by cultivated area and pesticide usage, which together reflect farm-level management intensity. Seasonal variation and annual rainfall show only marginal effects, the latter likely because irrigated systems reduce dependence on natural precipitation. Crop year carries the least reliable signal, as its standard deviation exceeds its mean weight, indicating high instability and limited predictive value.

Comparative performance analysis of models for production prediction

- features (X): crop, crop year, season, state, annual rainfall, area, fertilizer, pesticides.
- Target (y): Production

The comparison of various machine learning regression models used for crop production prediction, evaluated using four key performance metrics: MAE, MSE, RMSE, and R2_Score (Table 5 below). Together, these metrics offer a comprehensive assessment of each model's accuracy and reliability. Among all the models, the RFR achieved the best overall performance, with the lowest MAE (1432742.01), MSE (6.235554e+14), RMSE (24971091.98), along with a high R2_Score of 0.9918, indicating excellent predictive capability. Other high-performing models include GBR, which also yield low errors (MAE: 1400660.38, MSE: 6.383589e+14, RMSE: 25265765.27) and a strong R2_Score of 0.9916. KNR, DTR, XGBR, and CBR also show strong performance with R2_Scores above 0.98, reflecting their effectiveness in capturing complex patterns in the data. In contrast, the linear models – LR, RR, and Lasso Reg. – have relatively high MAE (ranging from 34745237.46 to 37558141.57) and RMSE values (211669841.00 to 212629615.82), indicating that they are less effective for this task due to their limitations in handling non-linear relationships. Elastic Net Regression shows the worst performance across all metrics, with the highest MAE (32678725.98), MSE (4.646412e+16), RMSE (215555367.05), and a low R2_Score of 0.3888, suggesting that it is not suitable for accurate crop yield prediction

Weight	Feature	Interpretation
1.8969 ± 0.2518	Crop	Crop type is overwhelmingly the most dominant predictor of yield. Its weight is nearly5.5x greater than the next feature, indicating that the intrinsic biological characteristics of each crop.
0.3423 ± 0.1417	State	Geographic region is the second most influential feature, capturing the combined effect of agro-climatic conditions, soil type, irrigation infrastructure, local agricultural policy, and farming practices that vary substantially across states.
0.0367 ± 0.0203	Fertilizer	Fertilizer application contributes a moderate and stable signal. Among the agronomic input variables, it ranks highest, consistent with the agronomic principle that nutrient availability particularly nitrogen, phosphorus, and potassium is a primary limiting factor for crop productivity.
0.0184 ± 0.0117	Area	Cultivated area exerts a notable though secondary influence on yield prediction. Larger cultivated areas may reflect greater mechanisation, access to irrigation, or economy-of-scale advantages, but can also introduce yield dilution effects if marginal land is included.
0.0130 ± 0.0095	Pesticide	Pesticide usage contributes a small but measurable effect, reflecting its role in mitigating yield losses due to pest infestations, fungal diseases, and weed competition.
0.0125 ± 0.0030	Season	Seasonal variations slightly impact yield, suggesting some seasonal dependency, but it is not a dominant factor.
0.0108 ± 0.0072	Annual_Rainfall	Annual rainfall has a marginal direct effect on yield prediction, which may appear counterintuitive but is plausible in datasets dominated by irrigated agricultural systems where supplemental water supply decouples yield form natural precipitation.
0.0085 ± 0.0135	Crop_Year	The year of cultivation exerts minimal influence on yield, but its SD (±0.0135) exceeds its mean weight (0.0085), indicating high instability and unreliable signal.

Source: Authors´ own work

Table 4: feature importance and interpretation in crop yield prediction-ELI5.

Cross Validation	ML Models	Performance Metrics			
		MAE	MSE	RMSE	R2_Score
K-Fold=5, GridSearch CV	Random Forest Regression	1432742.01	6.24E+14	24971091.98	0.9918
	Gradient Boosting Regression	1400660.38	6.38E+14	25265765.27	0.9916
	K-Neighbor Regression	1637409.01	8.10E+14	28465197.09	0.9893
	Decision Tree Regression	1846917.51	8.63E+14	29372158.02	0.9887
	XGBoost Regression	1648909.12	9.44E+14	30724896.24	0.9876
	CatBoost Regression	2442648.76	1.27E+15	35576511.38	0.9834
	LGBM Regression	2986105.2	1.67E+15	40822064.44	0.9781
	AdaBoost Regression	3039801.11	1.76E+15	41899915.66	0.9769
	Lasso Regression	37557266.6	4.48E+16	211669932.5	0.4107
	Linear Regression	37558141.57	4.48E+16	211669841	0.4107
	Ridge Regression	34745237.46	4.52E+16	212629615.8	0.4053
	ElasticNet Regression	32678725.98	4.65E+16	215555367.1	0.3888

Source: Authors´ own work

Table 5: Comparison of ML model performance in agricultural forecasting with production as target.

Author Name	Algorithms	Best Model	Input Parameters	Evaluation Results
Shams et al. (2024)	XAI_CROP, GB, DT, RF, GNB, MNB	XAI-CROP	District_Name, Season, Area, Production, Crop	R2_Score=0.94
Sathya and Gnanasekaran (2023)	MLR-LSTM, MLR, SVM, LSTM, RF	MLR-LSTM	pH range, rainfall, maxtemp, mintemp, block, Windspeed	R2_Score=0.93
Malashin et al. (2024)	DNN, GA, LIME	DNN	Crop, Season, Area, State, Annual_ Rainfall, Pesticide, Fertilizer, Production, Yield	R2_Score=0.92
Mohan et al. (2024)	DTR, RFR, LGBMR, SHAP, LIME	LGBMR	Rainfall, Temperature, Fertilizer, Phosphorus, Nitrogen, Potassium	R2_Score=0.92
Proposed	LR, RR, LassoR, DTR, RFR, LGBMR, CBR, ABR, ENR, GBR, XBR, KNR, SHAP, LIME, ELI5	KNR	Crop, Crop_Year, Season, Area, State, Annual_Rainfall, Pesticide, Fertilizer, Yield	R2_Score=0.94

Source: Authors´ own work

Table 6: Crop Yield Prediction comparing with various State-of-the-Art (SOTA) Models.

in this context. ABR shows the mid performance of MAE (3039801.11), MSE (1.755603e+15), RMSE (41899915.66), and R2_Score (0.9769). LGBM fall in the middle tier. LGBM has decent MAE (2986105.20) and a respectable R2_Score (0.9781), is slightly less accurate with higher errors but still better than linear models.

Comparison with the state-of-the-art models

The proposed work compared with the state-of-the-art model (SOTA) as shown in below (Table 6 above).

The comparative analysis of various crop yield prediction models based on the algorithms used, the input parameters considered, and the corresponding evaluation results in terms of R2_Score (Table 6). References (Shams et al., 2024; Sathya and Gnanasekaran, 2023; Malashin et al., 2024; Jagan Mohan et al., 2024) represent existing studies, each employing a combination of traditional ML, DL, and XAI techniques. Input parameters vary across models, including geographic attributes, environmental conditions, soil characteristics, and crop-specific data. The proposed model, in contrast, integrates a broader set of ensemble and regression algorithms, including LR, RR, Las. Reg., DTR, RFR, LGBMR, CBR, ABR, ENR, GBR, XBR, KNR, and advanced explainability tools such as SHAP, LIME, and ELI5. It uses a rich set of input features covering crop details, temporal factors, environmental inputs, and geographical identifiers. The proposed model achieves the highest R2_Score of 0.94, demonstrating superior performance and accuracy in crop yield prediction compared to previously reported models.

Conclusion

The present study successfully demonstrates that integrating Explainable Artificial Intelligence (XAI) with advanced machine learning techniques significantly enhances both the accuracy and transparency of crop yield prediction systems. By evaluating a wide range of regression algorithms, the research establishes a comprehensive comparative framework, ensuring that model selection is not only performance-driven but also analytically justified. Among all the models, the K-Nearest Neighbour Regression (KNR) model emerged as the most effective, achieving a high R2_Score of 0.9827 along with low error metrics (MAE: 9.03, MSE: 13824.73, RMSE: 117.58). These results confirm the robustness and predictive strength of the proposed approach in handling complex agricultural datasets.

A key contribution of this work lies in its emphasis on interpretability through the integration of XAI techniques like SHAP, LIME, and ELI5. Unlike traditional "black-box" models, the proposed framework provides clear insights into how different features influence crop yield predictions. This transparency is crucial in agriculture, where decision-making often relies on trust and understanding. By identifying the most influential factors affecting yield, the system empowers farmers, agronomists, and policymakers to make informed, data-driven decisions that can improve productivity and reduce uncertainty.

Furthermore, the use of preprocessing techniques such as One-Hot Encoding for categorical variables, along with 5-Fold Cross-Validation and Grid Search optimization, ensures the reliability and generalizability of the models. These methodological choices reduce overfitting and enhance the stability of predictions across different data subsets. The study, therefore, not only focuses on predictive performance but also maintains strong experimental rigor, making the findings more credible and applicable in real-world agricultural scenarios.

The integration of KNR with SHAP-based explanations stands out as a particularly effective combination, offering both high predictive accuracy and meaningful interpretability. This dual advantage addresses a major challenge in modern AI applications—balancing performance with explainability. The ability to visualize feature contributions enables stakeholders to understand the reasoning behind predictions, fostering greater adoption of AI-driven systems in agriculture. As a result, the proposed model supports sustainable farming practices by enabling optimized resource utilization and improved crop management strategies.

Further the results confirm that ensemble-based models, particularly Random Forest Regression, are highly effective for accurate crop production prediction and significantly outperform traditional linear regression methods.

In conclusion, this research contributes significantly to the field of precision agriculture by bridging the gap between high-performance machine learning models and the need for transparency. It sets a strong foundation for the development of intelligent, explainable, and reliable agricultural decision-support systems. Future work can further enhance this framework by incorporating region-specific models and integrating dynamic environmental factors such as rainfall, temperature, and soil conditions. Such advancements will improve the adaptability and scalability of the system, ultimately promoting sustainable and resilient agricultural practices in diverse farming environments.

Acknowledgements

The authors express their sincere to Vellore Institute of Technology (VIT) for their assistance with this research.

Corresponding author:
Dr. Kumaresan, P.
Centre For Artificial Intelligence Research(C-FAIR)
Vellore Institute of Technology, Vellore-632014, Tamilnadu, India
Phone: +91-9842582301, E-mail: pkumaresan@vit.ac.in

References

[1] Abbas, F., Afzaal, H., Farooque, A. A. and Tang, S. (2020) "Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms", *Agronomy*, Vol. 10, No. 7, p. 1046. ISSN 2073-4395. DOI 10.3390/agronomy10071046.

[2] Abdel-Salam, M., Kumar, N. and Mahajan, S. (2024) "A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning", *Neural Computing and Applications*, Vol. 36, No. 33, pp. 20723-20750. ISSN 1433-3058. DOI 10.1007/s00521-024-10226-x.

[3] Abekoon, T., Hirushan, S., Namal, R., Imesh, U. E., Anuradha, J. and Upaka, R. (2025) "A Novel Application with Explainable Machine Learning (SHAP and LIME) to Predict Soil N, P and K Nutrient Content in Cabbage Cultivation", *Smart Agricultural Technology,* Vol. 11 (March), p. 100879. ISSN 2772-3755. DOI 10.1016/j.atech.2025.100879.

[4] Abhasha, J., Pradhan, B., Chakraborty, S., Varatharajoo, R., Alamri, A., Gite, S. and Lee, C.-L. (2025) "An Explainable Bi-LSTM Model for Winter Wheat Yield Prediction", *Frontier in Plant Science*, Vol. 15, pp. 1-17. ISSN 1664-462X. DOI 10.3389/fpls.2024.1491493.

[5] Abasha, K., Pradhan, B., Gite, S. and Chakraborty, S. (2023) "Remote-Sensing Data and Deep-Learning Techniques in Crop Mapping and Yield Prediction: A Systematic Review", *Remote Sensing*, Vol. 15, No. 8, p. 2014. ISSN 2072-4292. DOI 10.3390/rs15082014.

[6] Agarwal, N., Choudhry, N. and Tripathi, K. C. (2025) "A novel hybrid time series deep learning model for forecasting of cotton yield in India", *International Journal of Information Technology (Singapore)*, pp. 1745-1752. ISSN 2511-2104. DOI 10.1007/s41870-024-02327-6.

[7] Akkem, Y., Biswas, S. K. and Varanasi, A. (2024) "Streamlit-based enhancing crop recommendation systems with advanced explainable artificial intelligence for smart farming", Neural Computing and Applications, Vol. 36, No. 32, pp. 20011-20025. ISSN 1433-3058. DOI 10.1007/s00521-024-10208-z.

[8] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R and Herrera, F. (2020) "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, Vol. 58, pp. 82-115. ISSN 1566-2535. DOI 10.1016/j.inffus.2019.12.012.

[9] Bhattacharya, A. (2022) "*Applied Machine Learning Explainability Techniques*", Birmingham, UK: Packt Publishing Ltd., 1st ed.. ISBN 978-1803246154.

[10] Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F. and Cho, Y. I. (2024) "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms", *Mathematics*, Vol. 12, No. 16. ISSN 2227-7390. DOI 10.3390/math12162553.

[11] Burdett, H. and Wellen, C. (2022) "Statistical and Machine Learning Methods for Crop Yield Prediction in the Context of Precision Agriculture", *Precision Agriculture,* Vol. 23, No. 5, p. 1553-74. ISSN 1385-2256. DOI 10.1007/s11119-022-09897-0.

[12] de Villiers, C., Mashaba-Munghemezulu, Z., Munghemezulu, C., Chirima, G. J. and Tesfamichael, S. G. (2024) "Assessing Maize Yield Spatiotemporal Variability Using Unmanned Aerial Vehicles and Machine Learning", *Geomatics*, Vol. 4, No. 3, pp. 213-236. ISSN 2673-7418. DOI 10.3390/geomatics4030012.

[13] Dilli, P., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C. and Athanasiadis, I. N. (2021) "Machine Learning for Large-Scale Crop Yield Forecasting", *Agricultural Systems,* Vol. 187, (June 2020), p. 103016. ISSN 0308-521X. DOI 10.1016/j.agsy.2020.103016.

[14] Fegade, T. K. and Pawar, B. V. (2020) "Crop prediction using artificial neural network and support vector machine", In *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019*, Springer Singapore, Vol. 2, pp. 311-324. ISBN 978-981-13-9363-1.

[15] Jagan Mohan, R. N. V., Rayanoothala, P. S. and Sree, R. P. (2024) "Next-Gen Agriculture: Integrating AI and XAI for Precision Crop Yield Predictions", *Frontiers in Plant Science*, Vol. 15, pp. 1-16. ISSN 1664-462X. DOI 10.3389/fpls.2024.1451607.

[16] Khaki, S. and Wang, L. (2019) "Crop yield prediction using deep neural networks", *Frontiers in Plant Science*, Vol. 10, pp. 1-10. ISSN 1664-462X. DOI 10.3389/fpls.2019.00621.

[17] Khan, M. and Noor, S. (2019) "Irrigation runoff volume prediction using machine learning algorithms", *European Journal of Science, Innovation and Technology*, Vol. 8, pp. 1-22. ISSN 2304-9693.

[18] Khosla, E., Dharavath, R. and Priya, R. (2020) "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression", *Environment, Development and Sustainability*, Vol. 22, No. 6, pp. 5687-5708. ISSN 1387-585X. DOI 10.1007/s10668-019-00445-x.

[19] Kuruguntu Mohan, K., Maheswari, N. and Sivagami, M. (2022) "Models for Feature Selection and Efficient Crop Yield Prediction in the Groundnut Production", *Research in Agricultural Engineering,* Vol.68, No. 3, pp. 131-41. ISSN 1212-9151. DOI 10.17221/15/2021-RAE.

[20] LeCun, Y., Bengio, Y. and Hinton, G. (2015) "Deep learning", *Nature*, Vol. 521, pp. 436-444. ISSN 0028-0836. DOI 10.1038/nature14539.

[21] Liaw, A. and Wiener, M. (2002) "Classification and regression by randomForest", *R News*, Vol. 2, No. 3, pp. 18-22. ISSN 1609-3631.

[22] Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., Borodulin, A. and Tynchenko, Y. (2024) "Predicting Sustainable Crop Yields: Deep Learning and Explainable AI Tools", *Sustainability (Switzerland),* Vol. 16, No. 21, pp. 1-29. ISSN 2071-1050. DOI 10.3390/su16219437.

[23] Maya Gopal, P. S. and Bhargavi, R. (2019) "Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms", *Applied Artificial Intelligence*, Vol. 33, No. 7, pp. 621-642. ISSN 1087-6545. DOI 10.1080/08839514.2019.1592343.

[24] Mishra, P., Khan, R. and Baranidharan, D. B. (2020) "Crop Yield Prediction using Gradient Boosting Regression", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 9, No. 3, pp. 2293-2297. ISSN 2278-3075. DOI 10.35940/ijitee.c8879.019320.

[25] Mitra, A., Beegum, S., Fleisher, D., Reddy, V. R., Sun, W., Ray, C., Timlin, D. and Malakar, A. (2024) "Cotton Yield Prediction: A Machine Learning Approach with Field and Synthetic Data", *IEEE Access*, Vol. 12, p. 101273-88. ISSN 2169-3536. DOI 10.1109/ACCESS.2024.3418139.

[26] Molnar, C. (2020) "*Interpretable machine learning*", Lulu.com. ISBN 978-0244768522. [Online]. Available: https://christophm.github.io/interpretable-ml-book. [Accessed: 10 Jan. 10, 2022].

[27] Nguyen, N. and Ngo, D. (2025) "Comparative Analysis of Boosting Algorithms for Predicting Personal Default", *Cogent Economics and Finance,* Vol. 13, No. 1. ISSN 2332 - 2039. DOI 10.1080/23322039.2025.2465971.

[28] Oikonomidis, A., Cagatay Catal, C. and Kassahun, A. (2022) "Hybrid Deep Learning-Based Models for Crop Yield Prediction", *Applied Artificial Intelligence*, Vol. 36, No. 1. ISSN 1087-6545. DOI 10.1080/08839514.2022.2031823.

[29] Pavithra, M. and Soundrapandiyan, R. 2024. "Yield Prediction for Crops by Gradient-Based Algorithms", *PLoS ONE,* Vol. 19, No. 8, pp. 1-20. ISSN 1932-6203. DOI 10.1371/journal.pone.0291928.

[30] Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S. and Stodtmann, S. (2024) "Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development", *Clinical and Translational Science*, Vol. 17, No. 11, pp. 1-15. ISSN 1752 - 8062. DOI 10.1111/cts.70056.

[31] Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) ""Why Should I Trust You?": Explaining the Predictions of Any Classifier", In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* ACM. pp. 1135-1144. ISSN 2157-0161. DOI 10.1145/2939672.2939778.

[32] Saadio, C. D., Adoni, W. Y. H., Aworka, R., Zoueu, J. T., Mutombo, F. K., Krichen, M. and Kimpolo, C. L. M. (2022) "Crops Yield Prediction Based on Machine Learning Models: Case of West African Countries", *Smart Agricultural Technology,* Vol. 2 (December 2021). ISSN 2772-3755. DOI 10.1016/j.atech.2022.100049.

[33] Samira, S. S., Shahhosseini, M., Huber, I., Hu, G. and Archontoulis, S. V. (2022) "County-Scale Crop Yield Prediction by Integrating Crop Simulation with Machine Learning Models", *Frontiers in Plant Science*, Vol. 13, pp. 1-16. ISSN 1664-462X. DOI 10.3389/fpls.2022.1000224.

[34] Sarr, A. B. and Sultan, B. (2023) "Predicting crop yields in Senegal using machine learning methods", *International Journal of Climatology*, Vol. 43, No. 4, pp. 1817-1838. ISSN 0899-8418. DOI 10.1002/joc.7947.

[35] Sathya, P. and Gnanasekaran, P. (2023) "Paddy Yield Prediction in Tamilnadu Delta Region Using MLR-LSTM Model", *Applied Artificial Intelligence*, Vol. 37, No. 1. ISSN 1087-6545. DOI 10.1080/08839514.2023.2175113.

[36] Shams, M. Y., Gamel, S. A. and Talaat, F. M. (2024). "Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making", *Neural Computing and Applications*, Vol. 36, No. 11, pp. 5695-5714. ISSN 1433-3058. DOI 10.1007/s00521-023-09391-2.

[37] Sharma, P., Dadheech, P., Aneja, N. and Aneja, S. (2023) "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning", *IEEE Access*, Vol. 11, No. October, pp. 111255-111264. ISSN 2169-3536. DOI 10.1109/ACCESS.2023.3321861.

[38] Sudha, M. K., Manorama, M. and Aditi, T. (2022) "Smart agricultural decision support systems for predicting soil nutrition value using IoT and ridge regression", *AGRIS on-line Papers in Economics and Informatics*, Vol. 14, No. 1, pp. 95-106. ISSN 1804-1930. DOI 10.7160/aol.2022.140108.

[39] Uribeetxebarria, A., Castellón, A. and Aizpurua, A. (2023) "Optimizing Wheat Yield Prediction Integrating Data from Sentinel-1 and Sentinel-2 with CatBoost Algorithm", *Remote Sensing*, Vol. 15, No. 6. ISSN 2072-4292. DOI 10.3390/rs15061640.

[40] Vijayan, S. and Chowdhary, C. L. (2025) "Hybrid feature optimized CNN for rice crop disease prediction", *Scientific Reports*, Vol. 15, No. 1, pp. 1-20. ISSN 2045-2322. DOI 10.1038/s41598-025-92646-w.