Smart Agricultural Technology 13 (2026) 101904

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

# Farm-level yield prediction for maize, rice, and beans in Tanzania using machine learning and multi-source agricultural data

Bally S. Omary [a,b,*], Mussa A. Dida [a], Devotha G. Nyambo [a]

[a] School of Computational and Communication Science and Engineering (CoCSE), Nelson Mandela African Institution of Science and Technology (NM-AIST), P.O. Box 447, Arusha, Tanzania
[b] Department of Computer Science and Engineering, Mbeya University of Science and Technology (MUST), P.O. Box 131, Mbeya, Tanzania

ARTICLE INFO

*Keywords:*
Crop productivity
Machine learning
Smallholder farmers
Multi-crop
Climate change

ABSTRACT

Accurate predictions of crop yields at the farm level are essential for improving agricultural productivity, enhancing food security, and supporting informed decision-making among smallholder farmers. However, conventional field assessments and simple statistical models are often time-consuming, limited in scope, and unable to capture complex interactions among climatic and soil factors. To address these challenges, this paper proposes a machine learning-based model for predicting the productivity of multiple crops, including maize, rice, and beans, using multi-source farm-level data from Tanzania. The dataset integrates climate variables such as temperature and rainfall, soil type, farm size, and crop type. Four ensemble learning models, namely Random Forest, Gradient Boosting, Extreme Gradient Boosting, and Extra Trees, were evaluated using an 80/20 train–test split on 9,897 farm-level records acquired from the Mbeya, Ruvuma, and Songwe regions between 2022 and 2024. Hyperparameter tuning with a fivefold cross-validation was applied to improve model generalization and reduce overfitting. Among the evaluated models, the Extra Trees ensemble achieved the highest performance, with a pooled multi-crop $R^2$ of 95%, while crop-specific $R^2$ values ranged from 79% to 81% for maize, rice, and beans. These findings demonstrate the potential of the proposed approach to support farm-level cultivation planning and climate adaptation decisions for smallholder farmers.

## 1. Introduction

Crop productivity is the measurable output of a crop per unit area and serves as a key indicator of agricultural efficiency and overall farm performance. However, climate change continues to undermine crop productivity, particularly among smallholder farmers in developing countries [1]. Since 1961, crop productivity has declined by about 21% around the world, with the biggest drops happening in warmer tropical areas [2]. Climate change has also significantly affected specific staple crops, causing productivity declines of 7.4% for maize, 3.2% for rice, and 3.1% for beans. Looking ahead, the Intergovernmental Panel on Climate Change warns that, without effective adaptation strategies, global crop productivity could decline by an additional 5–25% by mid-century [3].

Africa remains highly vulnerable to declining crop productivity because over 60% of its population depends on rain-fed agriculture, while adaptive capacity and technological integration remain limited [4]. Overall, crop productivity on the continent is expected to decline by about 20%. Maize, rice, and beans are some of the most affected staples [5]. The African Development Bank's regional assessment reports that these losses in productivity could cost East Africa's economy more than 1% of its GDP each year by 2050 [6]. These kinds of drops put not only food security at risk but also the jobs and income stability of millions of smallholder farmers.

Evidence indicates a comparable decline in Tanzania. Maize productivity has already decreased by 10–20% across several agro-ecological zones [7]. Projections indicate further declines, with maize productivity expected to decline by approximately 13% and rice by 7.6% by 2050 [8]. Local studies reinforce this pattern, reporting that yields of major food crops, including maize, rice, and beans, have fallen by 12–51% in traditional farming systems over the past four decades [9]. These findings highlight the magnitude of Tanzania's crop productivity challenge, particularly given that agriculture contributes 26% of the national GDP and employs more than 65% of the population [10].

The integration of artificial intelligence (AI), particularly machine learning techniques, presents a promising opportunity to improve

* Corresponding author at: Nelson Mandela African Institute of Science and Technology, Tanzania
*E-mail address:* ballyo@nm-aist.ac.tz (B.S. Omary).

https://doi.org/10.1016/j.atech.2026.101904
Received 26 October 2025; Received in revised form 18 February 2026; Accepted 18 February 2026
Available online 19 February 2026

prediction accuracy, operational efficiency, and the accessibility of agricultural insights [11]. Convolutional Neural Networks (CNNs) have been successfully applied in several contexts, such as predicting strawberry yield using Faster R-CNN [12], estimating maize yield from multispectral imagery [13], and forecasting rice yield based on Sentinel-2 satellite data [14]. Beyond CNNs, models such as artificial neural networks, support vector machines, random forests, and AdaBoost have demonstrated strong performance in predicting county-level yields of crops including beans, potatoes, barley, and wheat [15].

Despite these advancements, existing solutions remain centered on single-crop, aggregated predictions at district, regional, or national scales, which fail to capture the variability and complexity of individual farm conditions. As a result, there is a growing need for predictive solutions that operate at the farm level, generating context-specific and actionable insights. This study responds to that need by developing a machine learning–based model capable of predicting the yields of multiple crops at the individual farm scale in Tanzania.

The following research questions guide this study accordingly:

RQ1: How accurately can machine learning models predict farm-level yields of maize, rice, and beans in Tanzania using temperature, rainfall, soil type, farm size, and crop type?

RQ2: Which machine learning algorithm yields the most accurate predictions for various crops?

RQ3: What is the relative importance of key features in influencing farm-level yield predictions?

RQ4: How well do the machine learning models generalize when validated using unseen farm-level datasets from different regions in Tanzania?

The specific contributions of this study are:

- Compilation of a farm-level, multi-source dataset integrating weather, soil, and farmer-reported production data for maize, rice, and beans in Tanzania.
- Application of machine learning–based ensemble models to predict yields of multiple crops (maize, rice, and beans) at the farm level.
- Comparative evaluation of established ensemble learning models (Random Forest, Gradient Boosting, Extreme Gradient Boosting, and Extra Trees) to assess their predictive performance and computational efficiency under Tanzanian climatic conditions.

The remainder of this paper is organized as follows: Section 2 reviews related work and identifies existing research gaps. Section 3 presents the material and methods, covering the study area, data acquisition, data exploration and analysis, feature engineering, the workflow of the proposed method, and model development. Section 4 outlines the experimental results. Section 5 discusses the experimental results. Finally, Section 6 concludes the study by summarizing the key contributions.

## 2. Related works

Recent advances in machine-learning approaches offer promising solutions to agricultural challenges by enabling more accurate and scalable crop yield predictions [16]. Studies across diverse regions and crop types have widely adopted machine learning applications in agriculture, consistently improving prediction accuracy [17].

### 2.1. Deep learning approaches

Deep learning has played a major role in improving crop yield prediction by making it possible to combine high-dimensional spatial and temporal data. Kaneko et al. [18] demonstrated the value of multispectral satellite imagery for predicting maize yields at the district level across six African countries, achieving moderate accuracy ($R^2$ of approximately 0.56). Their histogram-based deep learning pipeline showed how rich spectral information can compensate for limited training labels. Sun et al. [19] also proposed a multilevel RNN–CNN architecture that accurately predicts corn yields at the county level across the U.S. Corn Belt from 2013 to 2016 by capturing both spatial variability and temporal growth dynamics.

Together, these studies highlight the effectiveness of deep learning in regions where large-scale remote-sensing data are available and where prediction is needed at a regional or national scale. They also provide evidence that integrating temporal and spatial signals enhances model performance. These insights inform the present study by showing the value of combining multiple feature types; however, unlike these large-scale remote-sensing applications, this study focuses on multi-crop prediction at the farm level in Tanzania, where fine-grained, on-farm data are needed rather than broad satellite-only inputs.

### 2.2. Ensemble and classical machine learning methods

Classical and ensemble machine-learning methods have been widely adopted for yield prediction, offering interpretability and strong performance across heterogeneous agricultural settings. Champaneri et al. [20] developed a web-based tool using Random Forest to provide farmers with pre-cultivation yield estimates, demonstrating the practicality of digital advisory systems. Bhoj et al. [21] integrated GPS-enabled mobile data with five algorithms and reported high predictive accuracy for Random Forest (approximately 95%), illustrating how location-aware applications can support fertilizer scheduling and planting decisions. Jhajharia and Mathur [22] compared ensemble algorithms for Rajasthan and found Random Forest to be consistently superior, while Jhajharia et al. [23] broadened the scope to seven crops using SVM, LSTM, Random Forest, Gradient Descent, and LASSO, again identifying Random Forest as the most robust. Aworka et al. [1] extended ensemble modeling to national-level predictions across East Africa using climate and production data, highlighting the method's reliability across diverse agroecological zones. Bao et al. [24] proposed a maize yield prediction model integrating vegetation indices and agronomic trait data with machine learning algorithms, showing improved model accuracy when combining the leaf area index, relative chlorophyll content, and normalized difference red-edge index.

These studies collectively show the strong alignment between ensemble models and mixed-type agricultural datasets, and they highlight the increasing shift toward mobile and web-based advisory platforms. While their contexts span regional, national, and multi-crop settings, the present study builds on this foundation by adapting the modeling approach to farm-level, multi-crop prediction using local climate and farm characteristics relevant to Tanzanian production realities.

### 2.3. Research in East Africa

Recent studies in East Africa provide important regional insights into how machine learning captures relationships between climate variables and crop yield. Kuradusenge et al. [25] showed that Random Forest performs reliably when predicting maize and Irish potato yields in Rwanda using rainfall and temperature inputs. Patrick et al. [26] examined Tanzania's banana sector using correlation analysis, regression models, SARIMAX, state space models, and LSTM, demonstrating how climate factors strongly drive yield fluctuations. Tende et al. [27] combined satellite imagery with climate data using deep learning models deployed through SMS and web interfaces, offering an accessible digital prediction solution for maize at the district scale.

Across these studies, climate variables, particularly rainfall and temperature, consistently emerge as the dominant predictors, and ensemble models demonstrate strong performance relative to linear or shallow learning alternatives. These findings guide the present research by confirming the central role of climate in yield prediction within East African contexts. While previous studies focus on single crops or district-level predictions, the current study extends this work by developing a

multi-crop, farm-level model designed for Tanzanian agricultural conditions.

### 2.4. Hybrid approaches

Researchers are increasingly combining complementary algorithms with hybrid and neuro-fuzzy methods to enhance prediction performance. Bali et al. [28] reviewed over 80 studies on crop yield prediction and found strong performance in artificial neural networks, adaptive neuro-fuzzy inference systems, and hybrid models, emphasizing the importance of temperature and humidity as key drivers of productivity. Khaki et al. [29] combined CNNs, RNNs, Random Forest, feedforward networks, and LASSO to predict corn and soybean yields in the U.S. Corn Belt, attaining RMSEs of 9% and 8%, respectively.

These studies show that hybrid models are particularly useful when multiple data sources (spatial, temporal, and environmental) are jointly analyzed. Their insights are valuable for contexts requiring rich multi-source inputs, though their data-intensive nature contrasts with the more locally grounded, farm-level dataset used in the present research. This study instead adopts methods suited to mixed-type, medium-scale agricultural datasets in Tanzania while incorporating lessons from hybrid modeling regarding feature integration.

### 2.5. Synthesis and link to the present study

Across deep learning, ensemble, classical machine learning, and hybrid approaches, prior research demonstrates substantial progress in modeling crop yield. Most studies rely on climate variables as core predictors and often operate at district, regional or national scales. The reviewed evidence highlights opportunities to address farm-level needs in Tanzania by producing multi-crop, fine-scale, and decision-ready predictions. Building on the contextual insights, methodological strengths, and empirical trends from prior work, this study develops a machine learning model optimized for local agricultural realities and practical deployment in Tanzanian contexts.

## 3. Material and methods

### 3.1. Study area

This study was conducted in three regions of Tanzania's Southern Highlands (Mbeya, Ruvuma, and Songwe) (Fig. 1). These regions are among the country's leading agricultural zones, contributing substantially to national maize, rice, and bean production. They were selected because of their favorable climatic conditions, diverse soil types, and the predominance of smallholder farming systems that are highly sensitive to climatic change.

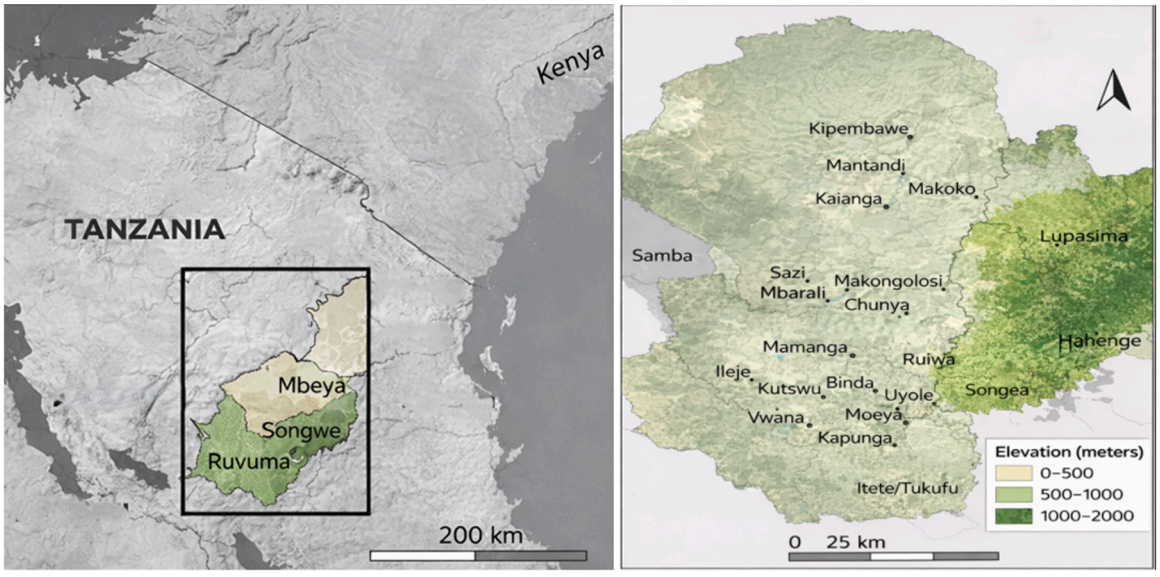
*Mbeya region*

Mbeya is one of Tanzania's thirty-one administrative regions, located in the Southern Highlands. It borders Tabora to the north, Malawi and Lake Nyasa to the south, Songwe to the west, and Njombe to the east, with Mbeya City as the regional capital. The region covers approximately 35,954 km$^2$ and has a population of over 2.7 million. Administratively, it comprises seven districts, 178 wards, 533 villages, and 181 streets [30].

Agriculture is the dominant livelihood activity, supported by clay and loam soils that are highly suitable for maize, rice, and beans, particularly in districts such as Mbarari and Mbeya Rural. Annual rainfall ranges from 800 to 1,200 mm, with the main rainy season occurring from November to April. Highland districts such as Rungwe and Kyela receive higher rainfall due to their elevation. Recent climate changes have led to increasing rainfall variability, prolonged droughts, and occasional flooding, posing challenges for smallholder farmers. In the cool, fertile highlands of Rungwe, Mbozi, and Mbeya Rural [31], Mbeya is also a major producer of coffee and tea. Despite its high agricultural potential, the region faces constraints such as limited irrigation infrastructure, low uptake of modern technologies, and post-harvest losses.

*Ruvuma region*

Ruvuma lies in Tanzania's southernmost zone, bordering Mozambique to the south, Njombe to the northwest, and Lindi to the northeast. It covers approximately 63,669 km$^2$ and has a population of about 1.5 million, with Songea as the regional capital. More than 85% of the people who live there work in agriculture, which is the main industry. The region is known for cultivating maize, beans, cassava, and coffee, supported by fertile volcanic soils and reliable rainfall averaging 1,000–1,400 mm annually. The climate is predominantly sub-humid, with a rainy season from November to May and a dry season from June to October. Ruvuma is also an important area for coffee production and livestock keeping [32].




(a) Location of the study area within Tanzania (b) Overview of the study region

**Fig. 1.** Study area and spatial context (Data sources: SRTM, HydroSHEDS, ESA land cover, and GADM).

*Songwe region*

Songwe, established in 2016 after separating from Mbeya, is located in southwestern Tanzania. It borders Zambia to the southwest, Malawi to the south, and Mbeya to the east. The region covers about 27,500 km$^2$ and has a population of more than 1 million. Vwawa serves as the regional capital. Songwe's economy is largely agrarian, dominated by smallholder farmers cultivating maize, beans, and sorghum. The region receives moderate rainfall (700–1,100 mm per year), suitable for both subsistence and commercial agriculture. Its sandy-loam soils support a variety of food and cash crops. Nonetheless, irregular rainfall, declining soil fertility, and limited access to agricultural inputs continue to constrain productivity [33].

### 3.2. Data acquisition

The dataset consists of 9,897 farm-level records collected from the Mbeya, Ruvuma, and Songwe regions between 2022 and 2024. These records represent individual farms and include three major crops commonly grown in Tanzania, with the following distribution: maize (4,621 records), rice (2,969 records), and beans (2,307 records). Each record contains six key features as indicated below.

- **Rainfall (mm):** total precipitation received during the season.
- **Temperature (°C):** average ambient temperature.
- **Farm size (ac):** cultivated area.
- **Soil type:** soil classification based on texture and structure.
- **Crop type:** a crop category grown by the farmer.
- **Production (kg):** total harvested quantity per unit area.

Data were compiled at the farm level using structured reports obtained from ward extension officers and government authorities. Rainfall and temperature data were sourced from the Tanzania Meteorological Authority (TMA) and aggregated to seasonal values corresponding to each farm's location. Farm plot sizes and soil types were taken directly from extension officers' routine monitoring records, while production data for maize, rice, and beans were sourced from farmers' production reports documented by the officers.

To ensure representativeness, a stratified random sampling technique was used, with regions serving as strata. Within each region, farm lists maintained by extension officers served as the sampling frame, enabling proportional selection of farms of varying sizes and production characteristics. Only farms with complete and consistent production records, and those actively cultivating at least one major crop, were included.

To make the data more reliable, quality control steps were taken. Reported farm sizes were validated by cross-checking with official boundary measurements maintained by the district councils. Soil properties and production data were verified using historical extension officer records to ensure internal consistency. Additional data cleaning steps included resolving duplicate entries, checking for logical inconsistencies, handling missing values, and harmonizing categorical labels. To process the large and heterogeneous dataset, we used the StreamSets Data Collector engine to implement the ETL (Extract, Transform, Load) workflow.

### 3.3. Data exploration and analysis

#### 3.3.1. Basic scatterplots and histograms

The complete dataset was examined to understand its structure, characteristics, and statistical behavior. It was identified as structured and primarily composed of both numerical and categorical variables, with no missing entries detected. Data exploration was performed on numerical variables using Seaborn and Matplotlib to assess feature distributions, linearity, variance, and skewness. Matplotlib provides a flexible, low-level 2D plotting API widely used in scientific research for producing publication-quality figures [34]. Seaborn, built on top of Matplotlib, offers high-level statistical graphics with improved aesthetics and simplified syntax, enabling quick generation of distributions, categorical plots, heatmaps, and other analytical visuals. Both libraries are open-source, well-tested, and widely used in peer-reviewed research, so they are reliable and trustworthy tools for the exploratory and presentation plots used in this study [35].

The visualization analysis revealed clear non-linear relationships between rainfall and production, as well as between temperature and production. Furthermore, the distributions of rainfall, temperature, and production were non-normal, as shown in Figs. 2 and 3. The presence of values deviating from the main clusters indicated noticeable variance within the dataset. Frequency distribution plots and histograms (Figs. 4 and 5) also revealed skewness in both rainfall and temperature variables.

Figs. 2 and 3 illustrate non-linear relationships between rainfall and temperature versus production, while Figs. 4 and 5 show skewed, non-normal feature distributions. These patterns suggested that linear models would struggle to capture the underlying variability. Ensemble algorithms such as Random Forest, Gradient Boosting, Extreme Gradient Boosting, and Extra Trees are more appropriate, as they can model non-linearity, handle skewed inputs, and remain robust to outliers, providing a solid foundation for accurate yield prediction.

#### 3.3.2. Summary statistics

Summary statistics for numerical variables (rainfall, temperature, farm size, and production) were computed using descriptive measures including the mean, standard deviation, minimum, maximum, and interquartile range (Table 1). Categorical variables (soil type and crop type) were summarized using frequency distributions, indicating the number of unique classes and the most frequently observed categories (Table 2).

#### 3.3.3. Correlation matrices

To examine relationships among the study variables, a Pearson correlation analysis was conducted, as shown in Fig. 6. Correlation coefficients were calculated for all numerical variables, including rainfall, temperature, and production, as well as for categorical variables such as soil type and crop type, following one-hot encoding. This analysis helped identify patterns of association, both positive and negative, and informed subsequent exploratory and feature importance analyses.

#### 3.3.4. Outlier justification

Outlier detection was performed only on numerical variables because categorical variables do not produce statistical outliers. The interquartile range (IQR) method was used to identify unusually high or low values in rainfall, temperature, farm size, and production. Observations falling outside the 1.5 × IQR threshold were flagged as potential outliers. Outliers caused by recording errors were removed, while those representing genuine extreme climatic or production events were retained to preserve the natural variability of rain-fed agricultural systems.

#### 3.3.5. Feature importance analysis

Feature importance analysis was performed using a trained tree-based Extra Trees model. Model-derived importance scores were calculated for all input variables to quantify each predictor's contribution to overall model accuracy. To validate the robustness of these importance scores, permutation importance analysis was also conducted (Fig. 7). This procedure provided a systematic method for identifying the relative influence of individual predictors on crop yield predictions and informed subsequent interpretation of model outputs.

#### 3.3.6. Agronomic interpretation

The agronomic interpretation of the feature importance results highlights how biophysical and crop-specific factors influence productivity in smallholder rain-fed farms. Crop type emerged as the most

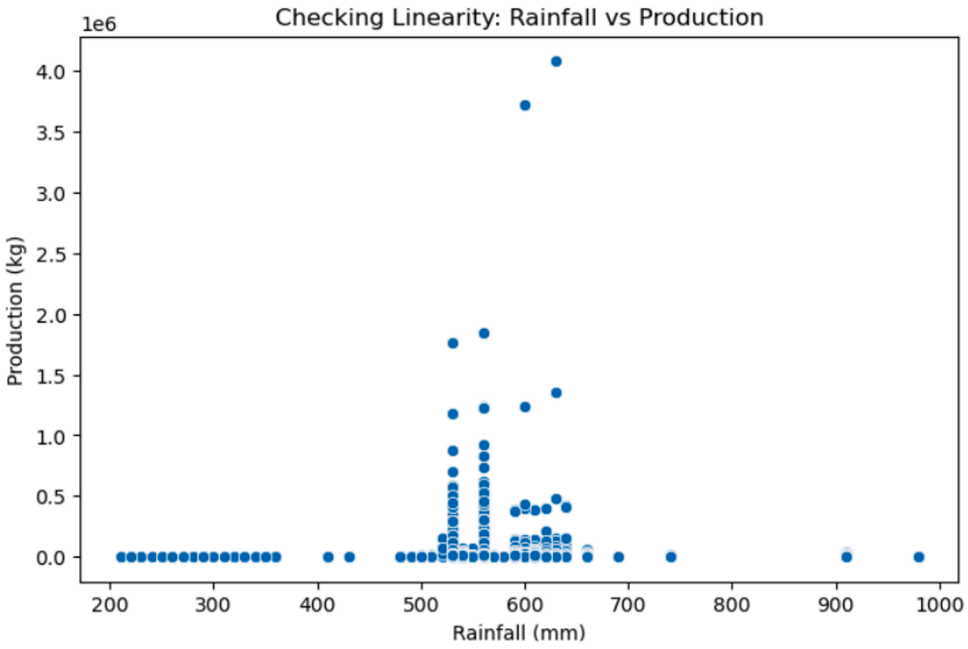



**Fig. 2.** Correlation between rainfall and production.

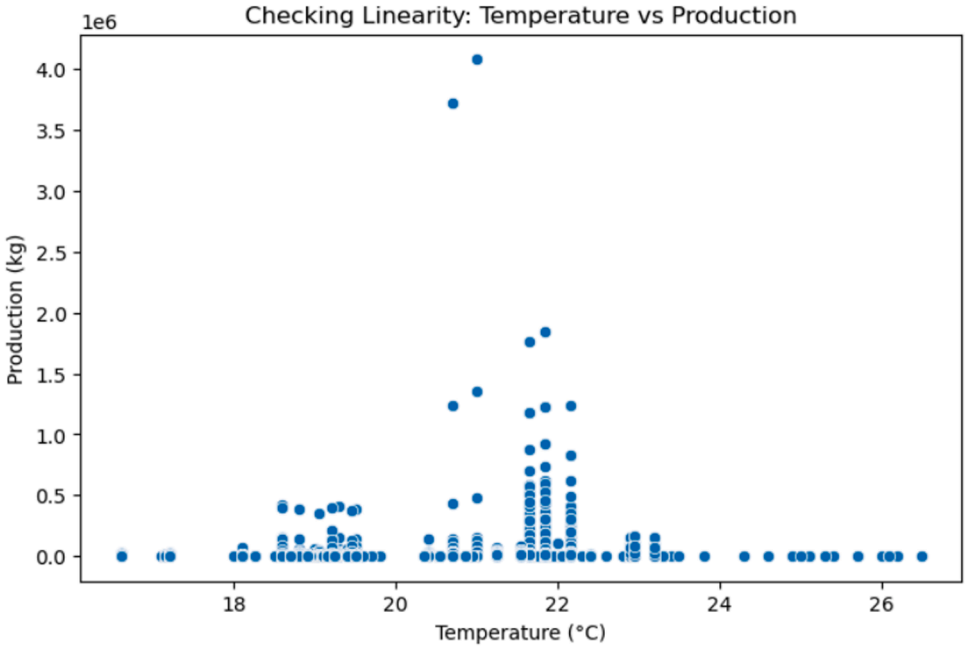



**Fig. 3.** Correlation between temperature and production.

important predictor, contributing more than 70% of the model's predictive power. This dominance reflects inherent physiological and genetic differences among rice, maize, and beans, including growth duration, nutrient requirements, canopy architecture, and tolerance to climatic stress factors that directly affect per-acre yield potential.

Rainfall was the second most important factor, accounting for approximately 25% of the model's predictive power. Its significance emphasizes the critical role of water availability in supporting key physiological processes such as germination, vegetative growth, flowering, and grain filling. Insufficient rainfall can limit biomass accumulation and reduce yields, particularly for water-demanding crops like rice and maize.

Other factors, including soil type, temperature, and farm size, contributed minimally, suggesting limited variability or influence within the study area. While soil type affects nutrient availability and water retention, its low importance indicates either relatively uniform soil conditions or effective buffering through farmers' management practices. Temperature showed limited effect, likely because the study area falls within a narrow thermal range that supports all three crops. Farm size influences total production but has minimal impact on per-acre productivity, implying that small and large farms achieve comparable yields per unit area.

These findings indicate that intrinsic crop characteristics and adequate water availability are the primary drivers of productivity. Consequently, interventions aimed at enhancing yields should prioritize the selection of high-yielding, well-adapted varieties and climate-

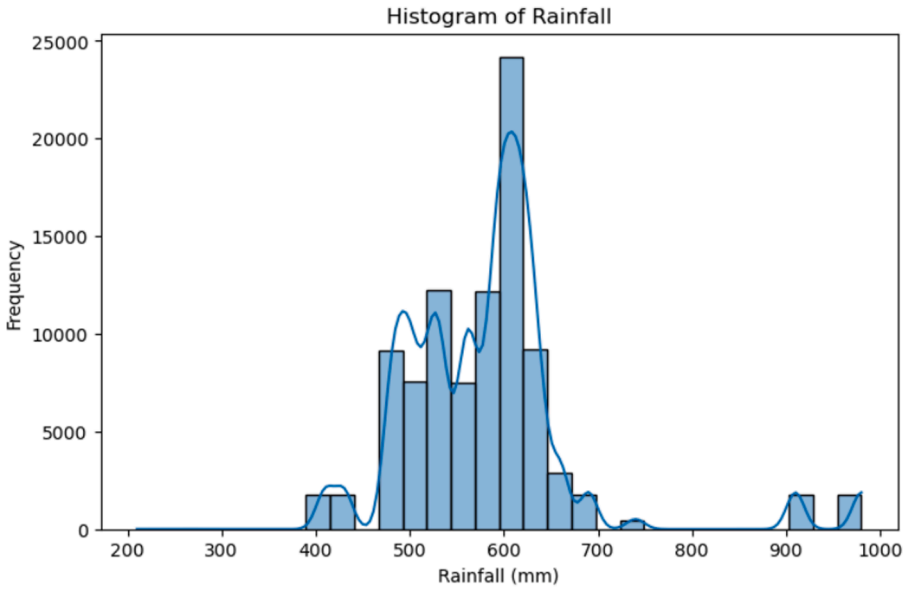



**Fig. 4.** Frequency distribution of rainfall feature.

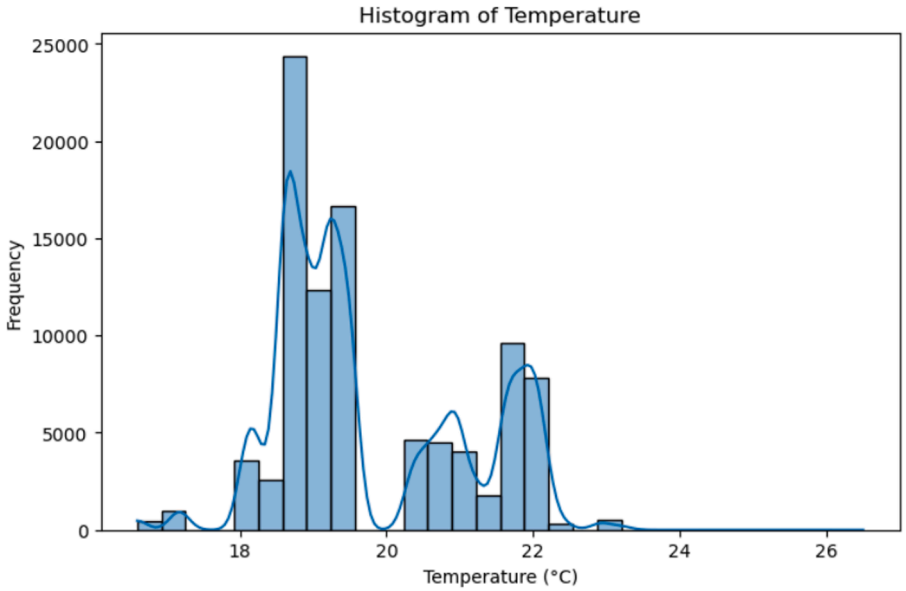



**Fig. 5.** Frequency distribution of temperature feature.

**Table 1**
Summary statistics for numerical variables.

| | Rainfall (mm) | Temperature (°C) | Farm size (ac) | Production (kg) |
|---|---|---|---|---|
| mean | 850.62997 | 22.9957 | 2.7735 | 1287.1555 |
| std | 201.119798 | 2.892222 | 1.313153 | 339.971237 |
| min | 500.1 | 18 | 1 | 482 |
| 25% | 679.7 | 20.5 | 2 | 1028 |
| 50% | 851.05 | 23 | 3 | 1266.5 |
| 75% | 1022.45 | 25.5 | 4 | 1544 |
| max | 1199.9 | 28 | 5 | 2172 |

adaptive water management practices, such as timely planting and the adoption of drought-tolerant cultivars, rather than focusing solely on expanding farm area or modifying soil types.

**Table 2**
Frequency distribution summary for categorical variables.

| | Soil type | Crop type |
|---|---|---|
| unique | 5 | 3 |
| top | Sandy Loam | Maize |
| freq | 2076 | 4621 |

### 3.4. Feature engineering

Feature engineering was performed to ensure data quality and model reliability. The final feature set consisted of rainfall (mm), temperature (°C), soil type, farm size (ac), and crop type, which were used as inputs to predict total crop productivity (kg) per farm plot. Outliers in rainfall, temperature, and farm size were examined to distinguish genuine extreme climatic events from erroneous measurements. Categorical variables (soil type and crop type) were transformed into numerical

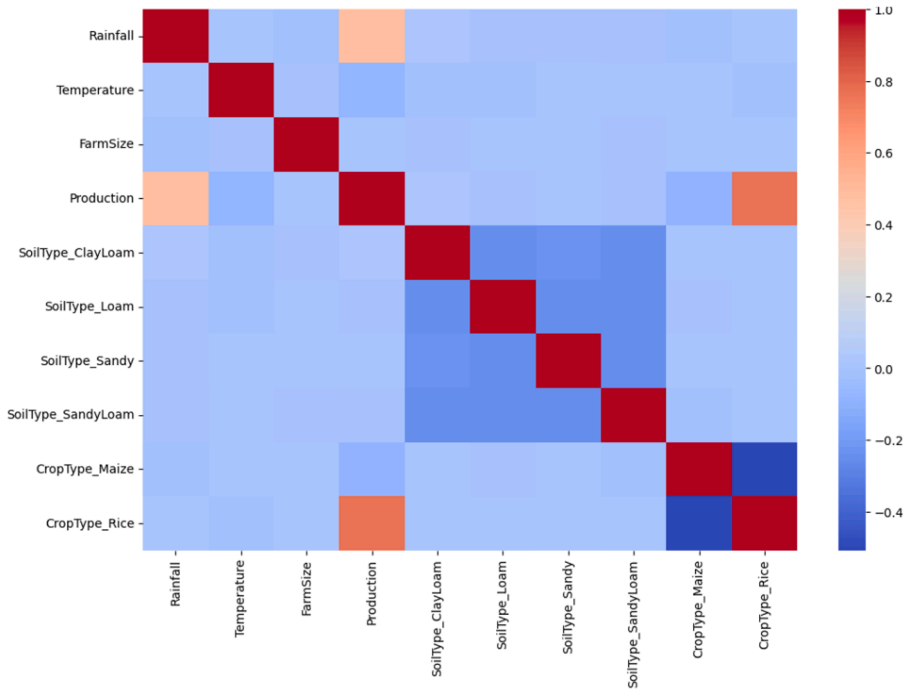



**Fig. 6.** Pearson correlation among features.

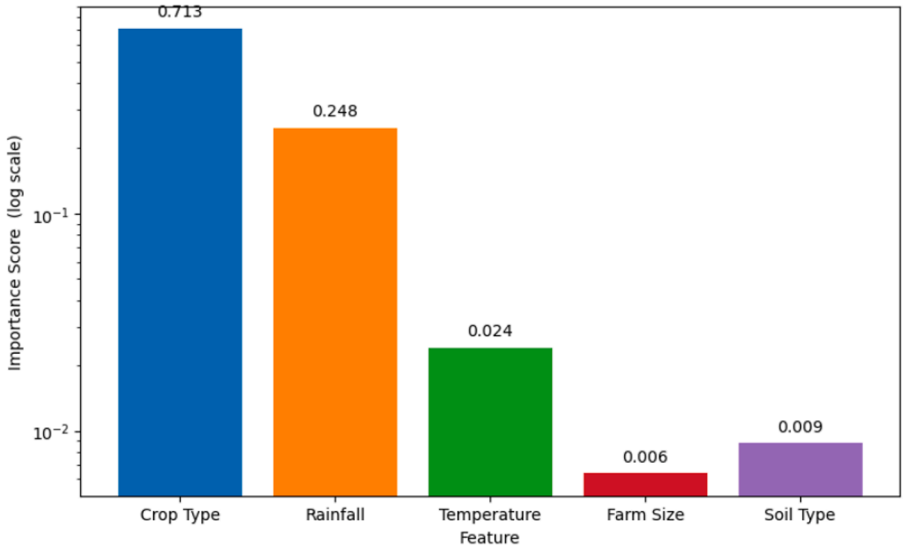



**Fig. 7.** Feature importance of crop yield predictors.

format using one-hot encoding. No scaling or normalization was applied to numerical features because tree-based algorithms, including Extra Trees, Random Forest, Gradient Boosting, and Extreme Gradient Boosting, are inherently insensitive to feature scaling. To avoid bias from a single train–test split, the initial 80/20 approach was replaced with stratified 5-fold cross-validation, ensuring proportional representation of crop types across folds and enabling a more robust evaluation. Model performance was assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and $R^2$, averaged across all folds.

### 3.5. Workflow of the proposed method

Fig. 8 illustrates the workflow for developing the farm-level machine learning model for crop yield prediction, outlining the full process from data acquisition to predictive analytics. Crop yield, farm size, soil type, and climate data were acquired from multiple sources and integrated into a centralized database. The data were then preprocessed, explored, and analyzed to uncover underlying patterns. Feature engineering techniques were applied to prepare the dataset for model training. Four machine learning models (Random Forest, Gradient Boosting, Extreme Gradient Boosting, and Extra Trees) were tested and evaluated. The final predictive model was designed to generalize to unseen data, enabling yield prediction for individual farms.

### 3.6. Model development

The experiment was conducted on a computer running the Windows 10 operating system, equipped with an Intel Core i7 processor and 8 GB of RAM. Model development was performed in the open-source Jupyter

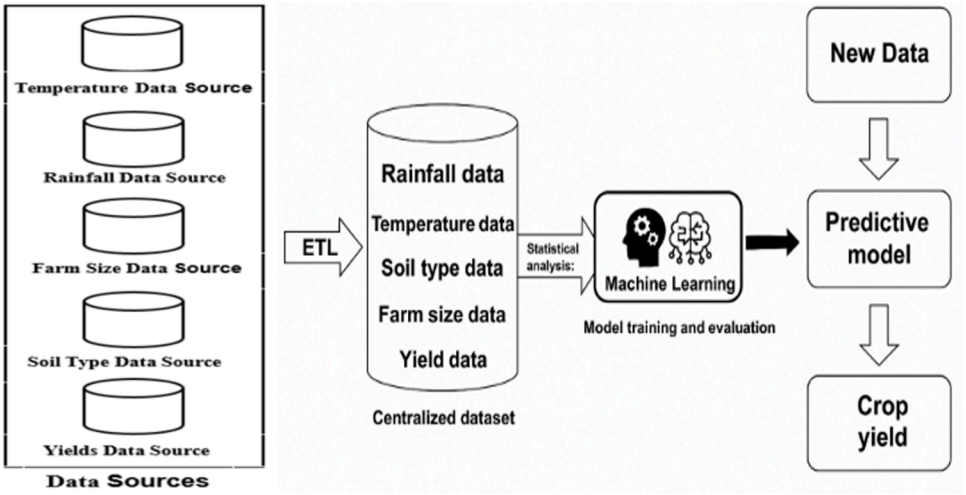



**Fig. 8.** Overall architecture of the proposed method.

Notebook environment using Python 3.8 as the programming kernel. Data processing and cleaning were carried out using the Pandas library, while NumPy supported numerical computations and matrix operations. Seaborn and Matplotlib were used for data visualization. Machine learning algorithms and evaluation metrics were accessed via the Scikit-learn library. Collectively, these tools ensured efficient data handling, model training, and performance evaluation.

## 4. Results

This section presents the performance of the machine learning models evaluated in the study. Four ensemble models were assessed using standard regression metrics ($R^2$, MAE, MSE, and RMSE). Table 3 summarizes the results obtained using default model parameters.

### 4.1. Random forest model

The Random Forest model achieved an MAE of 106.54 units, an MSE of 40563095.8 units, an RMSE of 6368.92 units, and an $R^2$ of 95%. The MAE indicates that predictions differed from actual values by an average of 106.54 units, while the RMSE reflects a larger deviation (6368.92 units) due to the squared error penalty. The $R^2$ value of 95% shows that the model explains most of the variance in the dataset, indicating strong predictive capability, though further improvement is possible to reduce prediction errors.

### 4.2. Gradient boosting model

The Gradient Boosting model achieved an MAE of 646.15 units, an MSE of 21399949.63 units, an RMSE of 4626.01 units, and an $R^2$ of 97%. The MAE represents an average deviation of 646.15 units between predicted and actual values, while the RMSE indicates a deviation of 4626.01 units. The $R^2$ of 97% demonstrates strong predictive accuracy. Despite the satisfactory performance, the observed errors indicate room for further tuning to enhance predictive precision.

**Table 3**
Performance results of the tested crop yield prediction models.

| Model Name | MAE | MSE | RMSE | $R^2$ (%) |
|---|---|---|---|---|
| Random Forest | 106.54 | 40563095.8 | 6368.92 | 95 |
| Gradient Boosting | 646.15 | 21399949.63 | 4626.01 | 97 |
| Extreme Gradient Boosting | 527.37 | 21899970.41 | 4679.74 | 97 |
| Extra Trees | 33.54 | 6494169.76 | 2548.37 | 99 |

### 4.3. Extreme gradient boosting model

The Extreme Gradient Boosting model recorded an MAE of 527.37 units, an MSE of 21899970.4 units, an RMSE of 4676.74 units, and an $R^2$ of 97%. This means predictions differed from true values by an average of 527.37 units, with an RMSE of 4676.74 units. The $R^2$ value shows that the model explains 97% of the variance. Although its performance is strong and consistent, further refinements could reduce prediction error margins.

### 4.4. Extra trees model

The Extra Trees model had an MAE of 33.54 units, an MSE of 6494169.76 units, an RMSE of 2548.37 units, and an $R^2$ of 99%. The very low MAE and high $R^2$ indicate exceptional predictive performance. With an $R^2$ of 99%, the model explains nearly all variance in the data, outperforming all other tested algorithms, despite minor prediction inconsistencies remaining.

### 4.5. Hyperparameter tuning

The results in Table 3 suggested potential overfitting. To improve generalization, hyperparameter tuning was performed using GridSearchCV with 5-fold farm-level cross-validation. All preprocessing steps, including normalization and one-hot encoding of soil and crop types, were fitted exclusively on the training folds and then applied to the validation folds to prevent data leakage. An 80/20 train–test split was used such that later seasons were excluded from training, preserving temporal independence and farm-level variability. Cross-validation folds were stratified by crop type to ensure balanced representation.

Grid specifications:

- **Extra Trees:** estimators {200, 400, 800, 1200}, max depth {10, 20, 30, None}, min samples split {2, 5, 10}, min samples leaf {1, 2, 4}, max features {“auto”, “sqrt”, “log2”}.
- **Random Forest:** same grid as Extra Trees, excluding Extra Trees’ randomized node-split behavior.
- **Gradient Boosting:** estimators {100, 200, 300, 500}, learning rate {0.01, 0.05, 0.1, 0.2}, max depth {3, 5, 7, 10}, subsample {0.6, 0.8, 1.0}.
- **Extreme Gradient Boosting:** learning rate {0.01, 0.05, 0.1, 0.3}, max depth {3, 6, 9, 12}, subsample {0.5, 0.7, 1.0}, colsample_bytree {0.5, 0.7, 1.0}, estimators {200, 400, 600, 800}.

Table 4 summarizes the performance of the optimized models. The

**Table 4**
Performance results of the tested crop yield prediction models after parameters tuning.

| Model Name | MAE | MSE | RMSE | $R^2$ (%) |
|---|---|---|---|---|
| Random Forest | 71.45 | 8058.08 | 84.77 | 93 |
| Gradient Boosting | 65.37 | 6674.18 | 81.70 | 94 |
| Extreme Gradient Boosting | 69.20 | 7498.68 | 86.59 | 94 |
| Extra Trees | 65.54 | 6719.43 | 80.97 | 95 |

Extra Trees model achieved the best performance (MAE = 65.37 units; MSE = 6719.43 units; RMSE = 80.97 units; $R^2$ = 95%). Gradient Boosting and Extreme Gradient Boosting both attained $R^2$ values of 94%, with slightly higher error metrics. Random Forest achieved an $R^2$ of 93%, accompanied by moderately higher MAE and RMSE.

#### 4.5.1. Model performance by crop type

Figs. 9–11 illustrate the crop-level performance of the four ensemble models. Random Forest achieved $R^2$ values of 76% for maize and beans and 74% for rice, with MAE ranging from 70.71 to 74.36 units and RMSE from 88.68 to 91.86 units, indicating stable but moderately variable accuracy across crops.

Gradient Boosting showed slightly improved performance, with $R^2$ values between 78% and 80%. Beans recorded the highest accuracy (80%), followed by maize (79%) and rice (78%), accompanied by lower MAE (66.30–69.83 units) and RMSE (83.37–86.57 units) values than Random Forest. Extreme Gradient Boosting demonstrated comparable stability, achieving $R^2$ values of 77% for maize and beans and 76% for rice, with MAE ranging from 68.91 to 72.22 units and RMSE from 86.08 to 89.00 units. The Extra Trees model delivered the strongest crop-specific performance, attaining $R^2$ values of 81% for maize and beans and 79% for rice. It also produced the lowest error metrics (MAE: 63.94–68.77 units; RMSE: 80.43–85.37 units), indicating superior generalization and consistent accuracy across all three crops.

## 5. Discussion

The Extra Trees model consistently outperformed other algorithms, likely due to its randomized split selection and decorrelated tree structure, which reduce variance and improve robustness to noisy, heterogeneous farm-level data. Random Forest and boosting-based models are more sensitive to local noise or sequential error propagation, which can limit performance when farmer-reported yields vary. Extra Trees effectively captures complex non-linear interactions among climatic, soil, and farm-level variables while maintaining computational efficiency suitable for farm-level deployment. Although Random Forest, Gradient Boosting, and Extreme Gradient Boosting also demonstrated strong performance, Extra Trees offered the best balance between predictive accuracy and runtime efficiency. Its ensemble structure stabilizes predictions across diverse agroecological zones and varying management practices, which is particularly important in datasets with high heterogeneity and noise.

Across the four models, beans achieved the highest predictive accuracy, followed by maize, whereas rice recorded the lowest $R^2$ values. Beans and maize are typically grown under more uniform agroecological conditions, reducing noise and enabling more stable model learning, whereas rain-fed rice is highly sensitive to water availability, soil moisture, and temperature variability, which increases yield heterogeneity and challenges prediction. The higher accuracy of the pooled multi-crop model reflects methodological advantages associated with a larger sample size, broader feature diversity, and shared cross-crop relationships among key climatic and farm-level variables. Ensemble methods, particularly Extra Trees, exploit these shared patterns through ensemble averaging and decorrelated tree construction, thereby

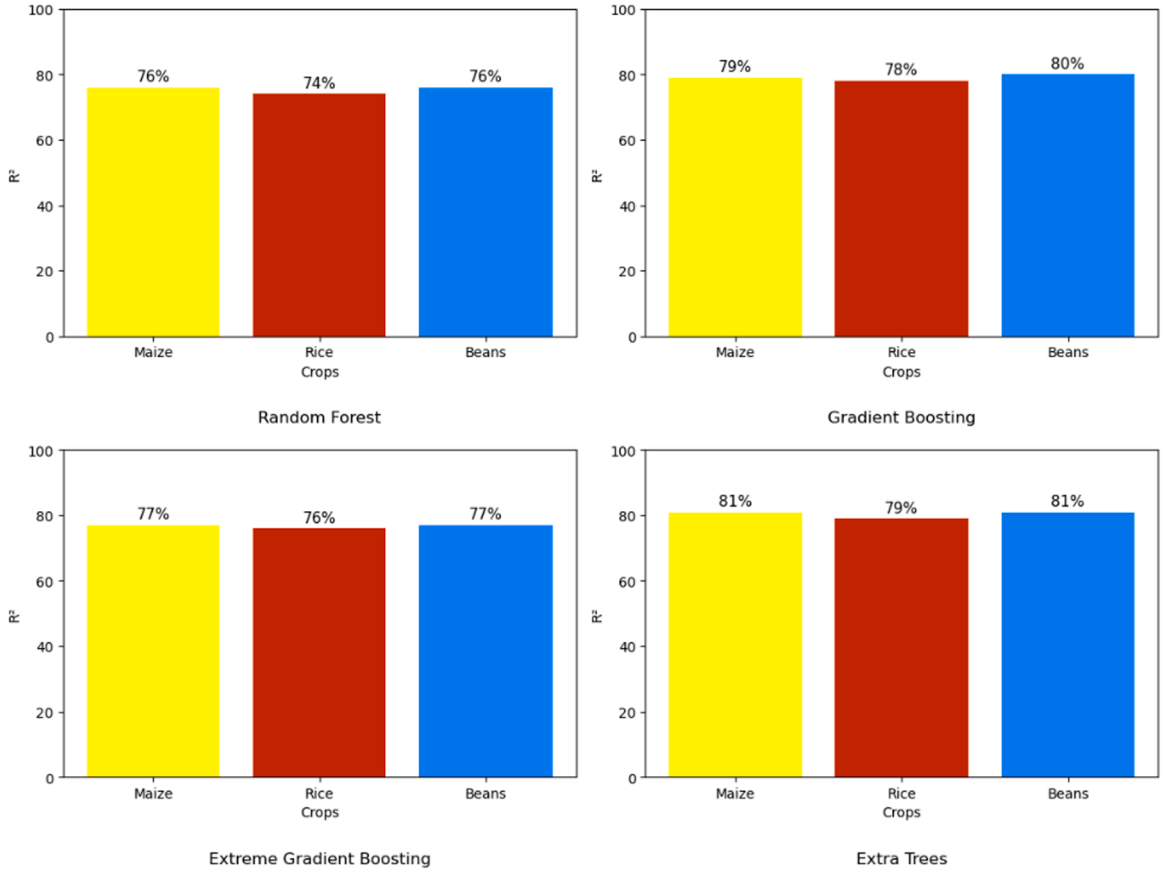



**Fig. 9.** $R^2$ values for maize, rice, and beans across four ML models.

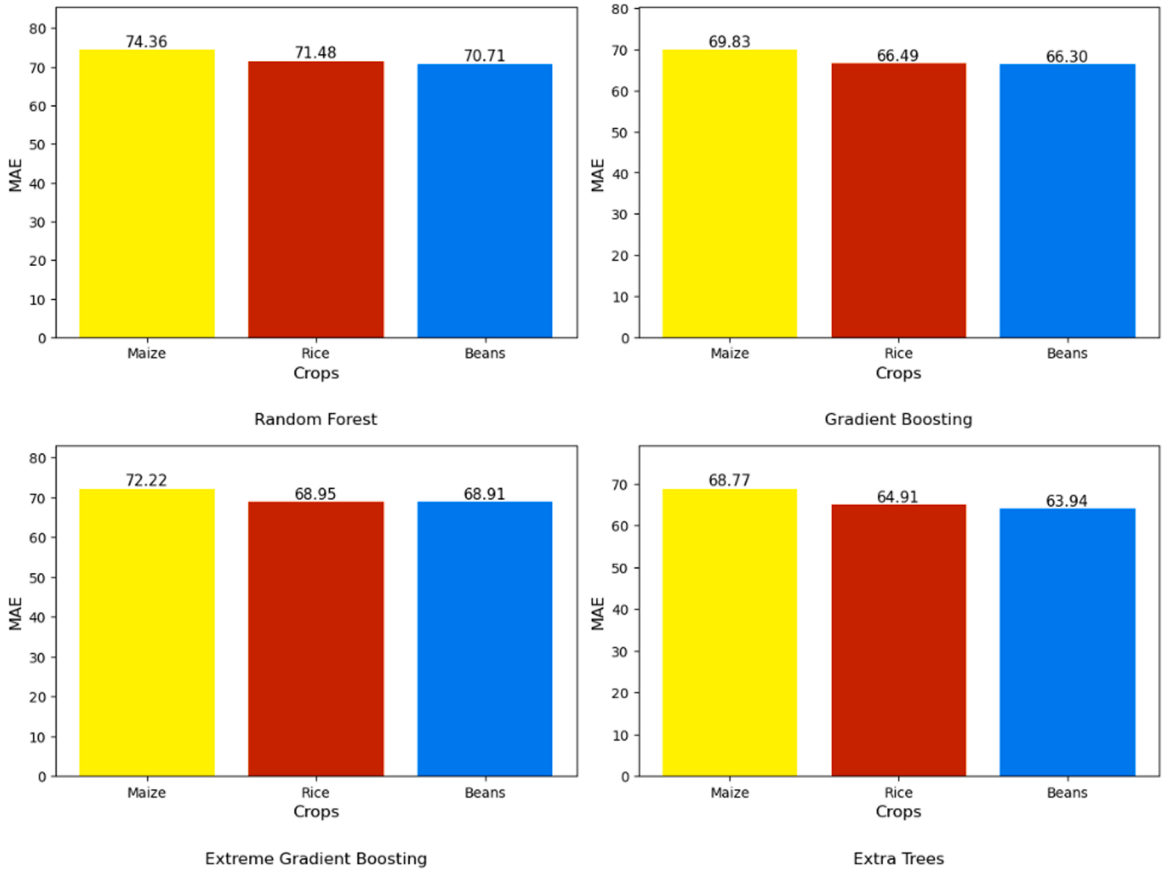



**Fig. 10.** Mean Absolute Error (MAE) for maize, rice, and beans across four ML models.

improving generalization across crops. In contrast, crop-specific models rely on smaller datasets with limited variability, constraining their ability to capture complex non-linear interactions and reducing predictive performance. Nonetheless, the pooled approach may mask fine-grained crop-specific dynamics, and remaining limitations include incomplete representation of microclimatic variability and potential bias in farmer-reported yields. Despite these constraints, well-calibrated ensemble models offer practical value for climate-adaptive crop management and for guiding extension services in prioritizing interventions.

## 6. Conclusion

This study demonstrates that machine learning models can provide reliable farm-level yield predictions for maize, rice, and beans in Tanzania using key climatic and farm-level variables, including rainfall, temperature, soil type, farm size, and crop type. Among the evaluated models, the Extra Trees ensemble achieved the highest predictive performance, highlighting the effectiveness of ensemble learning in capturing complex non-linear interactions. Despite limitations related to reliance on farmer-reported yield data and potential generalizability across heterogeneous agroecological zones, the results confirm the strong potential of ensemble-based machine learning models for supporting climate-adaptive agricultural decision-making.

To enhance both robustness and interpretability, future work should integrate machine learning with process-based crop simulation models such as Decision Support System for Agrotechnology Transfer (DSSAT) and Agricultural Production Systems Simulator (APSIM). One approach is hybrid modeling, where outputs from crop growth models are used as additional inputs for ensemble machine learning models, allowing the models to leverage both empirical data and biophysical crop processes. Model coupling is another strategy, in which machine learning algorithms calibrate or correct systematic biases in simulated yields. Stacked ensemble frameworks combining multiple ML models and crop simulation outputs could further improve accuracy while mitigating overfitting risks.

Explainable AI techniques, such as SHAP values or feature attribution methods, can quantify the influence of climatic and soil factors on predicted yields. When paired with process-based indicators, these methods provide agronomically meaningful insights that are understandable to farmers, extension officers, and policymakers. Integrating machine learning with process-based crop models offers a pathway toward accurate, interpretable, and transferable systems for predicting farm-level yields across agroecological contexts, addressing the challenges of data limitations and climate change in climate-vulnerable regions like Tanzania.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used QuillBot and Grammarly to improve the language to ensure the clarity and readability of the manuscript. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

## Data availability

Due to confidentiality restrictions associated with farm-level agricultural records, the raw data used in this study cannot be publicly released. To support replicability, anonymized, aggregated, or synthetic datasets that mirror the statistical properties of the original data will be made available by the authors upon reasonable request.

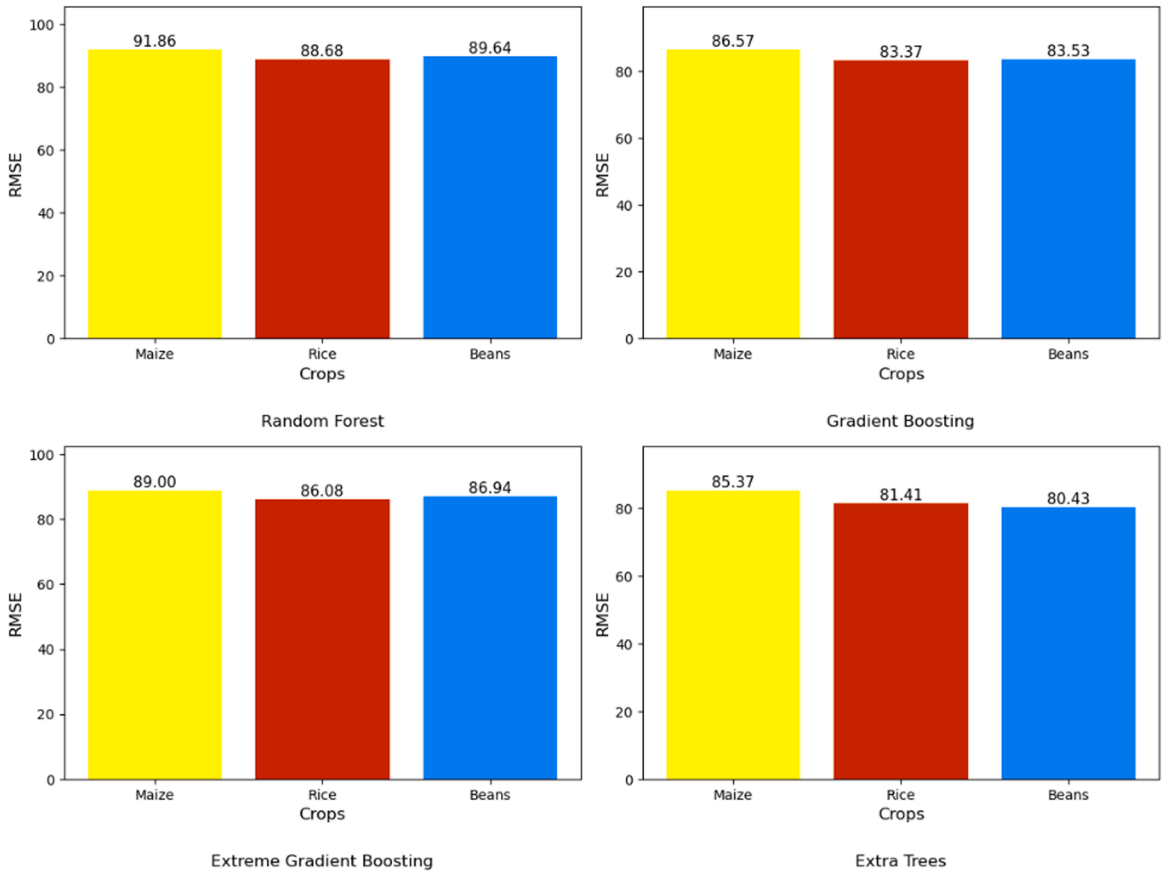



**Fig. 11.** Root Mean Squared Error (RMSE) for maize, rice, and beans across four ML models.

## Ethics Statement

Not applicable: This manuscript does not include human or animal research.

## CRediT authorship contribution statement

**Bally S. Omary:** Writing – review & editing, Writing – original draft, Resources, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mussa A. Dida:** Writing – review & editing, Supervision. **Devotha G. Nyambo:** Writing – review & editing, Supervision, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was carried out with financial support from UK International Development and the International Development Research Centre, Ottawa, Canada as part of the AI for Development: Responsible AI, Empowering People Program (AI4D).

## Data availability

The authors are unable or have chosen not to specify which data has been used.

## References

[1] R. Aworka, L. Cedric, W. Adoni, J. Zoueu, F. Mutombo, C. Kimpolo, T. Nahhal, M. Krichen, Agriculture decision system based on advanced machine learning models for yield prediction, Smart Agric. Technol. (2022).
[2] A. Ortiz-Bobea, T.R. Ault, C.M. Carrillo, R.G. Chambers, D.B. Lobell, Anthropogenic climate change has slowed global agricultural productivity growth, Nat. Clim. Chang. 11 (4) (2021) 306–312.
[3] C. Zhao, B. Liu, S. Piao, X. Wang, D.B. Lobell, Y. Huang, M. Huang, Y. Yao, S. Bassu, P. Ciais, J.-L. Durand, J. Elliott, F. Ewert, I.A. Janssens, T. Li, E. Lin, Q. Liu, P. Martre, C. Müller, S. Peng, J. Peñuelas, A.C. Ruane, Wall, Temperature increase reduces global yields of major crops in four independent estimates, Proc. Natl. Acad. Sci. (PNAS) 114 (35) (2017) 9326–9331.
[4] FAO, The State of Food Security and Nutrition in the World, Food and Agriculture Organization, Rome, 2021.
[5] O. Serdeczny, S. Adams, F. Baarsch, D. Coumou, A. Robinson, W. Hare, M. Schaeffer, M. Perrette, J. Reinhardt, Climate change impacts in Sub-Saharan Africa: from physical changes to their social repercussions, Reg. Environ. Change 17 (2017) 1585–1600.
[6] AfDB, The Economics of Climate Change in Africa, African Development Bank Group, 2020.
[7] P. Rowhani, D.B. Lobell, M. Linderman, N. Ramankutty, Climate variability and crop production in Tanzania, Agric. For. Meteorol. 151 (4) (2011) 449–460.
[8] D.B. Lobell, M.B. Burke, C. Tebaldi, M.D. Mastrandrea, W.P. Falcon, R.L. Naylor, Prioritizing climate change adaptation needs for food security in 2030, Science (1979) 319 (5863) (2008) 607–610.
[9] E.K.G. &. S. C. Mugalavai, Climate variability and crop productivity in Tanzania, Tanzania J. Agric. Sci. 22 (3) (2023) 145–160.
[10] URT, Tanzania Economic Survey Report, United Republic of Tanzania, Ministry of Finance and Planning, Dodoma, 2022.
[11] J. Kihoro, L. Nhamo, Artificial intelligence applications for sustainable agricultural transformation in Africa: opportunities and challenges, Smart Agric. Technol. 4 (2023).
[12] Y. Chen, W. Lee, H. Gan, N. Peres, C. Fraisse, Y. Zhang, Y. He, Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages, Remote Sens. (2019).
[13] M. Danilevicz, P. Bayer, F. Boussaid, M. Bennamoun, D. Edwards, Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection, Remote Sens. (2021).
[14] R. Fernandez-Beltran, T. Baidar, J. Kang, F. Pla, Rice-yield prediction with multi-temporal sentinel-2 data and 3D CNN: A case study in Nepal, Remote Sens. (2021).

[15] Y. Wang, Z. Zhang, L. Feng, Q. Du, T. Runge, Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States, Remote Sens. (2020).
[16] S.E.F. Shawon, A. Mahi, F. Niha, H. Zubair, Crop yield prediction using machine learning: an extensive and systematic literature review, Smart Agric. Technol. (2025).
[17] A. &. P.-B. F. X. Kamilaris, Deep learning in agriculture: A survey, Comput. Electron. Agric. 147 (2018) 70–90.
[18] A. Kaneko, T. Kennedy, L. Mei, C. Sintek, M. Burke, S. Ermon, D. Lobell, Deep Learning For Crop Yield Prediction in Africa (2019).
[19] J. Sun, Z. Lai, L. D, Multilevel deep learning network for county-level corn yield estimation in the U.S. Corn belt, Select. Top. Appl. Earth Observ. Remote Sens. (2020).
[20] M. Champaneri, D. Chachpara, C. Chandvidkar, M. Rathod, Crop yield prediction using machine learning, Int. J. Sci. Res. 9 (4) (2020) 2020.
[21] J. Bhoj, G. Bharte, C. Bhalerao, S. Ahire, B. Thakare, Crop recommendation system using machine learning algorithms, Int. Res. J. Moderniz. Eng. Technol. Sci. (2023).
[22] K. Jhajharia, P. Mathur, Machine learning based crop yield prediction model in Rajasthan region of India, Iraq J. Sci. 65 (2024) 390–400.
[23] K. Jhajhariaa, P. Mathur, S. Jain, S. Nijhawana, Crop yield prediction using machine learning and Deep learning techniques, Procedia Comput. Sci. (2023).
[24] S. Bao, Y. Wang, S. Ma, H. Liu, X. Xue, Y. Ma, M. Zhang, D. Wang, Field-scale maize yield estimation using remote sensing with the integration of agronomic traits, Agriculture (2025).
[25] M. Kuradusenge, E. Hitimana, D. Hanyurwimfura, P. Rukundo, K. Mtonga, A. Mukasine, C. Uwitonze, J. Ngabonziza, A. Uwamahoro, Crop yield prediction using machine learning models: case of Irish potato and maize, Agriculture (2023).
[26] S. Patrick, S. Mirau, I. Mbalawata, J. Leo, Timeseries and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change, Resourc., Environ. Sustain. (2023).
[27] I.G. Tende, K. Aburada, H. Yamaba, T. Katayama, N. Okazaki, Development and evaluation of a deep learning based system to predict district-level maize yields in Tanzania, Agriculture (2023).
[28] N. Bali, A. Singla, Deep learning based wheat crop yield prediction model in Punjab Region of North India, Appl. Artif. Intell. (2021).
[29] S. Khaki, L. Wang, Crop yield prediction using deep neural networks, Front. Plant Sci. (2019).
[30] URT, Mbeya Region profile: Welcome to Mbeya Region, Mbeya Regional Secretariat, Mbeya, 2025.
[31] URT, Mbeya Region investment guide 2020, Rungwe District Council, Mbeya, 2020.
[32] R.R. Secretariat, Ruvuma Region Tourism Strategic Plan 2023–2032, Government of the United Republic of Tanzania, Songea, 2023.
[33] URT, Songwe Region socio-economic profile, Songwe Regional Secretariat, Vwawa, 2025.
[34] J.D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. (2007) 90–95.
[35] M.L. Waskom, seaborn: statistical data visualization, J. Open. Source Softw. (2021).