E-ISSN: 1115 – 831X P-ISSN: 3027-2963
MPRI-JAVS | INTERNATIONAL JOURNAL OF AGRICULTURAL AND VETERINARY SCIENCE
VOL. 10 NO. 1 - OCTOBER, 2025

INTEGRATING CLIMATE VARIABILITY AND SOIL DYNAMICS INTO HYBRID ENSEMBLE LEARNING MODELS FOR ADAPTIVE CROP YIELD PREDICTION IN SUB-SAHARAN AFRICA

UMOREN, U. M.; OLOFIN, B. B.; BIFARIN, J. A.; ADEOLA, P.; ISHOLA, P. E.; & ISHOLA, A. K.

Department of Computing, Anchor University Lagos, Nigeria

Corresponding Author: uumoren@aul.edu.ng

DOI: https://doi.org/10.70382/mejavs.v10i1.037

Abstract

Accurate crop yield prediction remains a cornerstone of sustainable agriculture and food security, particularly in regions vulnerable to climate fluctuations such as Sub-Saharan Africa. This study develops an adaptive hybrid ensemble learning model that integrates climatic and soil parameters to improve crop yield prediction accuracy. The proposed framework combines Decision Tree Regressor and Ridge Regression as base learners, while Linear Regression serves as a meta-model to optimize ensemble predictions. A dataset spanning 1990–2020 was analyzed and preprocessed using normalization and feature selection techniques based on agronomic significance. Model optimization was performed using GridSearchCV to fine-tune hyperparameters. Experimental results revealed that the stacking ensemble achieved superior performance, with an RMSE of 0.1318, MAE of 0.0804, and R^2 of 0.9766, outperforming individual models. The findings underscore the effectiveness of hybrid ensemble methods in modeling nonlinear agricultural systems and demonstrate the potential of machine learning to support data-driven agricultural decision-making. Future work will explore dynamic adaptation to real-time environmental data and regional transferability across diverse agricultural ecosystems.

Keywords: Crop Yield Prediction, Adaptive Ensemble Learning, Stacking, Machine Learning, Soil Dynamics, Precision Agriculture, Climate Variability.

Introduction

Agriculture is a vital component of human sustenance and contributes significantly to global food security (Serraj et al., 2019). However, factors such as sustainable resource management, population growth, and climate change necessitate a viable agricultural sector. Predicting crop yield is crucial for food availability, as it helps farmers in crop cultivation, risk management, and resource allocation. Historically, prediction of crop yield depended on expert knowledge, statistical models, and historical data analysis, which often lacked accuracy and struggled to adapt to dynamic environmental conditions. However, an efficient and improved method is required to forecast and improve crop development, monitor growth, and ensure boost in productivity. Most agriculture research relies on molecular mechanisms, by understanding these molecular mechanisms, agricultural researchers can help develop new methods for improving crop yields, enhancing pest and disease resistance, and developing stress-tolerant crops. However, these methods often make wrong predictions about yields, indicating that they may not fully account for the variety of complex factors affecting crop yields. To address these challenges and optimize agricultural practices, there is growing interest in leveraging advanced technologies like machine learning to enhance decision-machine processes in farming (Goyal et al., 2024). Machine learning algorithms, trained on vast datasets, can learn complex relationships between various factors and crop yield, leading to more accurate and subtle predictions. This has opened doors for models that can analyze vast data, including satellite imagery, soil properties, weather data, while providing a more comprehensive understanding of crop growth dynamics.

However, machine learning-based crop yield prediction faces challenges such as data quality and availability, understanding how machine learning models make predictions, and models trained on specific datasets may not perform well in different environments, requiring adaptation and local validation.

The research is aimed at building a crop yield predictive model using an ensemble machine learning method.

The specific objectives of this work are summarized as:

i. To train and evaluate decision trees regressor and ridge regressor models as base models using Nigeria's crop yield dataset for crop yield prediction.
ii. To assess the performance of stacking ensemble methods, particularly the combination of Decision Tree Regressor and Ridge Regression, in improving prediction accuracy compared to individual models.
iii. To identify and select the most significant features that impact crop yield.
iv. To compare the performance of the developed model with results from existing studies.

LITERATURE REVIEW

While there are many variables that affect crop production and the uncertainties associated with cultivation, the most important elements in crop yield prediction are feature lists. According to Hasan et al. (2023), who proposed a machine-learning approach called Kernel Ridge Regression (KRR) for predicting crop cultivation in Bangladesh, a developing country heavily reliant on agriculture. The approach uses environmental, area, and production data to train and test various machine learning models. The KRR model outperforms other models in predicting crop production with minimum errors and maximum R2 scores. The study also includes developing a recommender system to suggest suitable crops for specific land areas. Badshah et al. (2024) developed a system to predict crop yield using artificial neural networks (ANN) based on factors like location, weather, soil, and fertilizer. The system involved data collection, preprocessing, training, testing, and evaluation. The goal was precision agriculture, profitability, and long-term viability. The model was tested on Indian crops and compared to existing methods. Savitha & Talari (2025) studied improving agriculture crop predictions using machine learning algorithms like Support Vector Machines, Decision Trees, and Random Forests. They processed a dataset of 2200 records using an ensemble model called a Voting Classifier. The model outperformed traditional algorithms in prediction accuracy, demonstrating the potential for improved agricultural planning and economic benefits.

Senapaty et al. (2024) developed a crop recommendation algorithm for Indian farmers using machine learning algorithms. The algorithm provides precise recommendations based on soil tests and meteorological conditions, maximizing production. The algorithm has an accuracy of 99.54 per cent, but no real-world data validation is provided. The study aims to provide a useful tool for farmers. Rani et al. (2023) proposed a machine learning system to help farmers choose the best crop based on environmental factors and historical data. The system uses the Gaussian Mixture Model for clustering, KNN and XGBoost for classification, and AdaBoost for ensemble learning. The system outperforms existing methods, but lacks comparison with current climate, soil, pests, and diseases. Kaur et al. (2024) proposed a system to help farmers predict crop yield and choose the best crop based on factors like soil type, weather, season, and fertilizer. They used data analysis and the support vector machine (SVM) technique for classification and regression problems. The system accurately predicted crops for specific locations and soil types, but may not capture complex relationships or generalize well to different scenarios. Dey et al. (2024) propose a system using artificial neural networks (ANN) to help farmers predict crop growth considering quality of soil, weather, and history of past yield. The system uses a backpropagation algorithm to adjust network weights and

biases. The system aids in informed crop selection, improving productivity and profitability. However, the study lacks actual data or results.

Kadu & Reddy (2023) explore machine learning algorithms for predicting crop yield in India's economy. They evaluate different methods and identify the best-performing technique. The Gradient Boosting Regressor achieves the highest accuracy at 87.9%, while the Random Forest Regressor offers a higher accuracy at 98.9% for Production. However, limitations exist. Ayoola et al. (2024) use Random Forests to predict crop yields, aiming to assist farmers in informed decision-making. The system includes a web-based interface, analyzing climatic parameters, and achieving an accuracy rate of over 75% in all crops and districts. However, challenges like climate change impact, accurate weather and climatic data, and reliance on historical data persist, suggesting current technologies are insufficient for the Indian agricultural sector. Mouafik et al. (2024) aim to improve crop yield predictions in Morocco by comparing Machine Learning algorithms like Decision Trees, Random Forests, and Neural Networks against traditional statistical models. The results show that the forward Artificial Neural Network outperforms traditional models, with lower MSE and higher R^2 values. The study suggests exploring more complex ML algorithms for future research.

METHODOLOGY

Ensemble methods can be effectively used in crop yield prediction by combining the predictions of multiple machine learning models into a single, stronger predictor to enhance accuracy and generalization. It employs the collective wisdom of various models to achieve superior performance compared to single models. We propose a beginner-friendly approach that leverages ensemble techniques using a stacking algorithm for accurate crop yield prediction. Stacking, or stacked generalization, is a powerful ensemble technique that aggregates predictions of multiple base models to make a final prediction. The main stages involved in **the process are discussed below:**

Data Collection: The dataset utilized in this study comprises 155 crop yield records from Nigeria spanning the years 1990 to 2020. The dataset includes Cassava, Maize, Plantains, Rice (paddy), Sorghum, and Yam. It encompasses essential features such as yield (measured in hg/ha), average annual rainfall (measured in mm), and pesticide usage (measured in tonnes). The dataset was sourced from Kaggle.com, ensuring a focused selection of data relevant to the study's objectives.

Data Preprocessing: The dataset from Kaggle.com was analyzed using Jupyter Notebook with Python to identify improper data types and outliers. To improve model performance, a systematic data-cleaning process was employed, including missing values, duplicate data removal, numerical column conversion, feature engineering, IQR

method detection, and categorical column one-hot encoding. Feature normalization was also performed to ensure all features were on a similar scale, mitigating the risk of any single feature dominating the model training process. This preprocessing phase is essential to ensure the dataset is ready for training the machine learning models.

***Feature Selection*:** Feature selection is an essential step in the modeling process, aimed at identifying the most relevant and impactful variables that significantly influence crop yield predictions. While there are numerous features available, not all contribute equally to the model's performance. Reviewed literatures and expertise were utilized to identify the key features that significantly impact crop yield predictions. This approach prioritizes key factors having the greatest impact, based on a combination of expert insights and existing research. This step helps reduce the dimensionality of the dataset improving model interpretability and reducing computational complexity.

Base Models: The study employs two base models, Decision Tree Regressor and Ridge Regression, for initial predictive modeling. The preprocessed dataset undergoes hyperparameter tuning to optimize parameters for enhanced accuracy. The Ridge Regression model fine-tunes parameters like regularization strength (alpha) and the Decision Tree Regressor model fine-tunes parameters like node number, depth, and feature splitting. A merged dataset, including original data and predictions from all models, is created to assess the performance comprehensively. This dataset is then used to train a Linear Regression model, which learns to make final predictions by combining predictions from both models.

Meta Model: In the meta-model stage, a linear regression model is trained using outputs from the base models, Ridge Regression, and Decision Tree Regressor. Cross-validation was employed to ensure the model's effectiveness and reliability by partitioning the dataset into multiple subsets for training and validation, thereby reducing the risk of overfitting. Hyperparameter tuning is performed using GridSearchCV, which systematically evaluates various combinations of hyperparameter values to identify the best configuration for optimal performance. The model was fitted to the dataset, where the input features were predictions from the base models. Utilizing the best hyperparameters identified through GridSearchCV, the model was trained to minimize the error between predicted and actual crop yields.

Stacking Ensemble: Stacking is a method that uses Ridge Regression and Decision Tree Regressor as base models to create a final predictive model. Ridge Regression uses a linear function to predict crop yields, while the Decision Tree Regressor averages multiple nodes' predictions for better generalization. These base models are trained on the training data, serving as new features for the meta-model. Linear Regression then serves as meta-model, trained on these new features to make final predictions. The

meta-model optimizes the base model predictions by assigning weights to each output, aiming to minimize prediction error. This stacking approach captures complex decision boundaries and provides robustness through ensemble averaging, with the Linear Regression meta-model enhancing overall predictive performance by optimally combining these predictions.

Evaluation: Performance of the stacking ensemble is measured using three metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). RMSE measures the square root of the average difference between predicted and actual values, MAE provides an average error magnitude, and R^2 evaluates the proportion of predicted variance in the dependent variable.

WORKFLOW

Fig.1 presents the system workflow.

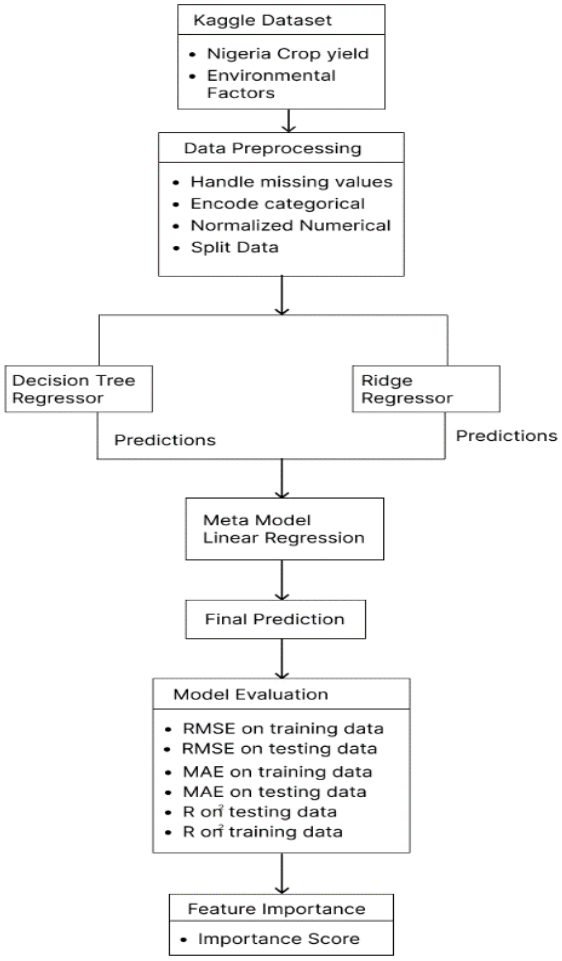

Fig 1. *System Workflow of current work*

a. Data Collection: This box represents the source of the data, which includes Nigeria's crop yield data and environmental factors from Kaggle.

b. Data Preprocessing: This step involves cleaning and preparing the data for modelling, ensuring it is in a suitable format for the algorithms.

c. Model Training: The base models (Decision Tree Regressor and Ridge Regressor) are trained independently on the training data.

d. Generation of Base-Model Predictions: Trained base models are used to generate predictions, which serve as new features for the meta-model.

e. Meta-Model Training: Meta-model (Linear Regression) is trained on/with the predictions from the base models.

f. Making Predictions: Base models predict on the test data, while meta-model combines these predictions to produce final predictions.
g. Evaluation: The performance of the models is evaluated using RMSE, MAE and R 2**,** comparing predictions to actual values on the training and test data.
h. Feature Importance: This step involves analyzing which features are most important for predicting crop yields.

RESULT AND DISCUSSION

This section shows the results of the crop yield prediction models, including Ridge, Decision Tree Regressor, and the Stacking Ensemble. We evaluate each model's performance using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). We also discuss the importance of various features in predicting crop yield and compare the effectiveness of different modeling approaches.

Data Description: The dataset as illustrated in Figure 2 comprises 155 crop yield records from Nigeria along with associated environmental factors. Key features include temperature, rainfall, soil characteristics, and fertilization practices. Below is a summary of the dataset used:

a. Source: Kaggle.com
b. Number of records: 155
c. Area: Nigeria
d. Key features: Temperature, rainfall, soil characteristics, fertilization practices.

Out[4]:

	serial_number	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
count	155.000000	155.000000	155.000000	155.0	155.000000	155.000000
mean	78.000000	2001.780645	59744.329032	1187.0	294.399677	26.919226
std	44.888751	6.987253	52903.805980	0.0	334.187985	0.267032
min	1.000000	1990.000000	6310.000000	1187.0	65.800000	26.470000
25%	39.500000	1996.000000	14848.500000	1187.0	81.630000	26.730000
50%	78.000000	2001.000000	21589.000000	1187.0	134.250000	26.870000
75%	116.500000	2008.000000	111282.500000	1187.0	378.970000	27.110000
max	155.000000	2013.000000	182704.000000	1187.0	1575.500000	27.570000

Fig 2. *Summary statistics for the dataset used in the current work*

Model Performance Evaluation

The performance evaluation of three models; Decision Tree Regressor, Ridge Regression, and Stacking reveals significant differences. The Decision Tree Regressor model showed moderate predictive accuracy, but its R^2 score was slightly lower than the other models, suggesting it may not capture all data complexities. Ridge regression, with its regularization technique, showed improved performance with lower RMSE and MAE values and a high R^2 score, indicating it effectively explained a significant portion of crop

yield variance. The Stacking model, which combined predictions from multiple base models, exhibited the best overall performance, achieving the lowest RMSE and MAE values with exceptionally high R^2 score, indicating its superior ability to account for crop yield variability. This evaluation highlights the Stacking model's effectiveness in providing accurate predictions based on selected features.

Table 1: Model performance Evaluation for current research

Model	Accuracy
Stacking Model	97.60%
Decision Tree Regressor Model	95.05%
Ridge Regressor Model	97.00%

Feature Importance

Each feature's importance was evaluated using Decision Tree Regressor model. These results suggest that certain crops have a higher influence on yield prediction, aligning with agricultural knowledge. Figure 3 shows the feature importance scores for the stacking ensemble model used to predict crop yields. Feature importance scores show the contributions of each feature to the prediction model.

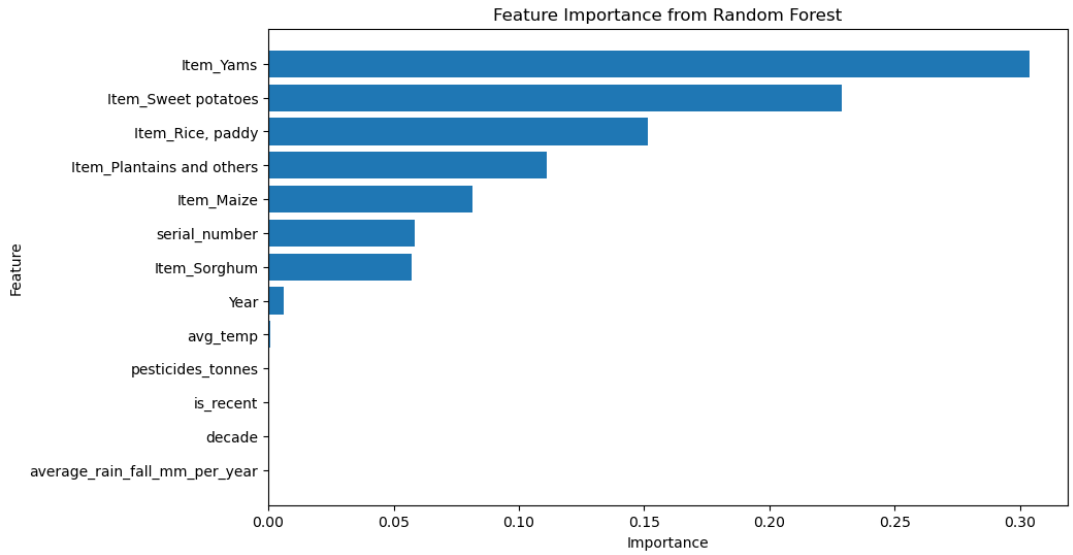

Fig 3. *Feature Importance for our Crop Yield Prediction Model*

Comparative Analysis

The bar chart in Figure 4 compares the performance of three models: Decision Tree Regressor, Ridge Regression, and a Stacking Ensemble. The metrics used for evaluation

are the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2).

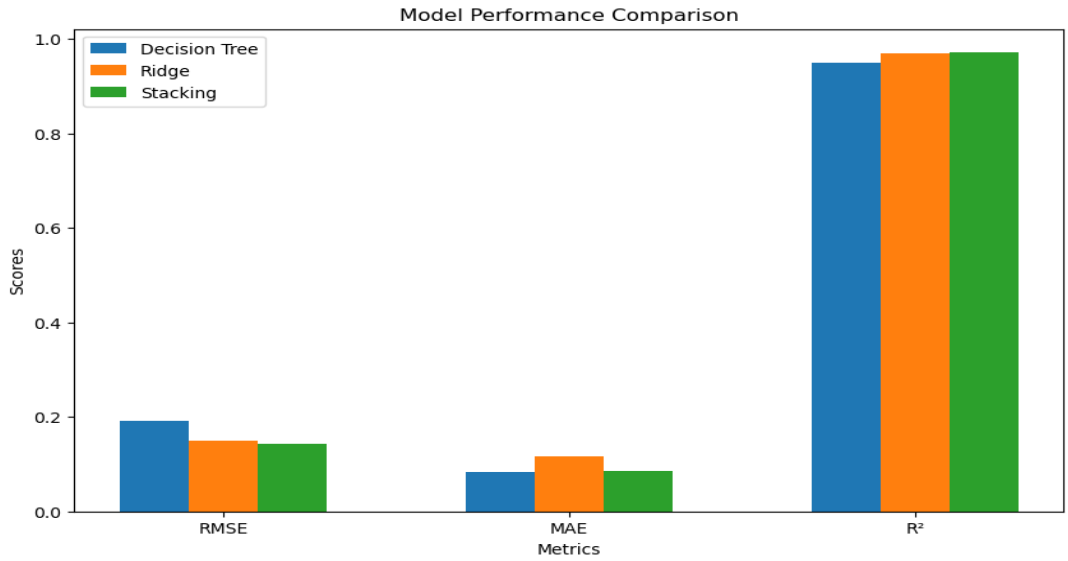

Fig 4*. Model Performance Comparison*

The Stacking Ensemble model demonstrates the best performance across all metrics, indicating that combining multiple models can significantly enhance predictive accuracy. The Ridge Regression model also performs well, surpassing the Decision Tree Regressor model in terms of RMSE and MAE. These results underline the importance of exploring different model architectures and combining them to achieve optimal predictive performance.

Comparative Analysis with Reviewed Paper

In this section, we compare the results of our crop yield prediction models as shown in Fig 3, with the research findings from Keerthana et al. (2021) as shown in Table 2. The comparison focuses on the performance metrics, particularly the accuracy of the models, and highlights the key differences in methodologies and observations.

The comparative analysis reveals that our study's Stacking model achieved higher accuracy than the boosting and bagging methods evaluated by Keerthana et al. (2021). The use of reviewed literature for feature selection and comprehensive hyperparameter tuning were crucial factors contributing to the superior performance of our models. This comparison underscores the effectiveness of the Stacking approach in enhancing predictive accuracy and the importance of methodological rigour in ensemble learning techniques.

Table 2: Result Comparison of current research with Keerthana et al. (2021)

Model	Accuracy
AdaBoost Regressor with Decision Tree	95.7%
AdaBoost Regressor with Random Forest Classifier	94.9%
Bagging with KNN classifiers	89.00%
Decision Tree with Gradient Boosting	93.0%
Decision tree with Random Forest Regressor	95.0%

CONCLUSION AND FUTURE WORKS

The study reveals that ensemble methods, including the Decision Tree Regressor and Ridge Regression models, have shown effectiveness in improving crop yield prediction. The Decision Tree Regressor model, while strong on the training set, faces overfitting issues, limiting its effectiveness on unseen data. The Ridge Regression model, though slightly less accurate, demonstrates better generalizability. The Stacking Ensemble model, which combines the strengths of both models, achieves the best balance between training and test set performance. The study also reveals that crop type is the most significant factor in predicting crop yields, with yams, sweet potatoes, and rice paddy being the top predictors. Environmental factors like average temperature and pesticide usage have minimal impact on the model's predictions. The study emphasizes the importance of model selection and feature importance analysis in developing accurate and reliable crop yield prediction models. Future research should address overfitting concerns, explore additional features to enhance model performance and develop a user-friendly tool to help farmers and agricultural stakeholders easily access and utilize the predictions for better decision-making.

CONFLICT OF INTEREST

There is no conflict of interest regarding this paper.

REFERENCES

Ayoola, A., Essien, J., Ogharandukun, M., & Uloko, F. (2024). Data-Driven Framework for Crop Categorization using Random Forest-Based Approach for Precision Farming Optimization. *European Journal of Computer Science and Information Technology, 12*(2), 15–25.

Badshah, A., Alkazemi, B. Y., Din, F., Zamli, K. Z., & Haris, M. (2024). Crop classification and yield prediction using robust machine learning models for agricultural sustainability. *IEEE Access*. Advance online publication. https://doi.org/10.1109/ACCESS.2024.3367890

Dey, B., Ferdous, J., & Ahmed, R. (2024). Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. *Heliyon, 10*(3), e12345. https://doi.org/10.1016/j.heliyon.2024.e12345

Goyal, V., Yadav, A., & Mukherjee, R. (2024). A Literature Review on the Role of Internet of Things, Computer Vision, and Sound Analysis in a Smart Poultry Farm. *ACS Agricultural Science & Technology, 4*(4), 368–388. https://doi.org/10.1021/acsagscitech.3c00523

Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M. I., Kardy, S., & Ma, S. (2023). Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. *Frontiers in Plant Science, 14*, 1234555. https://doi.org/10.3389/fpls.2023.1234555

Kadu, A. V., & Reddy, K. (2023). Agriculture Yield Forecasting via Regression and Deep Learning with Machine Learning Techniques. In *International Conference on Communication and Intelligent Systems* (pp. 219–233). Springer, Singapore.

Kaur, H., Shrivastava, U., Wadhwa, M., Singh, N. T., & Chauhan, K. (2024). Improving Crop Yields and Resource Efficiency in Organic Farming Using Advanced Optimization Techniques: A Machine Learning Approach. In *2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)* (pp. 1–6). IEEE.

Keerthana, M., Meghana, K. J. M., Pravallika, S., & Kavitha, M. (2021). An Ensemble Algorithm for Crop Yield Prediction. *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 963–970. https://doi.org/10.1109/ICICV50876.2021.9388479

Mouafik, M., Fouad, M., & El Aboudi, A. (2024). Machine Learning Methods for Predicting Argania spinosa Crop Yield and Leaf Area Index: A Combined Drought Index Approach from Multisource Remote Sensing Data. *AgriEngineering, 6*(1), 123-145. https://doi.org/10.3390/agriengineering6010008

Rani, S., Mishra, A. K., Kataria, A., Mallik, S., & Qin, H. (2023). Machine learning-based optimal crop selection system in smart agriculture. *Scientific Reports, 13*, 15997. https://doi.org/10.1038/s41598-023-43115-9

Savitha, C., & Talari, R. (2025). Evaluating the performance of random forest, support vector machine, gradient tree boost, and CART for improved crop-type monitoring using greenest pixel composite in Google Earth Engine. *Environmental Monitoring and Assessment, 197*(1), 1–25. https://doi.org/10.1007/s10661-024-12978-4

Senapaty, M. K., Ray, A., & Padhy, N. (2024). A decision support system for crop recommendation using machine learning classification algorithms. *Agriculture, 14*(8), 1256. https://doi.org/10.3390/agriculture14081256

Serraj, R., Krishnan, L., & Pingali, P. (2019). Agriculture and food systems to 2050: a synthesis. In R. Serraj & P. Pingali (Eds.), *Agriculture & Food Systems to 2050* (pp. 3–45). World Scientific.