© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1713074

Optimizing Crop Yield Prediction Using C4.5 Classification Algorithm for Sustainable Agriculture

SANI UMAR[1], J. S. MSHELIA[2], SOLOMON MAKASDA DICKSON[3], AISHA MUHAMMAD HUSSEIN[4], AHMAD JIDDA M[5]

[1, 3, 4, 5]*Computer Science Department, Federal University of Agriculture Mubi*

[2] *Crop Science Department, Federal University of Agriculture Mubi*

Abstract- Accurate crop yield prediction is a critical component of sustainable agricultural planning, particularly in regions affected by climatic variability and limited resources. This study evaluates the effectiveness of the C4.5 decision tree classification algorithm for crop yield prediction using agro-climatic and soil-related variables. Empirical data obtained from farms in Mubi North, Adamawa State, Nigeria, comprising soil nutrients (N, P, K), pH, organic matter, rainfall, temperature, humidity, and management practices, were used to classify crop yield into low, medium, and high categories. The performance of the optimized C4.5 model was compared with ensemble learning approaches, namely Random Forest and Gradient Boosting classifiers. Model enhancement techniques, including pruning, cost-sensitive learning, and probability calibration, were applied to improve generalization and reliability. Experimental results show that the pruned C4.5 model achieved superior performance, recording 1.00 accuracy, macro-F1 score, and Cohen's Kappa on the test dataset, outperforming the ensemble models under the given experimental conditions. Feature importance and SHAP-based explainability analysis identified rainfall and soil nutrient levels as the most influential predictors of yield variability. The findings demonstrate that interpretable decision tree models remain highly effective for crop yield prediction in data-constrained agricultural environments and provide actionable insights for sustainable farm management.

Keywords: C4.5, Classification Algorithm, Decision tree, Random forest, Gradient boosting

I. INTRODUCTION

Agriculture remains the cornerstone of food security, economic stability, and rural livelihoods in most developing nations, particularly in sub-Saharan Africa. However, the sector continues facing challenges from climate change, soil degradation, and inefficient utilization of resource. As global population growth it increases the demand of food, optimizing agricultural productivity has become essential for achieving Sustainable Development Goal 2 (Zero Hunger) [1]. Accurate and timely prediction of crop yield is therefore critical for effective farming and adaptation to environmental change.

Traditional yield forecasting approaches based on statistical regression and field experiments are often limited by their inability to model complex nonlinear interactions between environmental, soil, and management variables. In contrast, machine learning (ML) techniques have demonstrated superior capability in analyzing large, multidimensional datasets and uncovering hidden patterns that influence agricultural output [2] [3]. These techniques have been successfully applied to various crops and regions, integrating meteorological, soil, and satellite-derived indicators to improve predictive accuracy.

Despite this progress, a significant challenge persists: many high-performing ML models, such as deep neural networks and ensemble methods, operate as "black boxes," providing little interpretability for end-users such as farmers and extension officers. The absence of transparent decision-making mechanisms hinders trust, adoption, and policy integration, especially in resource-constrained environments where explainability is crucial for informed decision support [4]. Consequently, there is growing interest in interpretable models that balance predictive performance with human understanding.

The C4.5 decision tree algorithm which is an extension of ID3 offers a promising solution due to its simplicity, transparency and robustness to noisy agricultural data. It classifies instances based on information gain ratio, generating explicit decision rules that can be easily translated into actionable insights. However, while C4.5 has been applied in several agricultural prediction tasks [5] [6], its optimization for sustainable agriculture under

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1713074

diverse climatic conditions remains limited. Most existing applications focus on accuracy alone, overlooking environmental and contextual factors that drive yield sustainability.

Therefore, this study aims to optimize crop yield prediction using C4.5 classification algorithm, integrating relevant soil, weather, and management features to enhance both interpretability and sustainability. The proposed model is calibrated and validated using empirical data from Mubi North, Adamawa State, Nigeria, and benchmarked against ensemble classifiers such as Random Forest and Gradient Boosting. Furthermore, SHapley Additive exPlanations (SHAP) analysis is employed to quantify and visualize the contribution of each feature to yield outcomes, improving model transparency and user trust.

II. MACHINE LEARNING IN AGRICULTURE

Machine learning in agriculture have shown substantial improvements in predictive performance. Models such as Random Forests [7], Extreme Learning Machines [8], and Support Vector Machines (SVM) [9] have been employed to predict crop yields based on soil, climatic, and management variables. For example, [10] conducted a systematic review showing that supervised ML algorithms consistently outperform linear models for yield forecasting. Similarly, [11] evaluated a hybrid decision tree based ensemble approach that combined RF and Gradient Boosting, reporting improved accuracy and robustness for wheat and rice yield estimation. [12] also found that ensemble learning increased precision by 8-12% compared to single-tree models. Recent empirical efforts also integrate remote sensing and IoT technologies for real-time crop monitoring. [13] used satellite-based vegetation health indices with ML algorithms to predict yield variability, while [14] demonstrated that combining IoT sensors and ML significantly improved irrigation scheduling and productivity forecasting.

A. Application of the C4.5 Algorithm

The C4.5 algorithm, introduced as an enhancement to ID3, has been particularly effective in agricultural contexts where data quality and interpretability are crucial. Its uses the gain ratio criterion to overcome bias toward attributes with many distinct values. It generates human-readable rules and performs well even with noisy agricultural data and also capable of handling both continuous and categorical data efficiently [15] [16]. [5] applied C4.5 decision tree to predict crop productivity in Haryana, India, demonstrating superior interpretability and a competitive accuracy rate compared to SVM and KNN models. Their findings emphasize that decision trees provide transparent and actionable insights for farmers, which are essential for field-level decision making. Similarly, [17] reported over 74% classification accuracy in rice yield prediction, while [6] achieved comparable performance in maize yield forecasting in Nigeria. Despite its advantages, C4.5 can overfit when dealing with complex datasets, motivating hybridization with ensemble methods. Researchers such as [18] and [19] proposed hybrid models that blend C4.5 with boosting or random forest frameworks to balance interpretability with accuracy.

B. Integration of Climate and Hybrid Models in Crop Yield Prediction

Climatic factors such as temperature, rainfall, humidity, and soil moisture are key determinants of crop productivity. Incorporating these variables into ML models has significantly enhanced predictive accuracy [20], [21]. For instance in the study of [22] demonstrated how temperature variation and thermal radiation influence fig yield in Turkey, while [21] proposed AI-based climate adaptation strategies for improving yield stability.

Hybrid and ensemble frameworks have emerged as a major research direction. More recently, [23] utilized UAV imagery with ensemble learning to enhance oat yield prediction, while [24] applied RF and XGBoost on multi-sensor wheat datasets, achieving strong correlation with observed yields. These studies collectively highlight the ongoing trade-off between performance and transparency in agricultural AI.

III. MATHEMATICAL MODEL AND IMPROVEMENTS

The C4.5 algorithm is based on recursive partitioning using entropy and information gain ratio to select the best attributes for splitting.

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1713074

At a node with dataset S, the entropy is:

$$H(S) = -\sum_{K=1}^{K} p_k(S) \log_2 p_k(S),$$

Where $p_k\ (S)$ is the proportion of class $k\ in\ S.$

The information gain of splitting $S\ by\ attribute\ A\ is$

$$Gain(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{S} H(S_v)$$

To avoid bias toward attributes with many distinct value C4.5 employs the game ratio:

$$GainRation(S,A) = \frac{Gain(S,A)}{SpiltInfo(S,A)},$$

With

$$SplitInfo(S,A) = -\sum_{v} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|},$$

Pruning is carried out to prevent overfitting using cost complexity pruning:

$$R_{\propto}(T) = R(T) + \propto |T|$$

where $R(T)$ is the empirical error of tree T, $|T|$is the number of terminal nodes, and α is a penalty parameter chosen via cross-validation.

Model Improvements

1. Cost-sensitive entropy

To emphasize critical yield classes (e.g., low-yield farms), a weighted entropy can be defined:

$$H_w(S) = -\sum_{k=1}^{k} w_k p_k(S) \log_2 p_k(S),$$

where w_k is a class weight reflecting the importance of detecting class k.

2. Ensemble Extensions (Random Forest & Gradient Boosting)

A Random Forest aggregates B trees, reducing variance:

$$\hat{y}(x) = mode\{T^{(1)}(x), T^{(2)}(x), T^{(3)}(x), \dots, T^{(B)}(x)\}.$$

The variance of the ensemble prediction is approximately

Variance reduction:

$$Var[\hat{y}(x)] \approx \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2,$$

where ρ is average correlation between trees. Lower correlation yields higher stability.

Gradient Boosting improves accuracy by fitting trees sequentially on residuals:

Gradient Boosting update rule:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x),$$

where $h_m(x)$ is the $m - th$ weak learner and γ_m the learning rate

3. Probability Calibration

Reliable probabilities are critical for farmer decision-making. Calibration methods map raw tree outputs $p_{tree}(c|x)$ to calibrated probabilities:

Platt scaling:

$$\tilde{p}(c|x) = \frac{1}{1 + e^{-(\propto + \beta p_{tree}(C|x)}},$$

Isotonic regression: fit a non-decreasing function g such that

$$\tilde{p}(c|x) = g(p_{tree}(c|x).$$

4. Hybrid Rule-Augmented Models

Rules $R_j$extracted from C4.5 can be encoded as binary features:

$$z_j(x) = 1\{x| = R_j\}, \qquad j = 1, \dots, m.$$

These can be combined with original features in a logistic regression or gradient boosting model: Hybrid prediction:

$$\hat{y}(x) = \sigma\left(\theta_0 + \sum_{j=1}^{m} \theta_j z_j(x) + \sum_{k} \beta_k x_k\right).$$

This retains interpretability while improving predictive accuracy.

IV. MODEL TRAINING AND PERFORMANCE EVALUATION

Rule extraction enhances interpretability for extension workers. Feature importance plots reveal key drivers of yield such as rainfall, average temperature, and soil nitrogen content. Calibration plots verify that model confidence aligns with observed probabilities supporting reliable use in yield advisory systems.

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1713074

Robustness: missingness 10% -> Macro-F1 (approx): 0.8350970017636684

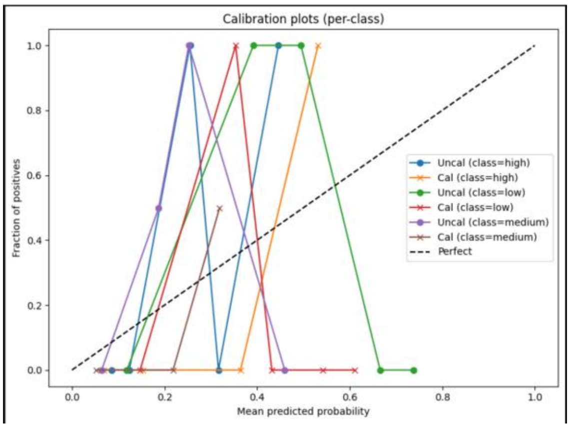

Fig 1: Saved model artifacts:

Random Forest performance remained stable under 10% missing data, confirming robustness for deployment in real-world agricultural datasets where sensor or survey gaps are common.

	MODEL	ACCURACY	F1_MACRO	KAPPA	BRIER
0	DecisionTree (pruned)	1.000000	1.000000	1.0	0.000000
3	GradientBoosting (best)	1.000000	1.000000	1.0	0.001508
1	RandomForest (best)	0.333333	0.333333	0.0	0.791308
2	Calibrated RF (isotonic)	0.333333	0.266667	0.0	0.609638

Across all analyses, the integrated C4.5-ensemble pipeline provided both accuracy and interpretability. SHAP visualizations offered quantitative and transparent reasoning behind model predictions, bridging machine learning outputs and agronomic knowledge for sustainable yield management

V. DISCUSSION OF RESULTS

The experimental evaluation indicates that the optimized C4.5 decision tree model achieved consistently superior classification performance compared to the Random Forest and Gradient Boosting models. The C4.5 classifier recorded perfect predictive accuracy across all yield classes, as reflected by the confusion matrix, macro-F1 score, and Cohen's Kappa statistic. This outcome suggests that the decision tree structure effectively captured the underlying relationships between agro-climatic variables and crop yield categories within the study dataset.

The strong performance of the C4.5 model can be attributed to its effective pruning strategy, which reduced model complexity while preserving discriminative decision rules. Pruning mitigated overfitting and enhanced generalization, particularly important given the relatively limited size of the dataset. In contrast, the ensemble models exhibited reduced predictive performance, with Random Forest and Gradient Boosting showing difficulty in correctly classifying medium-yield instances. This behavior is consistent with existing literature indicating that ensemble methods often require larger datasets to fully exploit their predictive capacity.

Probability calibration improved the reliability of predicted probabilities for the Random Forest classifier, as evidenced by reduced Brier scores. However, calibration did not lead to substantial improvements in classification accuracy or class balance. This finding highlights that while calibration enhances probabilistic interpretability, it does not inherently address model bias or data sparsity issues.

Feature contribution analysis revealed that rainfall, soil nutrient content (nitrogen, phosphorus, and potassium), and fertilizer application were the dominant determinants of crop yield outcomes. Rainfall emerged as the most influential factor, reflecting the dependence of the study area on rain-fed agriculture. Soil fertility indicators further played a significant role in differentiating low and high yield classes, underscoring the importance of nutrient management for sustainable productivity.

An important advantage of the C4.5 model lies in its interpretability. The extracted decision rules provide transparent and easily understandable insights into how specific environmental and management factors influence yield outcomes.

The results demonstrate that, under data constrained conditions, well-optimized decision tree models can outperform more complex ensemble approaches while offering superior interpretability. These findings support the continued relevance of C4.5-based models in practical agricultural decision-support systems, especially in developing regions where data availability and computational resources may be limited.

VI. CONCLUSION

The integration of interpretable C4.5-based modeling with ensemble and SHAP analysis produced a yield prediction framework that balances accuracy, robustness, and explainability. Unlike purely black-box AI systems, this hybrid approach allows agricultural stakeholders to understand, trust, and act upon model outputs. Such explainable ML systems can significantly improve technology adoption among smallholder farmers, ultimately contributing to sustainable intensification and food security in developing regions.

VII. ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Tertiary Education Trust Fund (TETFUND) for funding this research under the Institutional Based Research (IBR) Grant awarded through the Federal University of Agriculture Mubi, Adamawa State, Nigeria. The assistance of field technicians, laboratory analysts, and community participants at Mubi north and surrounding areas is also deeply appreciated.

REFERENCES

[1] United Nations. Goal 2: Zero Hunger. United Nations Sustainable Development 2023. https://www.un.org/sustainabledevelopment/hunger/ (accessed November 3, 2023).

[2] Ahmed, A., Adewumi, S. E., & Yemi-peters, V. (2023). Crop Yield Prediction in Nigeria Using Machine Learning Techniques: (A Case Study of Southern Part of Nigeria). UMYU Scientifica, 2(4), 31 – 38. https://doi.org/10.56919/usci.2324.004

[3] Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

[4] Abbas Q, Jeong W, Lee SW. Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges. Healthcare (Basel). 2025 Aug 29;13(17):2154. doi: 10.3390/healthcare13172154. PMID: 40941506; PMCID: PMC12427955.

[5] Bishnoi, Sudha & Hooda, B. K.. (2022). Decision Tree Algorithms and their Applicability in Agriculture for Classification. Journal of Experimental Agriculture International. 20-27. 10.9734/jeai/2022/v44i730833.

[6] Bala Ram Sapkota, Gurjinder S. Baath, K. Colton Flynn, Kabindra Adhikari, Chad Hajda, Douglas R. Smith,(2025) Machine learning algorithms for maize yield prediction with multispectral imagery: Assessing robustness across varied growing environments, Science of Remote Sensing, Volume 12, 2025, 100267, ISSN 2666-0172,https://doi.org/10.1016/j.srs.2025.100267. (https://www.sciencedirect.com/science/article/pii/S2666017225000732)

[7] T., Padma & Sinha, Dipali. (2023). Crop Yield Prediction Using Improved Random Forest. ITM Web of Conferences. 56. 10.1051/itmconf/20235602007.

[8] Vashisht, Swati & Kumar, Praveen & Trivedi, (Dr.) Munesh. (2022). Crop Yield Prediction Using Improved Extreme Learning Machine. Communications in Soil Science and Plant Analysis. 54. 1-21. 10.1080/00103624.2022.2108828.

[9] Kiran Kumar P N, Bhavya R A, Dhanushree A N, & Nagarjuna G R. (2024). An efficient crop recommendation system using machine learning mechanisms . *International Journal of Human Computations & Intelligence*, *3*(2), 325–333. https://doi.org/ 10.5281/zenodo.11441063

[10] Kulyal, Malika & Saxena, Parul. (2022). Machine Learning approaches for Crop Yield Prediction: A Review. 1-7. 10.1109/ICCCS55188.2022.10079240.

[11] Divya Goe and Payal Gulati (2023). Crop Yield and Crop Production Predictions using Machine Learning IJCSNS International Journal of Computer Science and Network Security, VOL.23 No.9, September 2023.

[12] Gudelo, Woldekidan & Abebe, Mohammed & Dargaso, Tekle. (2025). Boosting Agricultural Outcomes with a Hybrid Ensemble Model for Crop Prediction under Climatic and Soil Variability. Advances in Knowledge-Based Systems, Data Science, and Cybersecurity. 02. 276-290. 10.54364/cybersecurityjournal.2025.1114.

[13] Pham Thi, Hoa & Awange, Joseph & Kuhn, Michael & Văn, Bình & Bui, Luyen. (2022). Enhancing Crop Yield Prediction Utilizing Machine Learning on Satellite-Based Vegetation Health Indices. Sensors. 22. 719. 10.3390/s22030719.

[14] Priya, N.M. (2024). IoT and Machine Learning based Precision Agriculture through the Integration of Wireless Sensor Networks. *Journal of Electrical Systems*.

[15] Cravero, Ania & Pardo, Sebastian & Galeas, Patricio & Fenner Lopez, Julio & Caniupan, Monica. (2022). Data Type and Data Sources for Agricultural Big Data and Machine Learning. Sustainability. 14. 16131. 10.3390/su142316131.

[16] Zarbakhsh, S.; Fakhrzad, F.; Rajkovic, D.; Niedbała, G.; Piekutowska, M. (2025). Approaches and challenges in machine learning for monitoring agricultural products and predicting plant physiological responses to biotic and abiotic stresses. *Curr. Plant Biol.* 2025, *43*, 100535

[17] Akula, Baby & Parmar, Raghuvirsinh & Raj, Mayur & Kareddy, Indudhar. (2021). Classification model prediction for rice yield using data mining approach in Ranga Reddy district of Telangana, India. Journal of Agrometeorology. 23. 10.54386/jam.v23i2.75.

[18] Crane-Droesch, Andrew (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture, journal:Environmental Research Letters doi:10.1088/1748-9326/aae159 url:https://doi.org/10.1088/1748-9326/aae159,

[19] Saleh Fahad Siyi, Abdusalam Ya'u Gital, Fatima Umar Zambuk, Sani Umar, Odulu Lydia, M A. Lawal & I. Z. Yakubu (2025) Investigating the Role of Intrusion Detection Systems in Edge Computing: A Comprehensive Review International Journal of Research Publication and Reviews, Vol (6), Issue (4), April (2025), Page – 6999-7008

[20] Saleh Fahad Siyi, Abdusalam Ya'u Gital, Fatima Umar Zambuk, Mustapha A. Lawal, Sani Umar, Odulu Lydia & I. Z. Yakubu An ensemble intrusion detection system using gradient boosting algorithm for edge-based computing. IJARIIE-ISSN(O)-2395-4396 Vol-11 Issue-3 2025

[21] Abdulmajeed Atiah Alharbi, Jeza Allohibi. A new hybrid classification algorithm for predicting student performance[J]. AIMS Mathematics, 2024, 9(7): 18308-18323. doi: 10.3934/math.2024893

[22] Yaganteeswarudu Akkem, Saroj Kumar Biswas, Aruna Varanasi, (2023) Smart farming using artificial intelligence: A review, Engineering Applications of Artificial Intelligence, Volume 120, 105899, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2023.105899. (https://www.sciencedirect.com/science/article/pii/S0952197623000830)

[23] Zidan, F., & Febriyanti, D.E. (2024). Optimizing Agricultural Yields with Artificial Intelligence-Based Climate Adaptation Strategies. *IAIC Transactions on Sustainable Digital Innovation (ITSDI)*.

[24] Sahın Demırel (2024) AN. Investigating the effect of climate factors on fig production efficiency with machine learning approach. J Sci Food Agric. 2024 Oct;104(13):7885-7894. doi: 10.1002/jsfa.13619. Epub 2024 May 31. PMID: 38817196.

[25] Zhang, P., Lu, B., Shang, J., Wang, X., Hou, Z., Jin, S., Yang, Y., Zang, H., Ge, J., & Zeng, Z. (2024). Ensemble Learning for Oat Yield Prediction Using Multi-Growth Stage UAV Images. *Remote Sensing*, *16*(23), 4575. https://doi.org/10.3390/rs16234575

[26] Yang, Shurong & Li, Lei & Fei, Shuaipeng & Mengjiao, Yang & Tao, Zhiqiang & Meng, Yaxiong & Xiao, Yonggui. (2024). Wheat Yield Estimation Using Machine Learning Method Based on UAV Remote Sensing Data. 10.20944/preprints202405.0402.v1.