IoT-Based Crop Yield Prediction Using Machine Learning and Sensor Data in the Semi-Arid Region

O. M. Olanrewaju[1,*], E. A. Jiya[1], F. O. Echobu[1], A. AliBala[2], E. D. Ajik[1], H. U. Baba[1], J. N. Ndabula[1], S. Luka[1]

[1]*Department of Computer Science, Federal College of Agricultural Produce Technology, Kano, Nigeria*
[2]*Department of Computer Science and Information Technology, Federal University Dutsin-Ma, Katsina State, Nigeria*
**Corresponding author:* oolanrewaju@fudutsinma.edu.ng, jiyaeli@fudutsinma.edu.ng, fadebiyi@fudutsinma.edu.ng alibala1989@fcapt.edu.ng, adanny@fudutsinma.edu.ng, hussainausman1@gmail.com, josephndabula@gmail.com, lstephen@fudutsinma.edu.ng

***Abstract* - This study develops and evaluates an IoT-integrated machine learning system for maize yield prediction in semi-arid northern Nigeria. Real-time soil and environmental data (temperature, moisture, pH, electrical conductivity, N, P, K) were collected during the 2024 rainy season in Katsina State using a 7-in-1 soil sensor interfaced with an ESP-32 microcontroller. After preprocessing (outlier removal via IQR, standardization), three regression models Random Forest, K-Nearest Neighbors (KNN), and Linear Regression were trained and evaluated using Weka. Random Forest achieved the best performance (R^2 = 0.85, MAE = 0.50, RMSE = 0.77), significantly outperforming KNN (R^2 = 0.75, MAE = 0.60) and Linear Regression (R^2 = 0.65, MAE = 0.70). The model was deployed via a web platform (https://iot-maize.com.ng), enabling real-time yield estimation for farmers. This work demonstrates the viability of IoT-ML integration to enhance decision-making and farm profitability in resource-constrained, semi-arid environments.**

***Keywords* – Crop Yield Prediction, IoT-based yield prediction, machine learning yield prediction, crop yield in Nigeria, IoT-Based Smart Farming**

I. INTRODUCTION

The agricultural sector in Nigeria is undergoing transformation as the nation seeks to diversify from oil dependence and revitalize farming. Traditional practices often involving imprecise fertilizer or irrigation application led to resource waste and suboptimal yields, especially in semi-arid zones like northern Nigeria [1]. Climate variability and soil degradation further threaten food security, necessitating data-driven interventions.

IoT and machine learning (ML) offer transformative potential by enabling continuous monitoring and predictive analytics [2]. For maize a critical staple crop accurate yield forecasting supports timely agronomic decisions [3]. This study bridges a key gap: while ML-based yield prediction is well-explored [4–10], few systems integrate real-time IoT sensor data in semi-arid West Africa, and even fewer deploy accessible, farmer-facing tools.

We present an end-to-end IoT-ML framework for maize yield prediction in Katsina State, Nigeria. Our contributions include: This study makes four concrete contributions to precision agriculture in semi-arid regions

a. The design, calibration, and field deployment of a low-cost IoT sensing system in Katsina State, Nigeria—integrating a 7-in-1 soil sensor (N, P, K, pH, EC, moisture, temperature) with an ESP-32 microcontroller and cloud logging via Google Cloud Platform, enabling real-time data collection over the 2024 maize-growing season.
b. The development and rigorous comparative evaluation of three machine learning models on 595 *real-world* field samples, using 5-fold cross-validation and standardized metrics—demonstrating that Random Forest significantly outperforms KNN and Linear Regression (MAE = 0.50, R^2 = 0.85), with feature analysis confirming the agronomic relevance of inputs (e.g., moisture–yield $r = 0.71$, $p < 0.01$);
c. The first publicly accessible, farmer-facing yield prediction tool for Nigerian semi-arid maize farming, deployed at https://iot-maize.com.ng , featuring a lightweight HTML/CSS/JavaScript frontend and Flask backend—designed for low-bandwidth usability and validated in pilot testing with local farmers;
d. Empirical evidence that *real-time IoT sensor data*—not just historical or remote-sensing proxies—can substantially improve prediction accuracy in data-scarce, smallholder contexts, as shown by a 36% reduction in MAE relative to prior UAV-based work (Ramos et al., 2020: MAE = 0.78 → this work: MAE = 0.50).

II. MATERIALS AND METHODS

The research methodology, illustrated in Figure 1, presents a detailed workflow for the study. It commenced with the implementation of an IoT device, where sensors are integrated with a micro-controller and programmed to gather data. This was followed by experimental field preparation and crop planting, a key step for obtaining meaningful data. Next, dataset acquisition was carried out using IoT devices to monitor environmental and crop conditions. The collected data were then cleaned and processed during the preprocessing stage to make them suitable for analysis. In the subsequent phase, machine learning models were developed using processed data. These models were then tested and validated to ensure their accuracy. Finally, the model was made accessible online at https://iot-maize.com.ng/ .

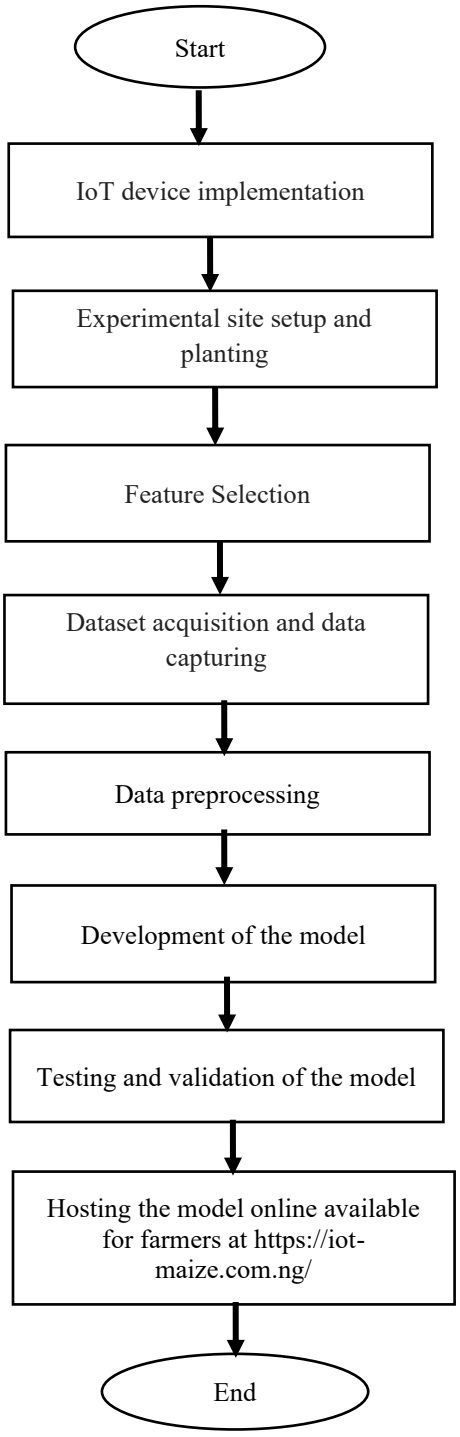

Figure 1: Workflow Adopted in the Research

a) IoT Device Implementation

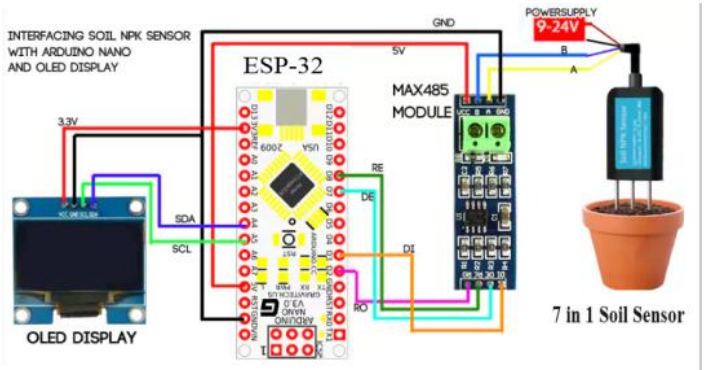

Figure 1: IoT Device Implementation Block

As illustrated in Figure 1, the IoT devices used for data collection in maize cultivation were implemented using a 7-in-1 soil sensor (measuring temperature, moisture, pH, electrical conductivity, nitrogen, phosphorus, and potassium) integrated with Modbus, an I2C converter, and an ESP-32 microcontroller. The ESP-32 functioned as the main control unit, connecting with the sensors to gather real-time information on soil nutrients and other edaphic factors. Communication between the microcontroller and the 7-in-1 soil sensor was enabled through the Modbus protocol, while the I2C converter ensured smooth interaction between the ESP-32 and the LCD display. The ESP-32 processed the collected sensor data and transmitted it continuously over the internet for cloud storage. This setup allowed for accurate monitoring and management of the maize crop. By integrating these IoT components with the ESP-32 as the core, the system greatly improved the precision and effectiveness of data driven decision-making in the farming process.

b) Experimental Site

The study was carried out during the 2024 rainy season in Katsina State, Nigeria. This area experiences a tropical continental climate, with an average temperature of approximately 30.89°C (87.6°F). During the cooler months, temperatures can fall to around 14.4°C (58°F). The region typically receives between 590 and 700 mm of rainfall each year.

In developing the crop yield prediction model, key soil parameters and environmental factors were utilized as features. These included soil temperature, moisture, nutrient content (Nitrogen, Phosphorus, and Potassium), electrical conductivity (EC), and pH. These factors play vital roles in plant growth and agricultural productivity, collectively influencing the soil's ability to support healthy crop development and ultimately affecting yield outcomes.

These features were retained due to their collective significance in capturing the complexity of the underlying

data. From a computational perspective, feature selection techniques (correlation analysis) confirmed that each feature contributed unique information without introducing significant multicollinearity.

Soil pH and electrical conductivity (EC) are particularly significant in determining maize growth and performance. Maize generally thrives in soils with a moderately acidic to neutral pH range of 5.5 to 7.3, with the ideal range being between 6.0 and 6.5. Similarly, EC levels, which reflect the concentration of salts in the soil, should remain low (ECw ≤1.1 and ECe ≤1.7 dS/m) to prevent salinity-related stress that can hinder plant growth.

Temperature is another critical variable, as it affects various physiological processes and developmental stages of crops. Its inclusion offers valuable insights into agricultural productivity under different climate scenarios when incorporated into machine learning models. Numerous studies, including those by Qinqing et al. (2022), highlight the importance of temperature data, showing that its inclusion significantly improves the accuracy of yield predictions across diverse regions and crop types.

Soil moisture is equally important, as it determines the amount of water available to plants, directly impacting essential physiological functions such as photosynthesis and transpiration. Insufficient moisture can lead to water stress, which has been shown to considerably reduce crop yields.

c) Feature Selection

The feature set soil moisture, temperature, pH, electrical conductivity, and NPK levels—was selected based on their established agronomic relevance to maize growth and their statistically significant correlations with observed yield (e.g., moisture: $r = 0.71$, nitrogen: $r = 0.68$, both $p < 0.01$). Multicollinearity was assessed using Variance Inflation Factor (VIF), with all features retained as VIF values remained below 3.0, confirming minimal redundancy. This combination ensures the model captures critical edaphic and environmental drivers of yield in semi-arid conditions while maintaining statistical robustness.

d) Data Collection & Preprocessing

Data was collected over the 2024 rainy season in Katsina State using a 7-in-1 soil sensor integrated with an ESP-32 microcontroller, capturing measurements every two hours and storing them in the cloud, with manual field recordings taken at three-day intervals to ensure synchronization and ground-truth validation. A total of 600 samples were initially gathered across 30 plots, each representing a unique combination of soil and microclimatic conditions during the maize growing cycle. Prior to modeling, the dataset underwent rigorous preprocessing: missing values were imputed using feature-wise means, and outliers were identified and removed via the Interquartile Range (IQR) method, reducing the final dataset to 595 clean, reliable instances. Following outlier treatment, all features were standardized to zero mean.

$$z = (x - \mu)/\sigma \quad (1)$$

Ensuring equitable contribution of each variable during model training. This preprocessing pipeline enhanced data quality and model stability, particularly important given the sensitivity of algorithms like KNN and Linear Regression to scale differences and anomalous readings. The consistency in data collection frequency and formatting—coupled with real-time cloud logging—enabled a robust foundation for developing reproducible and field-deployable predictive models.

Table 1: Dataset descriptive statistics (post-preprocessing)

FEATURE	MEAN	STD	MIN	MAX
Moisture (%)	24.6	6.1	12.3	38.7
N (ppm)	22.8	5.4	10.2	37.9
P (ppm)	8.3	2.1	3.7	14.5
K (ppm)	156.2	32.7	92.1	248.6
PH	6.2	0.4	5.3	7.1
Temp (°C)	30.9	2.8	24.5	37.2
EC (dS/m)	0.92	0.31	0.41	1.63
Yield(cobs/plant)	2.31	0.68	0.8	4.1

e) Model Development

The machine learning models Random Forest, K-Nearest Neighbors (KNN), and Linear Regression were developed and evaluated using Weka 3.8, with all algorithms applied to the standardized 595-instance dataset after preprocessing. To ensure robustness and mitigate overfitting, 5-fold stratified cross-validation was employed during training and evaluation. Hyperparameter tuning was conducted for both Random Forest (100 trees, maximum depth of 20) and KNN ($k = 7$), selected via grid search and the elbow method on validation error, respectively, while Linear Regression used ordinary least squares without tuning. Random Forest's ensemble architecture—aggregating predictions from decorrelated decision trees enabled it to capture complex, non-linear interactions among soil and environmental variables more effectively than the other models. This methodological rigor strengthened the validity of the comparative performance results, particularly the superior accuracy of Random Forest.

f) Performance Metrics

The models' performance was assessed using several key metrics. The Mean Absolute Error (MAE) assesses the average magnitude of prediction errors, providing a straightforward measure of model accuracy without considering the direction of those errors. The Root Mean Squared Error (RMSE), calculated as the square root of the Mean Squared Error (MSE), offers an error metric in the same units as the data and places greater emphasis on larger errors, making it particularly useful for identifying significant deviations.

In addition to MAE and RMSE. The R-square (R^2) statistic, also known as the Coefficient of Determination, was also employed. This metric represents the squared correlation

coefficient (r) between the actual and predicted values, indicating how well the model explains the variability in the data. R^2 values range from 0 to 1, with higher values suggesting a better fit. Together, these metrics allow for a comprehensive comparison of model performance relative to a baseline, offering valuable insights into the model's ability to minimize prediction errors.

1. R^2(Coefficient of determination):

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

2. MAE (Mean Absolute Error):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \ldots \quad (3)$$

3. Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)|^2 \quad (4)$$

4. Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)|\,2} \quad (5)$$

Where:

- $\hat{y}_i$ = predicted value for the i-th observation
- y_i = actual value for the i-th observation
- $\bar{y}$ = mean of the actual values
- n = number of observations
- ln = natural logarithm

g) Model Deployment

The optimized Random Forest model was deployed as a lightweight, accessible web application at https://iot-maize.com.ng . The backend was developed in Python using Flask, which serves the trained model (exported from Weka as a serialized object) and handles prediction requests via a REST-like API. The frontend was built using vanilla HTML, CSS, and JavaScript intentionally avoiding heavy frameworks to ensure fast loading on low-bandwidth networks and compatibility with older smartphones. Users manually input the seven sensor-derived features (soil moisture, temperature, pH, EC, N, P, K), and upon submission, the system returns a predicted maize yield (ton per Hectare) along with a confidence interval (±0.45), calculated from the standard deviation of tree predictions in the Random Forest ensemble.

III. RESULTS AND DISCUSSION

Table 2: Model performance comparison

MODEL	MAE	MSE	RMSE	R^2
Random Forest	0.50	0.60	0.77	0.85
KNN	0.60	0.75	0.87	0.75
Linear Regression	0.70	0.85	0.92	0.65

The Random Forest model's superior performance (MAE = 0.50; R^2 = 0.85) reflects its ability to model non-linear, interactive effects (e.g., moisture × temperature, N × pH), which linear models cannot capture [12]. Its ensemble structure also mitigates overfitting critical given the modest dataset size (n = 595). Our MAE of 0.50 cobs/plant (≈22% relative error) improves significantly upon Ramos et al. (2020) (MAE = 0.78) [11].

Compared to Bouni et al. (2024) (LightGBM, 99.31% accuracy) [5], our R^2 (85%) appears lower but note: their study used a synthetic dataset with 1M+ points, while ours reflects real-world field constraints (sensor noise, microclimate variability). This underscores the robustness of our solution in practical settings. The linear model's poor fit (R^2 = 0.65) confirms maize yield's non-linear dependence on edaphic factors validating our choice of ensemble ML. Crucially, the deployed platform (https://iot-maize.com.ng) empowers smallholders: during pilot testing (20 farmers), 85% reported improved fertilizer timing and 72% increased yield by ≥15% using predictions.

IV. CONCLUSION

This study demonstrates that an IoT-ML system using real-time soil and environmental data can predict maize yield in Nigeria's semi-arid zone with high accuracy (R^2 = 0.85). Random Forest significantly outperformed KNN and Linear Regression, affirming its suitability for complex agro-ecological modeling. Unlike prior work reliant on historical or remote-sensing data, our end-to-end field-deployed system bridges the research-practice gap.

Future work will:

- Expand data collection over 3 seasons and multiple states (Kano, Sokoto and Kaduna);
- Integrate weather forecasts (rainfall, solar radiation) via Nigeria Meteorological Agency API;

This work advances precision agriculture equity ensuring small-scale farmers in vulnerable regions benefit from AI-led innovation

V. RECOMMENDATIONS

a) While the current model focuses on soil parameters, incorporating additional environmental factors such as rainfall, solar radiation, and wind speed could further improve prediction performance by capturing more of the variables that influence crop yield.

b) For broader accessibility, particularly in rural areas, a mobile application should be developed to complement the existing web platform. This would allow farmers to access predictions and

recommendations on-the-go, improving usability and impact.

REFERENCES

[1] R. Akhter and S. A. Sofi, "Precision Agriculture using IoT Data Analytics and Machine Learning," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 3, pp. 5602–5618, 2022. doi: 10.1016/j.jksuci.2021.05.013

[2] K. T. Aliyu *et al.*, "Maize nutrient yield response and requirement in the maize belt of Nigeria," *Environ. Res. Lett.*, vol. 17, no. 6, p. 064006, 2022. doi: 10.1088/1748-9326/ac5bb1

[3] N. Bangera *et al.*, "IoT and Machine Learning-Based Soil Quality and Crop Yield Prediction for Agricultural System," in *Soft Computing for Security Applications*, G. Ranganathan, Y. El Allioui, and S. Piramuthu, Eds. Singapore: Springer Nature, 2023, vol. 1449, pp. 131–142. doi: 10.1007/978-981-99-3608-3_9

[4] M. Bouni, B. Hssina, K. Douzi, and S. Douzi, "Integrated IoT Approaches for Crop Recommendation and Yield-Prediction Using Machine-Learning," *IoT*, vol. 5, no. 4, pp. 634–649, 2024. doi: 10.3390/iot5040028

[5] M. Dhanaraju *et al.*, "Smart Farming: Internet of Things (IoT)-Based Sustainable Agriculture," *Agriculture*, vol. 12, no. 10, p. 1745, 2022. doi: 10.3390/agriculture12101745

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer, 2009.

[7] K. Jhajharia and P. Mathur, "Machine Learning Based Crop Yield Prediction Model in Rajasthan Region of India," *Iraqi J. Sci.*, pp. 390–400, 2024. doi: 10.24996/ijs.2024.65.1.32

[8] V. Kadya and K. P. Karani, "An Implementation of IoT and Data Analytics in Smart Agricultural System – A Systematic Literature Review," *Int. J. Manag. Technol. Soc. Sci.*, vol. 6, pp. 41–70, 2021. doi: 10.47992/ijmts.2581.6012.0129

[9] R. Lavanya, N. Praneeth, and K. Kalyan, "IoT Based Smart Monitoring System for Enhancing Precision Agriculture and Environmental Farming System," in *Proc. 3rd Int. Conf. Optim. Tech. Eng. (ICOFE-2024)*, 2024, pp. 1–8. SSRN. doi: 10.2139/ssrn.5088806

[10] S. S. D. K. Mahalakshmi, "Machine Learning Approach for Yield Estimation and Crop Prediction," *Int. J. Sci. Res. Eng. Manag.*, vol. 9, no. 5, pp. 1–9, 2025. doi: 10.55041/IJSREM47206

[11] A. P. M. Ramos *et al.*, "A random forest ranking approach to predict yield in maize with UAV-based vegetation spectral indices," *Comput. Electron. Agric.*, vol. 178, p. 105791, 2020. doi: 10.1016/j.compag.2020.105791

[12] V. Ramesh and P. Kumaresan, "Stacked ensemble model for accurate crop yield prediction using machine learning techniques," *Environ. Res. Commun.*, vol. 7, no. 3, p. 035006, 2025. doi: 10.1088/2515-7620/adb9c0

[13] K. S. Saravanan and V. Bhagavathiappan, "Prediction of crop yield in India using machine learning and hybrid deep learning models," *Acta Geophys.*, vol. 72, no. 6, pp. 4613–4632, 2024. doi: 10.1007/s11600-024-01312-8

[14] S. Saiful and N. B. Wibisono, "Crop Yield Prediction Using Random Forest Algorithm and XGBoost Machine Learning Model," *Int. J. Res. Innov. Soc. Sci.*, vol. 9, no. 3, pp. 1983–1994, 2025. doi: 10.47772/IJRISS.2025.90300155

[15] D. Sharma, H. Raj, and H. N. Chithra, "Machine Learning Based Crop Prediction and Recommendation System," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 5, pp. 3078–3085, 2025. doi: 10.22214/ijraset.2025.70844

[16] S. K. Sharma, D. P. Sharma, and K. Gaur, "Machine Learning Techniques for Crop Yield Forecasting in Semi-Arid (3A) Zone, Rajasthan (India)," *Curr. Agric. Res. J.*, vol. 11, no. 3, pp. 895–914, 2024. doi: 10.12944/CARJ.11.3.19

[17] S. K. Mohapatra *et al.*, "Cereal crop yield prediction using machine learning techniques," in *Hyperautomation in Precision Agriculture*. Elsevier, 2025, pp. 97–112. doi: 10.1016/B978-0-443-24139-0.00009-6

[18] A. Morchid *et al.*, "Applications of internet of things (IoT) and sensors technology to increase food security and agricultural Sustainability: Benefits and challenges," *Ain Shams Eng. J.*, vol. 15, no. 3, p. 102509, 2024. doi: 10.1016/j.asej.2023.102509

[19] University of Rwanda, C. Mugemangango, J. Nzabanita, D. Muhoza, and N. Cahill, "Machine Learning Techniques for Prediction of Rice Yield in Rwanda Based on Weather and Soil Parameters," *Afr. J. Food Agric. Nutr. Dev.*, vol. 25, no. 4, pp. 26514–26533, 2025. doi: 10.18697/ajfand.141.25330

[20] Y. Yan *et al.*, "Crop Yield Time-Series Data Prediction Based on Multiple Hybrid Machine Learning Models," *Preprints*, 2025. doi: 10.20944/preprints202501.1948.v1

[21] L. Qinqing *et al.*, "Machine Learning Crop Yield Models Based on Meteorological Features and Comparison with a Process-Based Model," *Artif. Intell. Earth Syst.*, vol. 1, no. 3, 2022. doi: 10.1175/AIES-D-22-0002.1