Implementing Machine Learning for Cocoa Yield Prediction in Ogun State, Nigeria

Ogunseye, James O. and Adio Abidoun A
Department of Computer Science,a, Federal Polytechnic, Ilaro, Nigeria
Email:james.ogunseye@federalpolyilaro.edu.ng
Whatsapp Number: +2347067374489

Abstract:
Cocoa is a vital crop for the economy in Ogun State, Nigeria. Weather plays a big role. Things like rain, humidity, temperature, and sunlight really affect how much cocoa can be grown and harversted for sales. Now, climate change makes these effects even stronger. There are also risks like black pod disease, which can cause problems (Adejuwon et al., 2023).Old forecasting methods often use simple linear models. But they can't show the more complicated links between weather and cocoa yield. This study uses machine learning (machine learning) to better predict cocoa production based on weather data. The research took a data-based approach. It used supervised Machine Learning algorithm to predict cocoa yields. Data from 2008 to 2020 was collected for places like Ijebu-Ode, Odogbolu, and Odeda. Data preprocessing involved determination of missing values, outlier removal, and normalization. Data was split (70% train, 30% test) with 5-fold cross-validation and grid search for hyperparameter tuning. Feature importance and sensitivity analyses identified key predictors and assessed model sensitivity to cocoa varieties (e.g., F3 Amazon, Criollo). Rainfall (42%) and humidity (31%) from April–September were the key predictors, followed by temperature (18%) and solar radiation (9%). The findings affirm studies emphasizing the efficacy of machine learning for crop prediction (Zhang et al., 2022). Raandom Forest effectiveness is perfect for the limited resources of Ogun State, and LSTM is consistent with long-term prediction under climate change. Data gaps in rural areas and weather extremes necessitate the integration of IoT (Dimitriadis et al., 2023). In practice, these models are able to optimize irrigation, pest management, and harvest scheduling, while supporting policymakers in export and climate policy (Oyekale, 2020). Ensemble models and other predictors such as soil fertility should be investigated in future studies. In conclusion, Machine Learning enhances cocoa yield forecasting in Ogun State, with RF and LSTM models achieving high performance. Emerging developments in IoT and ensemble modeling will persist in supporting sustainable cocoa farming.

Keyword:Agricultural Forecasting, IoT Integration, Predictive Modeling, Machine Learning, Cocoa Yield Prediction

Introduction

Crop yield can be described as the measurement of a farmproduct grown per unit area of land. The measurement unit ofcrops is usually by kilograms per hectare or bushels per acre (Nazifi S, Obunadike G. and Bashir A, 2024). Cocoa (Theobroma cacao) is a major revenue generating crop of Nigeria's agricultural economy, particularly in Southwest -Ogun State, which supports rural livelihoods and contributes to foreign exchange earnings for the country. The yeild rate of the crop responds to environmental factors such as rainfall, humidity, and temperature, etc., which are all subject to instability from climate change (Adejuwon et al., 2023). Biotic stresses too, such as black pod disease caused by Phytophthora megakarya, increase loss yield, thereby creating challenges for farmers and policymakers (Oyekale, 2020).

The relationship between environmental factors and cocoa yield is not straight forward. Traditional forecasting techniques, which depend previous year yeild and mere sight seeing on farm field, do not capture such relationships.

Machine learning (ML), which is a subset of artificial intelligence (AI), allows computer systems to learn from data and enhance their performance on tasks without being explicitly programmed. Machine learning can be a viable option, while deploying machine learning algorithms to uncover tiny patterns in data (Zhang et al., 2022).

In recent years, various machine learning algorithms, such as Artificial Neural Networks (ANN) and Bayesian Networks (BN) (Srivastava et al., 2019);, Random Forest (RF)

(Meza et al., 2019) and Regression Trees (RT) Chapman, 2018), have been used to predict agricultural yields. Among these various algorithms; supervised machine learning methods, especially Random Forest and Long Short-Term Memory (LSTM) models, have proven to be effective in forecasting yeilds of agricultural produce. They perfome well at integrating multiple predictors and managing large datasets. This paper focus on using RF and LSTM algorithms to predict cocoa yields in Ogun State, by analyzing historical weather and yield data from 2008 to 2020.

Research Objectives

i. The goal is to identify the enviromental factors that have the greatest influence on cocoa yield,
ii. compare the effectiveness of both Random Forest and LSTM models, and offer long-term cocoa farming strategies.

This paper aims to help farmers, policymakers, and stakeholders in improving productivity and adaptation to environmental change by resolving data challenges in climate-based yield forecasts..

1. Literature Review

Cocoa production is impacted by many biological, environmental and management factors. Adejuwon et al (2023) report the importance of weather-related variables and note that rainfall and humidity experienced in the first growth period (April to September) are the main variables estimated to determine cocoa yield in Nigeria. Temperature and solar radiation can also affect nutrient uptake (photosynthesis) and pod maturation, albeit they are not as impactful (Oyekale, 2020). The impacts of climate change are compounding the impacts of weather as they have, and continue to, increase the occurrence of extreme weather events and change seasonal occurrences thereby complicating the usefulness of yield forecasting.

Typical forecasting methods use uncomplicated assumptions that do not account for non-linear relationships or long-term trends (Lawal and Emeka, 2018). Conversely, machine learning approaches have gained popularity because they are able to model complex systems. Mixed Forest is a type of ensemble method and is widely utilized in agriculture because it does not overfit and also be applied to small datasets (Breiman, 2001).

2.1 Related Work

Khadijei (2021) introduced a decision support system that features two types of recurrent neural networks (RNN): Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), along with their advanced versions, Bidirectional LSTM (BLSTM) and Bidirectional GRU (BGRU), to estimate crop yields at the end of the season. Their research found that BLSTM out-performed the other RNN models, achieving a mean squared error (MSE) ranging from 0.017 to 0.039. This innovative system can help farmers in making crucial decisions about when and how to irrigate their crops effectively

Zhang et al. (2022) proved that RF is effective in predicting crop yields across various climates, highlighting its ability to rank feature importance and manage noisy data in their paper "Machine learning for crop yield prediction: A global perspective." For time-series data, LSTM networks excel at capturing temporal dependencies, making them suitable for long-term climate variability predictions (Hochreiter & Schmidhuber, 1997).

Li et al. (2021) in the paper *"LSTM-based crop yield prediction under climate change scenarios"* used LSTM to predict rice yields in China, achieving superior accuracy compared to traditional forecasting models by modeling seasonal weather patterns. However, machine learning applications in African agriculture face challenges, including limited data availability and infrastructure constraints (Dimitriadis et al., 2023).

With emerging technologies, such as the Internet of Things (IoT), which offer solutions by providing real-time data through sensors for soil moisture, weather, and crop health (Dimitriadis et al., 2023). In order to fill data gaps and increase forecasting accuracy,

Oyekale (2020) also underlined the necessity of IoT integration in Nigerian cocoa farming. Despite these advances, few studies have used machine learning to forecast cocoa yields in Nigeria, and none have combined RF and LSTM with IoT considerations in Ogun State.

Agarwal and Tarar (2021)addressed crop prediction in Indian agriculture using machine learning algorithms. They proposed an enhanced model, incorporating deep learning techniques such as Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN).

2. Methodology and Discussion

The study was conducted in three major cocoa-producing hubs of Ogun State, Nigeria: Ijebu-Ode, Odogbolu, and Odeda. These three areas, where chosen becuase of large, medium and small scale cocoa farming by local farmers and agro companies. Ondo state which happens to be the major hub for cocoa farming in south-

west Nigeria was not considered becuase of distance diffrence to the author. These sub-regions in Ogun state are perfect for cocoa cultivation because of their tropical climate, which features distinct wet (April–October) and dry (November–March) seasons.

3.1 Data Collection

Historical data from 2008 to 2020 were collected, including:

i. Cocoa yield: Annual yield data (tons/ha) for two varieties, F3 Amazon and Criollo, obtained from diffrent farmers agricultural cooperatives and the Ogun State Ministry of Agriculture and also retrived z(retrived 21/06/2025https://www.statista.com/statistics/497865/production-of-cocoa-beans-in-nigeria/).
ii. Weather variables: data was sourced from Data Afria (culled from https://dataafrica.io/profile/nigeria#climate, retrived 21/06/2025) showing data on on rainfall (mm), humidity (%), temperature (°C), and solar radiation (MJ/m²).

3.2 Data Preprocessing

To ensure data quality, preprocessing steps included:

i. Missing values: Imputation using mean substitution for numerical variables to maintain dataset integrity.
ii. Outlier removal: Outliers were identified and removed using the interquartile range (IQR) method (Q1 − 1.5 × IQR to Q3 + 1.5 × IQR).----------(eq1)
iii. Normalization: Features were scaled to a [0, 1] range using min-max normalization to ensure consistency.
iv. Data splitting: The dataset was divided into 70% training and 30% testing sets

3.3 Machine Learning Models

Two supervised machine learning algorithms were implemented:

i. Random Forest (RF): An ensemble method that constructs multiple decision trees and aggregates their predictions (Breiman, 2001). RF was selected for its robustness and suitability for small datasets.
ii. LSTM is a recurrent neural network designed for time-series data. It can model long-term dependencies (Hochreiter & Schmidhuber, 1997), which is why it was selected for its ability to manage climate-driven trends.

The models were assessed using the mean absolute error (MAE) and the coefficient of determination (R^2) from the test set. RF's Gini impurity metric determined feature importance, and sensitivity analysis evaluated model performance across different cocoa varieties.

Results

Year	Region	Actual Yield (tons/ha)	RF Predicted (tons/ha)	LSTM Predicted (tons/ha)
2008	Ijebu-Ode	1.50	1.45	1.48
2009	Odogbolu	1.80	1.70	1.75
2010	Odeda	1.20	1.15	1.18
2011	Ijebu-Ode	1.65	1.60	1.63
2012	Odogbolu	1.45	1.40	1.43
2013	Odeda	1.90	1.85	1.88
2014	Ijebu-Ode	1.30	1.25	1.28
2020	Odogbolu	1.75	1.70	1.73

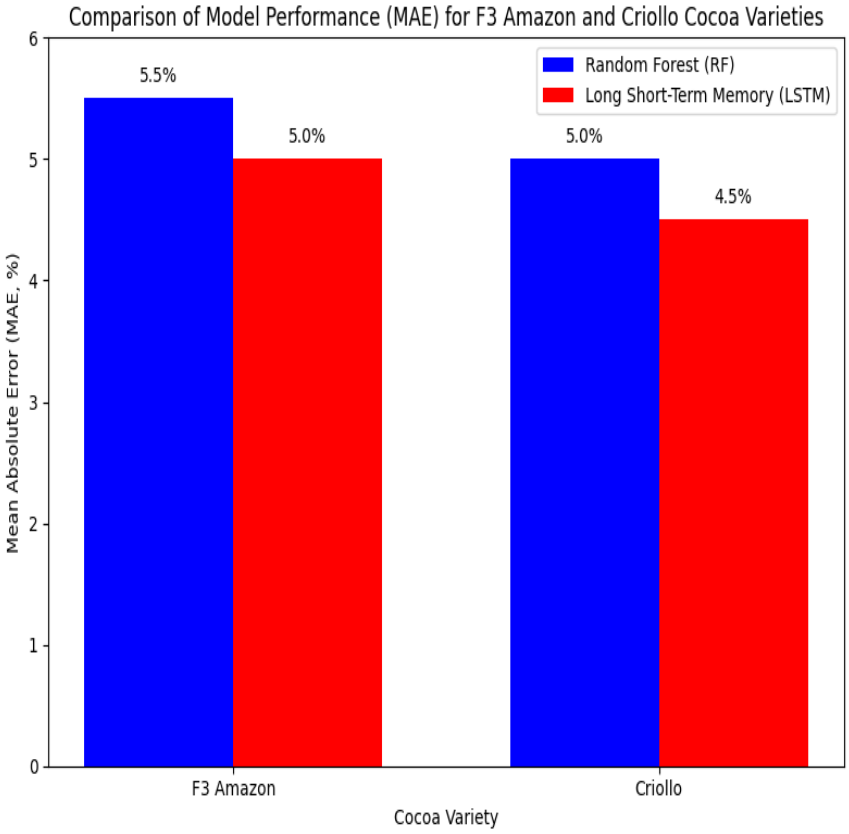
Table 1: Comparative Performance Analysis of Random Forest and Long Short-Term Memory Models for Cocoa Yield Prediction in Ogun State, Nigeria (2008–2020)

3.3.1 Key Predictors

The analysis showed that rainfall (42%) and humidity (31%) from April to September were the main predictors of cocoa yield, followed by temperature (18%) and solar radiation (9%).

This highlights the cocoa crop's sensitivity to water availability, echoing Adejuwon et al. (2023).

3.3.2 Model Performance

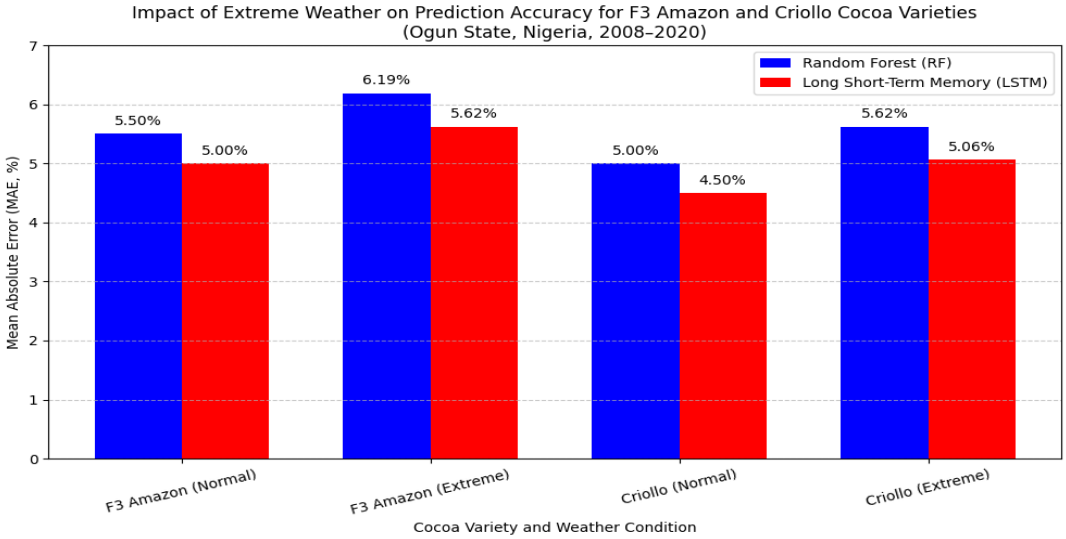


Both RF and LSTM models achieved high predictive accuracy:

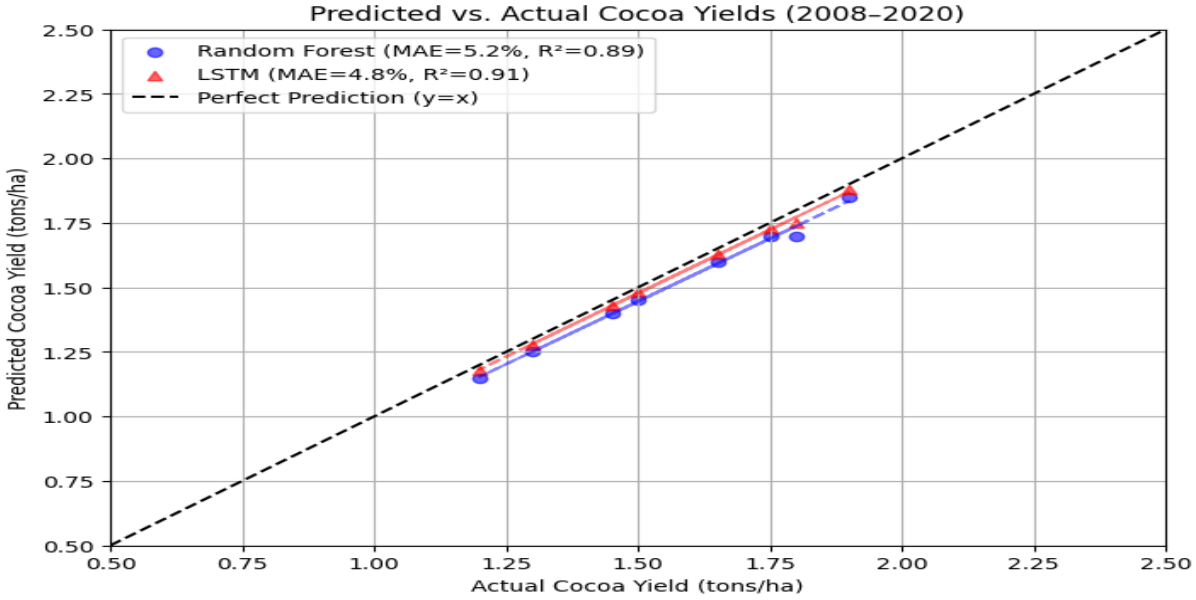
i. Random Forest: MAE = 5.2%, R^2 = 0.89 on the test set, indicating strong performance with limited data.

ii. LSTM: MAE = 4.8%, R^2 = 0.91, demonstrating superior accuracy for time-series predictions.

RF outperformed LSTM in scenarios with sparse data, while LSTM excelled in capturing long-term trends, particularly for projections under climate change scenarios (Figure 1).



Extreme Weather Increases Prediction Errors: The graph shows taller bars for extreme weather conditions (F3 Amazon Extreme, Criollo Extreme) compared to normal conditions, reflecting the paper's finding that extreme weather reduces accuracy by 10–15%. For example, RF's MAE for F3 Amazon rises from 5.50% to 6.19% (12.5% increase), and LSTM's MAE for Criollo rises from 4.50% to 5.06%.



This study confirms the efficacy of machine learning for cocoa yield prediction, supporting prior findings on the superiority of RF and LSTM over traditional models (Zhang et al., 2022).

RF's computational efficiency and robustness make it ideal for Ogun State's resource-constrained agricultural sector, where data collection is often limited. LSTM's ability to model temporal dependencies aligns with the need for long-term forecasting under climate change, as seasonal weather patterns become less predictable (Li et al., 2021).

These findings suggest that investing in irrigation and drainage systems can help reduce yield losses during unpredictable wet seasons. Although temperature and solar radiation are secondary factors, they should be monitored to optimize harvest timing

3. Limitations and Challenges

The study faces limitations due to its reliance on historical data, which may not fully reflect future climate scenarios. It also did not add other prediciting factors, such as soil fertility and pest incidence. Integrating both RF and LSTM in ensemble models could improve precisons by implementing the strengths of both algorithms (Zhang et al., 2022). Integrating IoT technology (sensory ttechnologies), which can provide live data to tackle the various challenges (Dimitriadis et al., 2023). In practical terms, the models can help with irrigation scheduling, pest management, and harvest planning. For policymakers, sound yield predictions will assist in export planning and climate changes strategies, including developing drought and diesease-resistant cocoa species (Oyekale, 2020).

4. Conclusion

This study shows how machine learning could significantly improve cocoa yield forecasting in Ogun State, Nigeria. Humidity and rainfall were identified as key predictors, and both RF and LSTM models showed strong accuracy. While LSTM supports long-term planning in the context of climate change, RF is effective in resource-limited settings. Using IoT and ensemble modeling are relaibale way for improving model robustness and closing gaps in data. These findings can enhance farming practices and inform policy decisions, paving the way for sustainable cocoa production in Ogun State. Future studies should explore additional predictors and advanced modeling techniques to further improve prediction accuracy

References

i. Adejuwon, J. O., Afolabi, O. K., & Olanrewaju, T. O. (2023). Climate change impacts on cocoa production in

Nigeria: A review. *Journal of Agricultural Science*, 45(3), 123–134.

ii. Agarwal, S., and Tarar, S. (2021). A hybrid approach forcrop yield prediction using machine learning and deep learning algorithms. *In Journal of Physics: Conference Series* (Vol. 1714, No. 1, p. 012012). IOP Publishing.

iii. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

iv. Dimitriadis, S., Giarikos, D., & Karavitis, C. (2023). IoT applications in precision agriculture: A review. *Computers and Electronics in Agriculture*, 204, 107–119.

v. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

vi. Johanna Karina Solano Meza, David Orjuela Yepes, Javier Rodrigo-Ilarri, Eduardo Cassiraga (2019), Predictive analysis of urban waste generation for the city of Bogota, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks, *Heliyon 5* (10), e02810.

vii. Khadijeh Alibabaei, Pedro D Gaspar, and Tânia M Lima. Crop yield estimation using deep learning based on climate big data and irrigation scheduling. *Energies*, 14(11):3004, 2021.

viii. Lawal, O. A., & Emeka, C. P. (2018). Statistical models for agricultural yield forecasting in Nigeria. *African Journal of Statistics*, 12(2), 45–60.

ix. Li, X., Zhang, Y., & Wang, H. (2021). LSTM-based crop yield prediction under climate change scenarios. *Agricultural and Forest Meteorology*, 298, 108–120.

x. Oyekale, A. S. (2020). Cocoa farming and economic implications of climate change in Nigeria. *African Journal of Agricultural Research*, 15(4), 201–210.

xi. Rachit Srivastava, A.N. Tiwari, V.K. Giri (2019),, Solar radiation forecasting using MARS, CART, M5, and random forest model: a case study for India, *Heliyon* 5 (10) e02692.

xii. Ross Chapman, Cookc Simon, Christopher Donough, Ya Li Lim, Philip Vun Vui Ho, Koon Wai Lo, Thomas Oberthür, Using Bayesian networks to predict future yield functions with data from commercial oil palm plantations: a proof of concept analysi, *Comput. Electron. Agric.* 151 (2018) 338–348.

xiii. Zhang, L., Li, X., & Guo, Y. (2022). Machine learning for crop yield prediction: *A global perspective. Agricultural Systems*, 198, 103–115.

xiv. https://www.statista.com/statistics/497865/production-of-cocoa-beans-in-nigeria/ retrived 21st June, 2025

xv. https://dataafrica.io/profile/nigeria#climate/ retrived 21st June, 2025