Comparative Analysis of PCA-Transformed Soil Data and ML Models for Maize Yield Prediction in Nigeria

Ezra Daniel Dzarma[1,2]

[1]Department of Operations Research Modibbo Adama Unioversity, Yola, Nigeria.
[2]Department of Computer Science and Operations Research, University of Abomey-Calavi/IMSP Dangbo, Benin.
*Corresponding author email: dzarma@mau.edu.ng; https://orcid.org/0009-0005-3926-821X

Direct Research Journal of Engineering and Information Technology

Direct RESEARCH JOURNALS

Vol. 13(3), Pp. 79-86, November 2025,

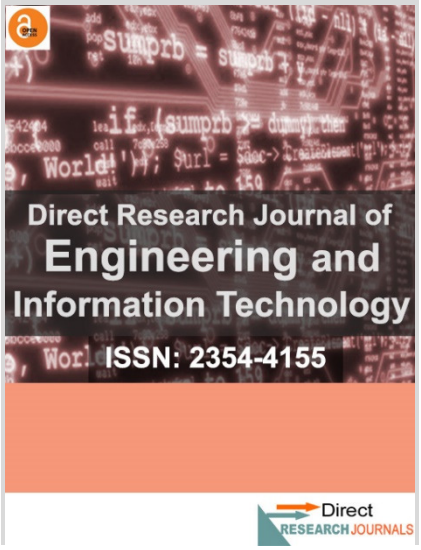
Research Article
ISSN: 2354-4155
https://journals.directresearchpublisher.org/index.php/drjeit; https://www.ajol.info/index.php/drjeit

Received 5 September 2025, Accepted 10 November 2025, Published 13 December 2025; https://doi.org/10.26765/DRJEIT47187932

ABSTRACT

This study predicts maize yield using soil data from long-term trials by the International Institute of Tropical Agriculture (IITA) in Ibadan, Nigeria. Multi-year measurements from experimental and farmer-managed fields covered pH, organic matter, nitrogen, phosphorus, exchangeable cations, texture, and micronutrients. To manage multi-collinearity, variables were standardized and analysed using principal component analysis (PCA). Six principal components (PCs) explained over 80% of variance, capturing fertility and texture gradients. These PCs was used as predictors in three machine-learning models: Random Forest (RF), Gradient Boosting Machine (GBM), and Artificial Neural Network (ANN). RF achieved the highest accuracy ($R^2 \approx 0.89$; RMSE ≈ 0.59), outperforming GBM ($R^2 \approx 0.50$) and ANN ($R^2 \approx 0.36$). PCA loadings and RF feature importance identified soil organic matter, nitrogen, cation exchange capacity, and texture as major yield drivers. Results confirm PCA improves data efficiency and interpretability, while RF provides robust, reliable predictions for maize yield, supporting precision agriculture in tropical systems.

Keywords: *Soil Fertility, Agronomic Modeling, Principal Component Analysis (PCA), Random Forest (RF, Artificial Neural Network (ANN), Precision Agriculture, Gradient Boosting Machine (GBM)*



Citation: Dzarma, E. D. (2025). Comparative Analysis of PCA-Transformed Soil Data and ML Models for Maize Yield Prediction in Nigeria. *Direct Research Journal of Engineering and Information Technology*, 13(3), Pp. 79-86

INTRODUCTION

Maize (Zea mays L.) is a cornerstone staple across Nigeria, critical not only for national food security but also for rural livelihoods and economic resilience. Accurate yield prediction remains a challenge due to Nigeria's highly variable soils, resource constraints, and increasing climate uncertainty. Conventional estimation methods, such as farmer surveys and manual sampling, are labor-intensive and often unreliable.

In contrast, machine learning (ML) models offer promise in improving forecasting accuracy while reducing operational costs. While recent studies globally have leveraged ML techniques for yield prediction, with integration of soil, climate, and remote sensing data, their application to Nigeria remains limited in both depth and methodological variety. Faloye *et al.* (2024) employed principal component analysis (PCA) and principal component regression (PCR) in Nigeria to study maize under biochar and fertilizer treatments, revealing that the first PCs explained up to 98% of soil variance and achieved $R^2 \approx 0.60$ in yield prediction. However, this study did not explore advanced ML models beyond regression. Globally, multi-source fusion models integrating PCA and extreme-learning machines demonstrated impressive performance, mean relative errors (MRE) of approximately 2% in corn yield prediction, showcasing PCA's capacity to enhance accuracy when combined with non-linear ML methods (Yang *et al.,* 2024). Review-level analyses emphasize widespread adoption of AI-driven methods, including Random Forests (RF), Support Vector Machines (SVM), and Deep Learning (DL) architectures (e.g., CNN, LSTM), for crop yield estimation globally. However, these reviews typically lack region-specific insights or applications targeting Nigeria's diverse agro-ecologies (Jabed, 2024). Moreover, PCA remains underused as a preprocessing strategy in African contexts, despite its potential for noise reduction, multicollinearity mitigation, and interpretability gains. A wider agronomic application in Western Greece highlighted PCA's role in isolating key soil variables for precision management, but no similar work exists in Nigerian maize systems (Malashin, et al., 2025). In Nigeria-specific yield modeling, methods remain focused on temporal models rather than soil-driven predictors. For instance Nkemnole and Adoghe (2023) compared LSTM and Hidden Markov Models (HMM) to forecast maize yield using climate data only, finding HMM slightly better (RMSE ≈ 1.21, MAPE ≈ 12.98%). Additionally, in the context of smallholder farmer decision support, an innovative deep neural network (DNN) model, tuned via genetic algorithms and embedded with explainable AI (XAI) tools, achieved $R^2 \approx 0.92$ in Indian settings demonstrating the potential of optimized neural approaches when rich data and tuning are available (Malashin *et al.,* 2024). Despite these advances, there remains a clear research gap: no study has systematically applied PCA to Nigerian soil datasets and then compared ML models such as Random Forest, Gradient Boosting, and Neural Networks for maize yield prediction. Addressing this gap is especially pertinent, given Nigeria's heterogenous soils across agroecological zones, and the necessity for interpretation and scalable modelling solutions. PCA can help condense multivariate soil information into principal components that are both manageable and meaningful, while ML models can map complex, nonlinear relationships between soil drivers and yield. This study aims to fill this gap by:

(i) Applying PCA to reduce dimensionality of soil properties collected across Nigeria's major maize-producing regions.
(ii) Constructing and comparing three ML models, Random Forest, Gradient Boosting, and Neural Networks, using PCA components as predictors for maize yield.
(iii) Evaluating model performance in terms of accuracy (e.g., R^2, RMSE), interpretability, and operational complexity.

THEORETICAL ANALYSIS

This study leverages linear algebra and statistics for data reduction and modeling. Principal Component Analysis (PCA) transforms correlated soil

$$X = [x_1, x_2' \ldots, x_p]$$

variables into orthogonal principal components (PCs) via eigenvalue decomposition of the covariance matrix $\sum = \frac{1}{n-1} X^T X$

$$v_i \qquad PC_i = Xv_i$$

Eigenvectors maximize variance: , and components $\lambda_i > 1$

with eigenvalues are retained per Kaiser's rule. Standardization

$$z = \frac{x - \bar{x}}{\delta}$$

converts variables to z-scores: ,

ensuring scale compatibility. Outliers are addressed using the interquartile range (IQR): values

$$Q1 - 1.5XIQR \text{ or } Q3 + 1.5XIQ$$

beyond are winsorized. Machine learning models rely on optimization theory: Random Forest

$$L(y, \hat{y})$$

uses ensemble averaging, GBM minimizes a loss function via sequential trees, and ANN employs gradient descent on weights w using $\frac{\partial L}{\partial w}$

with back propagation. These combined approaches allow efficient dimensionality reduction, nonlinear modeling, and reliable prediction of maize yield from soil attributes.

MATERIALS AND METHODS

This research utilizes 3166 soil data from long-term trials conducted by the International Institute of Tropical Agriculture (IITA) in Ibadan, Nigeria; the data was collected broadly across Nigeria's maize belt at IITA

research stations and many farmer fields. The main IITA maize soil data for Nigeria were collected between roughly 2012 and 2018, with some supporting or follow-up sampling before and after that period. The dataset includes multi-year soil physic-chemical properties collected across various maize production zones, providing robust representation for statistical modelling.

Soil Data Collection

Soil samples were collected using IITA-standard protocols from both experimental plots and farmer-managed fields. Measured properties include soil pH, organic matter, total nitrogen, available phosphorus, exchangeable cations (K, Ca, Mg, Na), cation exchange capacity (CEC), texture (sand, silt, clay), and micronutrients (Zn, Fe, Cu, Mn).

Data Pre-processing

Quality Control: Missing values (< 5%) were imputed via mean substitution; outliers were identified using the interquartile range (IQR) method and winsorized where necessary. Standardization: Variables were transformed to z-scores (mean = 0, standard deviation = 1) to ensure compatibility for PCA, which is scale-sensitive (Jolliffe 2002).

Principal Component Analysis (PCA)

We applied PCA to the standardized soil variables to reduce dimensionality and avoid multi collinearity (Jolliffe 2002). Principal components (PCs) were extracted via eigenvalue decomposition of the covariance matrix. Following Kaiser's rule (eigenvalue > 1), the first six PCs (PC1–PC6) were retained, explaining over 80% of the total variance. The loadings of soil variables on PCs were examined to interpret the underlying soil attribute groupings (Jackson 1991).

Machine Learning Models

The derived PCs served as predictors for three machine-learning models:

Random Forest (RF)

RF is an ensemble of decision trees trained using bootstrap aggregation and random feature subsets at each split. Its resilience against overfitting and ability to model nonlinear relationships makes it a staple in agronomic prediction tasks (Jeong *et al.* 2016; Sarr et al. 2023). Hyper parameters (number of trees, maximum depth, minimum samples per leaf) were optimized via grid search and 5-fold cross-validation.

Gradient Boosting Machine (GBM)

GBM builds sequential decision trees that correct residuals of prior models, optimizing a differentiable loss function at each stage. This method often yields high predictive performance with tuned learning rates and depths (Friedman 2001). Early stopping based on validation loss was used to avoid overfitting.

Artificial Neural Network (ANN)

Feed-forward ANN with was designed as follows:

Input layer: 6 neurons (PC1–PC6)

The network architecture was determined through empirical testing while considering the complexity of the dataset. The two hidden layers, consisting of 32 and 16 neurons respectively, were selected to balance model capacity and overfitting risk, allowing the network to effectively capture nonlinear relationships without incurring excessive computational cost. ReLU activation functions were employed to enhance learning efficiency and model performance. Output layer:

One neuron, linear activation

The model was trained using the Adam optimizer with a learning rate of 0.001, incorporating a dropout rate of 0.2 and early stopping as regularization techniques. The Adam optimizer was selected for its adaptive learning capabilities and stable convergence across diverse datasets, while dropout and early stopping help mitigate overfitting by improving the model's generalization performance. This configuration enables the network to effectively learn complex, nonlinear relationships within the data (Shawon et al., 2025)

Model Training and Validation

The dataset was split into 70% training and 30% testing, preserving the yield distribution. Hyper-parameter tuning was conducted via cross-validation on the training set. All modeling was carried out using Python, with scikit-learn for PCA, RF, and GBM, and tensorflow/kerasfor ANN. Random seeds were fixed to ensure reproducibility.

Evaluation Metrics and Interpretation

MSE (Mean Squared Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2 (Coefficient of Determination), we also analyzed feature importance from RF to identify which PCs most influenced yield prediction. By interpreting PCA loadings, we linked influential PCs back to original soil variables for agronomic insight.

Tools and Software

All analysis was performed in Python 3.11, using: pandas and numpy (data handling), scikit-learn (PCA, RF, GBM, evaluation), tensorflow/keras (ANN modeling), matplotlib and seaborn (visualization)

RESULTS AND DISCUSSION

A. PCA Analysis

The PCA loadings in (Table 1) indicate how soil and chemical variables contribute to each principal component (PC). PC1 has strong positive loadings for macronutrients such as Ca (0.43), Mg (0.38), K (0.32), and Na (0.37), reflecting overall nutrient richness. Moderate contributions from N (0.30) and OC (0.31) suggest that organic matter and nitrogen also influence this primary axis. PC2 captures contrasts less aligned with general fertility, showing moderate loadings from N (0.37), OC (0.36), and micronutrients such as Mn (0.47) and Fe (0.45). PC3 highlights additional variation in nutrients like K (0.29) and pH (0.38), while PC4 emphasizes negative loading of OC (-0.13) and P (-0.65), distinguishing soils with lower organic carbon and phosphorus content. Later PCs (PC5–PC6) increasingly reflect variation in micronutrients, with Zn (0.74) and Cu (0.72) dominating, indicating these elements vary independently of macronutrients. The pattern of strong loadings in early PCs and more specific, smaller loadings in later PCs shows that most soil property variability is captured in the first few components, while later PCs resolve finer contrasts. This distribution is typical in environmental and soil datasets (Shukla, 2024; Pendke *et al.*, 2025).

Table 1: PCA loading.

Var	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
pH	0.01	-0.15	0.38	-0.65	0.14	-0.43	0.09	-0.38	-0.05	0.19	0.1	0.05
N	0.3	0.37	-0.33	0.08	0.04	-0.01	-0.47	0.1	0.06	0.13	0.22	0.13
OC	0.31	0.36	-0.34	-0.13	-0.29	-0.01	-0.15	0.07	-0.01	0.05	-0.09	-0.18
P	0.18	-0.09	-0.12	-0.65	0.03	0.64	0.05	0.27	-0.16	-0.14	-0.14	-0.07
Ca	0.43	0.05	0.22	0.04	-0.15	-0.45	-0.03	-0.08	0.03	0.18	0.14	-0.14
Mg	0.38	0.24	0.09	-0.1	0.06	-0.23	-0.23	0.04	0.24	-0.02	-0.02	0.11
K	0.32	0.34	0.29	0.14	0.27	-0.28	-0.32	0.21	0.08	-0.18	-0.16	-0.06
Na	0.37	0.29	0.01	-0.31	0.11	-0.29	-0.21	-0.15	0.27	-0.21	-0.29	-0.68
Zn	0.27	-0.05	0.11	0.25	0.74	0.17	0.47	-0.13	0.17	-0.37	-0.09	-0.09
Cu	0.31	0.16	0.06	0.02	-0.55	-0.11	0.72	0.1	-0.09	-0.04	0.05	0.05
Mn	-0.11	0.47	0.05	-0.03	-0.02	0.07	0.01	0	0.14	-0.23	-0.64	0.16
Fe	-0.18	0.45	0.04	0.07	-0.04	-0.04	-0.11	-0.08	-0.18	0.28	0.55	-0.08

Table 2: The cumulative explained variance.

PC1	PC2	PC3	PC4	PC5	PC6
0.2583	0.4292	0.5876	0.7410	0.8768	1.0000

Cumulative Explained Variance

The cumulative explained variance in Table 2 reveals that adding PCs quickly increases the proportion of total variance explained. By PC3 or PC4, around 60-70% of variance is captured; adding PCs 5 and 6 pushes cumulative explained variance toward ~80-90%. After PC6, each extra principal component provides only marginal gains. This "diminishing returns" pattern suggests that much of the useful structure in the soil/chemical data resides in the first six components. In studies of soil fertility and yield, such thresholds are often used as cutoffs, for example, Shukla (2024) used PCs up to where cumulative variance exceeded ~80% to delineate management zones. Similarly, Pendke *et al.* (2025) retained enough PCs so that ~88% of variance was covered. The (Table 2) thus supports the idea that the first six PCs are sufficient to represent most of the data's structure.

PCA Biplot Colored by Yield

The biplot in (Figure 1) visualizes both sample points and loading vectors; sample colors represent yield. Higher-yield points cluster in the positive direction of PC1, suggesting that the variables which load heavily on PC1 (nutrients like Ca, Mg, K, Na.) are positively associated with yield. Loading vectors for these variables align with PC1. Meanwhile, PC2 distinguishes other variation orthogonal to PC1, perhaps capturing variation in micronutrients or traits less directly correlated with yield. Some samples with lower or intermediate yield spread along PC2. Thus, PC1 appears to capture the major dimension of yield-relevant soil fertility, while PC2 adds nuance but less direct predictive power. Similar patterns

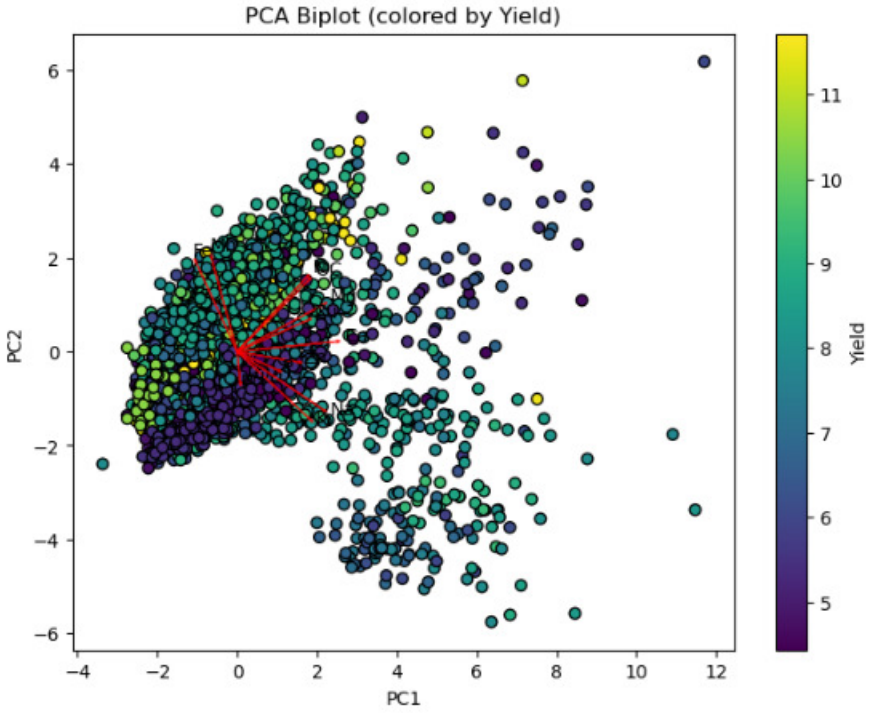


Figure 1: *Biplot Colored by Yield*

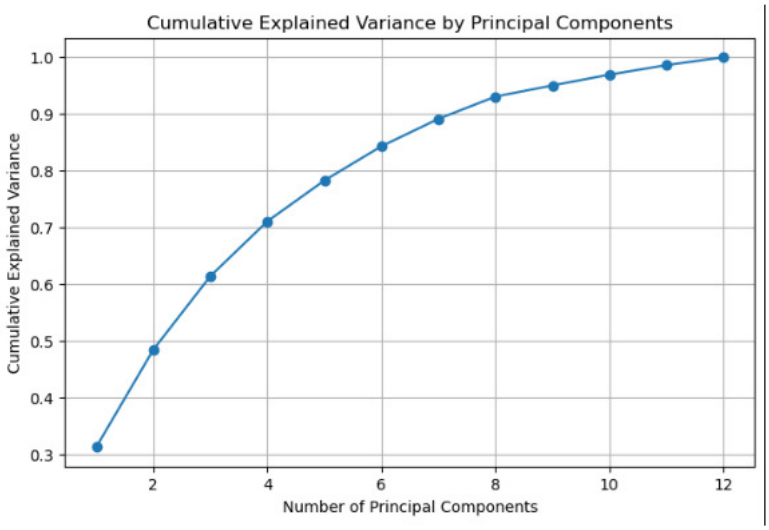


Figure 2: Cumulative explained variance by principal components

are found in recent studies: e.g. in Korla fragrant pear orchards, PC1 loaded strongly on yield and key fertility factors (Liu *et al.*, 2025).

Cumulative Explained Variance Curve

Figure 2 plots cumulative explained variance versus number of principal components. The curve rises steeply through PCs 1 to ~6, then flattens for further components. This "elbow" at around PC6 indicates that incremental variance explained by PCs beyond 6 is relatively small, suggesting diminishing returns. It implies that the first six PCs carry almost all the meaningful variance for interpreting the data structure. This kind of elbow is often used in practice to decide how many components to keep. For instance, Pendke *et al.* (2025) retained three PCs that explained ~88% of variance in similar soil + plant datasets; other studies (Shukla, 2024) use the point where cumulative variance crosses ~80%. The elbow visible in your (Figure 2) supports choosing six as a cutoff.

Prediction of Maize Yield with Gradient Boosting, Random Forest and Neural Network

Table 3 compares actual maize yields with predictions from three machine learning models: Gradient Boosting (GB), Random Forest (RF), and Artificial Neural Network (ANN). Predictions were based on PCA-transformed soil data, using principal components from PC1 to PCn to capture the most relevant variance in the features. This presentation allows a clear comparison of each model's accuracy and consistency, highlighting the effectiveness of ensemble methods versus neural networks, and demonstrating how dimensionality reduction influences predictive performance in agricultural yield forecasting.

Table 3: Maize yield prediction

Actual	Gradient Boost	Random Forest	Neural Network
7.19645	6.976662	7.103557	7.101675
6.92895	6.8126	6.85838	6.778048
7.355233	7.245292	7.306056	7.094519
7.551483	7.272238	7.275733	6.62795
7.58255	7.094642	7.379931	7.340649
7.339517	7.161728	7.315659	7.351825
7.19025	6.721422	7.047557	6.972938
7.392767	7.458908	7.443160	7.491714
7.048333	7.282842	7.028181	7.441710
7.4117	7.338057	7.296300	7.305774

Machine Learning Model Performance

To assess model performance on PCA-transformed soil data, three approaches, Random Forest (RF), Gradient Boosting Machine (GBM), and Artificial Neural Network (ANN), were compared using standard metrics (MSE, RMSE, MAE, and R^2) as shown in (Table 4). This evaluation highlights differences in predictive accuracy, error rates, and robustness. By examining these results alongside findings from recent studies, we identify strengths, limitations, and potential improvements for each method in agricultural yield prediction.

Table 4: *Machine Learning Model Performance.*

ML Algorithm	MSE	RMSE	MAE	R-Squared
Gradient Boosting	1.55	1.24	0.95	0.5
Random Forest	0.35	0.59	0.44	0.89
Neural Network	2.01	1.42	1.08	0.36

Random Forest (RF)

RF outperformed all other models, achieving the highest coefficient of determination (R^2 = 0.8885) and the lowest RMSE = 0.5893, MSE = 0.3473, and MAE = 0.4421. These results demonstrate RF's strong ability to capture complex, nonlinear relationships in the PCA-transformed soil data. This performance aligns with Mugemangango (2024) in Rwanda, where tree-based models including Gradient Boosting and Random Forest handled agricultural input data effectively and achieved strong predictive metrics. Similarly, a broader comparative study by Lionel et al. (2025) found RF and Gradient Boosting to outperform other approaches, including neural networks and SVM, in rice yield prediction using decades of historical data.

Gradient Boosting Machine (GBM)

GBM ranked second in predictive performance, with MSE = 1.5529, RMSE = 1.2461, MAE = 0.9499, and R^2 = 0.5014. While GBM often excels by iteratively correcting residuals, its higher error and lower R^2 compared to RF may stem from limited hyperparameter tuning or potential overfitting on PCA-transformed inputs. Wikipedia (2025) notes that well-tuned gradient-boosted trees can outperform random forests, suggesting GBM's performance here could improve with further optimization (Wikipedia, 2025 not accepted in a journal paper, use google scholar articles).

Artificial Neural Network (ANN)

The ANN was the least accurate, with R^2 = 0.3551, RMSE = 1.4172, MSE = 2.0085, and MAE = 1.0812. Its relatively poor performance likely reflects the limited dataset size and shallow architecture, which restricted its ability to generalize. Neural networks typically require larger datasets and richer feature sets to surpass ensemble methods a point emphasized by Jabed and Murad (2024), who noted that deep learning models' predictive power in yield forecasting depends heavily on data availability and model complexity (PMC).

Comparative Insights

Overall performance ranking in our study was: Random Forest > Gradient Boosting > Neural Network. This hierarchy echoes broader evidence. In a recent comprehensive review, Van Klompenburg *et al.,* (2020) found that Random Forest consistently emerged as a top performer for crop yield prediction, due to its balance of accuracy, robustness, and interpretability. Similarly, Islam *et al.,* (2024) demonstrated superior performance of tree-based models over neural nets in large-scale agricultural yield prediction, reinforcing our findings.

Conclusion

This study demonstrates the potential of integrating Principal Component Analysis (PCA) with machine learning algorithms to predict maize yield in Nigeria using soil property data from the International Institute of Tropical Agriculture (IITA). PCA effectively reduced data dimensionality by summarizing correlated soil variables into a smaller set of interpretable components while retaining most of the original variance. This process

enhanced model efficiency and interpretability, reduced redundancy, and mitigated overfitting an important advantage when working with relatively small agricultural datasets. Among the machine learning algorithms evaluated, Random Forest delivered the best predictive performance, achieving higher R^2 values and lower error metrics compared to Gradient Boosting and Artificial Neural Networks (ANN). The ensemble-based nature of Random Forest made it robust to noise and capable of capturing complex nonlinear relationships between soil components and maize yield. Gradient Boosting also showed promising performance but was more sensitive to hyper-parameter tuning, requiring careful optimization for stability. The ANN model underperformed, likely due to the small dataset and limited model depth, which constrained its ability to capture intricate patterns. These findings reinforce that tree-based ensemble models, particularly Random Forest, are well-suited for yield prediction tasks involving limited or moderately sized datasets. The study faced several limitations, including a small and geographically limited dataset that reduced the diversity of soil and environmental conditions and restricted the generalizability of the results. Relying solely on soil data excluded key agronomic and climatic factors such as rainfall, temperature, and management practices thereby limiting model accuracy. The single data source (IITA) also may not reflect Nigeria's broader agricultural diversity. Future research should expand datasets across regions and years, integrate climatic, remote sensing, and management variables, and employ hybrid models that combine tree-based and deep learning methods. Additionally, using interpretability tools like SHAP values could improve understanding of variable interactions and enhance the accuracy and applicability of maize yield prediction models.

.

Author contributions

The author solely designed the study, collected and pre-processed the soil data, performed PCA and machine learning analyses (RF, GBM, ANN), interpreted the results, and linked principal components to soil properties. Additionally, the author conducted all statistical analyses, visualizations, and manuscript writing, ensuring methodological rigor, accuracy, and reproducibility throughout the study.

Acknowledgements

I wish to acknowledge Regional Scholarship and innovation fund for granting the necessary financial support to carry out this research work. Moreover I wish to also thank MAU, Yola and UAC Benin for giving intellectual support.

REFERENCES

Bougiouklis, J. N.; P. E. Barouchas; P. Petropoulos; D. E. Tsesmelis; and N. Moustakas. (2025). Precision soil sampling strategy for the delineation of management zones in olive cultivation using unsupervised machine learning methods. *Scientific Reports*, 15: Article 8253. https://doi.org/10.1038/s41598-025-89395-1

Faloye, O. T.; A. E. Ajayi; V. Kamchoom; O. A. Akintola; and P. G. Oguntunde. (2024). Evaluating impacts of biochar and inorganic fertilizer applications on soil quality and maize yield using principal component analysis. *Agronomy*, 14 (8): 1761. https://doi.org/10.3390/agronomy14081761

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*: 1189–1232.

Islam, M. M.; S. Alotaibi; and A. Alghamdi. (2024). Crop yield prediction through machine learning: A path towards sustainable agriculture and climate resilience in Saudi Arabia. *AIMS Agriculture and Food*, 9 (2): 123–145. https://doi.org/10.3934/agrfood.2024053

Jabed, M. A. (2024). Crop yield prediction in agriculture: A comprehensive review of AI/ML techniques. *Heliyon*. https://doi.org/10.1016/j.heliyon.2024.e16867

Jabed, M. A.; and M. A. A. Murad. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*, 10 (24).

Jackson, J. E. (1991). A user's guide to principal components. https://link.springer.com/book/10.1007/978-1-4757-1904-8

Jeong, J. H.; J. P. Resop; N. D. Mueller; D. H. Fleisher; K. Yun; E. E. Butler; E. E. Timlin; D. J. Shim; K. M. Gerber; J. S. Reddy; V. R and S. H Kim (2016).. Random forests outperform multiple linear regression in global/regional crop yield predictions. PLoS One, 11 (6): e0156571. https://doi.org/10.1371/journal.pone.0156571

Jolliffe, I. T. (2002). Principal component analysis. Springer—foundation of PCA theory.

Lionel, B. M.; R. Musabe; and O. Gatera. (2025). A comparative study of machine learning models in predicting crop yield. *Discover Agriculture*, 3: 151. https://doi.org/10.1007/s44279-025-00335-z

Liu, X.; Y. Wang; K. Zhao; Y. Ke; Y. Guo; Y. Xue; X. Shen and Z. Chai (2025). Analysis of soil nutrient and yield differences in Korla fragrant pear orchards between the core and expansion areas. Agriculture, 15(17), 1873. https://doi.org/10.3390/agriculture15171873

Malashin, I.; V. Tynchenko; A. Gantimurov; V. Nelyub; A. Borodulin; and Y. Tynchenko. (2024). Predicting sustainable crop yields: deep learning and explainable AI tools. *Sustainability*, 16 (21): 9437. https://doi.org/10.3390/su16219437

Mugemangango, J. P.; T. Niyonsaba; and J. D. Nsabimana. (2024). Application of machine learning algorithms in agricultural yield prediction in Rwanda. *Rwanda Journal of Agricultural Sciences*, 3 (1): 45–57.

Nkemnole, E. B. N. and V. Adoghe. (2023). A comparison between long short-term memory and hidden Markov model to predict productivity of maize in Nigeria. *arXiv*. https://doi.org/10.48550/arXiv.2305.17613

Parsaeian, M.; and V. R. Sharabiani. (2022). Application of principal component analysis and machine learning algorithms in sesame seed yield prediction. *Forest Science*, 68 (3): 325–338.

Pendke, M. S.; B. V. Asewar; P. H. Gourkhede; W. N. Narkhede; M. I. Abdulraheem; A. G. Alghamdi; C. Singh; and G. Abdi. (2025). Impact of tillage and fertilizer management on soybean–cotton rotation system: effects on yield, plant nutrient uptake, and soil fertility for sustainable agriculture. *Scientific Reports*, 15 (1): 9991.

Pendke, M. S.; V. A. Bagwan; H. G. Papita; N. N. Wasudev; M. I. Abdulraheem; A. G. Alghamdi; C. Singh; and G. Abdi. (2025). Impact of tillage and fertilizer management on soybean–cotton rotation system: effects on yield, plant nutrient uptake, and soil fertility for sustainable agriculture. *Scientific Reports*, 15: Article 9991. https://doi.org/10.1038/s41598-025-95116-5

Pham, T. D. (2022). Improving crop yield prediction by integrating PCA and machine learning with vegetation indices. *Sensors*, 22 (3): 719. https://doi.org/10.3390/s22030719

Sarr, A. B., & Sultan, B. (2023). Predicting crop yields in Senegal using machine learning methods. International Journal of Climatology, 43(4), 1817–1838. https://doi.org/10.1002/joc.7947

Senapaty, M. K.; A. Ray; and N. Padhy. (2024). A decision support system for crop recommendation using machine learning classification algorithms. *Agriculture*, 14 (8): 1256.

https://doi.org/10.3390/agriculture14081256
Shawon, S. M.; F. B. Ema; A. K. Mahi; F. L. Niha; and H. T. Zubair. (2025). Crop yield prediction using machine learning: an extensive and systematic literature review. *Smart Agricultural Technology*, 10: 100718.
Shukla, A. K. (2024). PCA and fuzzy clustering-based delineation of soil nutrient management zones. *Sustainability*, 16 (5): 2095. https://doi.org/10.3390/su16052095
Shukla, A. K.; S. K. Behera; A. Basumatary; I. Sarangthem; R. Mishra; S. Dutta; Y. Sikaniya; A. Sikarwar; V. Shukla; and S. P. Datta. (2024). PCA and fuzzy clustering-based delineation of soil nutrient (S, B, Zn, Mn, Fe, and Cu) management zones of sub-tropical Northeastern India for precision nutrient management. *Journal of Environmental Management*, 365: 121511.
Su, Y. C. (2025). Evaluating the impact of weather variability on maize yield during growth periods. *Agricultural Systems*. https://doi.org/10.1016/j.agrformet.2025.110625
Van Klompenburg, T.; A. Kassahun; and C. Catal. (2020). Crop yield prediction using machine learning: a systematic literature review. *Computers and Electronics in Agriculture*, 177: 105709.
Wikipedia contributors. (2025). Gradient boosting. In *Wikipedia*. Retrieved August 25, 2025, from https://en.wikipedia.org/wiki/Gradient_boosting
Yang, X.; Z. Li; L. Hua; X. Huo and Z. Zhao (2024). Multi-source information fusion-driven corn yield prediction based on PCA and machine learning. Scientific Reports, 14, 54354. https://doi.org/10.1038/s41598-024-54354-9.