EPH – International Journal of Mathematics and Statistics
ISSN:2208-2212
https://eijms.com/index.php/ms/issue/archive

COMPARATIVE EVALUATION OF MACHINE LEARNING MODELS FOR CACAO YIELD PREDICTION USING SOIL PHYSICOCHEMICAL PROPERTIES

Ndukwu Eric Chima

Senior Lecturer, Educational Foundations, Psychology, University of Nigeria Nsukka State: Enugu, Nsukka , 410001, ORCID ID: https://orcid.org/0000-0002-5540-9968, Email ID: eric.ndukwu@unn.edu.ng

Corresponding author: eric.ndukwu@unn.edu.ng

A R T I C L E I N F O

Received: 05/07/2026
Revised: 22/09/2026
Accepted: 10/10/2026
Published: 27/10/2026

DOI: https://doi.org/10.69980/zc8fsm72
Volume:12, Issue:04, October 2026

Abstract

Cacao productivity is influenced by soil physicochemical characteristics, making accurate yield prediction essential for sustainable agricultural management. This study aimed to evaluate relationships between soil properties and cacao yield, compare machine learning models, and identify influential soil predictors. Secondary agricultural data were obtained from the publicly available Mendeley Data repository, comprising 150 cacao plantations across three climatic zones in Ghana. Following data preprocessing, 149 complete observations were included in predictive modelling. Descriptive statistics, Pearson correlation analysis, and four regression models like Linear Regression, Decision Tree Regression, Random Forest Regression, and Support Vector Regression were applied. Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). Results revealed significant correlations between cacao yield and exchangeable acidity (r = 0.601), soil pH (r = -0.600), and sand content (r = -0.504). Random Forest Regression demonstrated the best overall predictive performance, achieving an R^2 of 0.487 and RMSE of 375.07 kg/ha/year. Decision Tree Regression achieved the lowest MAE of 274.23 kg/ha/year. Feature importance analysis identified soil pH, exchangeable acidity, and sodium as influential predictors. The findings demonstrate the potential of soil-based machine learning approaches for cacao yield estimation, although moderate predictive accuracy highlights the need for additional environmental and management variables to improve agricultural prediction and decision-making.

Keywords: Cacao yield prediction; Soil physicochemical properties; Machine learning; Random Forest Regression; Precision agriculture

1. Introduction

Cacao (Theobroma cacao L.) is an economic significant agricultural crop, which plays a crucial role in the rural livelihoods, international trade and global chocolate industry. In West Africa, cocoa farming is a key enterprise, accounting for a significant portion of the world's cocoa production, especially in terms of smallholder farmers. Cacao production, however, is very variable from farming region to farming region due to variations in soil fertility, farming practices, and environmental conditions. Climatic variation, farm management and resource availability have been shown to be drivers of yield gap in smallholder cacao production systems in Ghana (Abdulai et al., 2020a).

Productive plantations of cacao trees need to be well managed, ensuring adequate soil nutrient availability, growth of the trees and sustainability of the crops. There is therefore a need to improve the soil health to help minimize yield gaps and limit the sustainable production of cacao beans (Amponsah-Doku et al., 2022). Also, agroforestry systems that incorporated appropriate shade trees have been shown to be productive under low input farming systems (Asitoakor et al., 2022).

The soil physicochemical properties are factors that have significant influence on the growth of cacao as they facilitate the availability of nutrients, root development, water retention and biological processes in agricultural soils. Soil pH affects nutrient solubility and availability, organic carbon and organic matter affect soil structure, microbial activity and nutrient cycling. Nitrogen is critical for vegetative growth, while phosphorus and potassium are important for root growth, crop physiological functions and yields. Soil properties can interact with ecological factors to affect cacao responses to soils (Doe et al., 2023). Soil fertility and nutrient availability in cacao agroforestry systems are also influenced by the different management practices of carbon and nutrient cycling processes (Agbotui et al., 2024). Balanced nutrient management is crucial as applying fertilizer has been linked to cacao pod development, nutrient accumulation and yield changes (Goudsmit et al., 2023). Soil moisture and its supply influence cacao productivity as well as availability of rainfall, especially when water is scarce during the rainy season (Adet et al., 2024).

In addition to the individual soil nutrients, soil texture, nutrient holding ability and soil associational factors of physical and chemical properties affect the productivity of cacao. Variations in sand and silt (clay) proportions can influence water-holding capacity, drainage and nutrient storage, which can be factors in yield differences among plantations. Agroforestry management practices can impact cacao production via interactions between vegetation, environment and resource availability (Konaté et al., 2024). The microbial resource availability and rhizosphere microbial communities are also strongly linked to the quality of soils in Ghanaian cacao farms (Schmidt et al., 2024). Cacao pod husk derived organic soil amendments can have a significant effect on soil fertility and nutrient mobility which highlights the importance of soil management strategies to ensure productive agricultural systems (Mwafulirwa et al., 2024).

The use of machine learning (ML) has proven to be a valuable technique to foretell agricultural productivity using environmental, climatic, and soil related factors. Machine learning algorithms can detect complex relationships, nonlinear interactions, and many more predictors than the traditional statistical methods in which one primarily considers the relations between a limited number of variables. Machine learning methods have been previously shown to be transferable to crop yield prediction over different agriculture systems (van Klompenburg et al., 2020). Linear Regression (LR) is a basic method to estimate the relationship between the predictors and yield, while Decision Tree Regression (DTR) discovers the nonlinear patterns by hierarchical decision rules. Random Forest Regression (RFR) is a collection of decision trees to create a more stable prediction model, and Support Vector Regression (SVR) is a regression model that uses kernel-based methods to model complex relationships. These techniques have been growing in application in the field of agriculture forecasting and sustainable agriculture management over the recent years (Jabed & Murad, 2024). Comparative evaluation of machine learning algorithms is especially crucial because the performance of the algorithms in predicting the results depends on the input characteristics, the number of samples and the modelling method (Shawon et al., 2025). The information related to soils has also shown potential in improving the estimation of crop production in advanced predictive models (Tripathi et al., 2022).

Even with these advances, there is still a need for research to determine the predictive role of soil physicochemical properties aimed at estimating yield of cacao. While the existing machine learning literature is often focused on general crop production, remote sensing and climatic variables, there are opportunities to develop soil-specific prediction models for specific crops. The selection of appropriate input variables, model evaluation, and transferability/harvesting in crop production prediction studies using deep learning have also been highlighted (Oikonomidis et al., 2023). Hence the purpose of the present study is to assess the relationships between soil physicochemical properties and the yield of cacao, to formulate LR, DTR, RFR and SVR models for predicting the yield of cacao, and to determine the most important soil indicators that contribute to the prediction of the yield of cacao. The study is to offer evidence to support soil-based assessment of productivity and improved agricultural decision-making, using secondary agricultural observations from Ghana.

2. Materials and Methods

2.1 Study Design and Data Source

The study used a quantitative, secondary data analysis research design to predict the yield of cacao using soil physicochemical properties using machine learning techniques. The data were obtained from the publicly available Mendeley Data repository (Abdulai et al., 2020b). The original study involved responses from farmers, observations in the fields and soil laboratory measurements from 150 cacao plantations in three climatic zones of Ghana.

2.2 Study Variables and Data Characteristics

Cacao yield (yield.kg.ha.yr) is the dependent variable which is measured in kilograms per hectare per year. Soil

characteristics used as independent variables were: soil pH, organic carbon, total nitrogen, soil organic matter, calcium, magnesium, potassium, sodium, phosphorus, exchangeable acidity, effective cation exchange capacity (ECEC), base saturation, sand, clay and silt percentages. These variables were soil acidity, nutrient availability, organic matter content, nutrient-holding capacity and soil texture. The observations were evenly spread over the different climatic zones dry, mid and wet with 50 plantations in each zone. Climatic zone was used descriptively for comparisons, but the main prediction models are limited to soil physicochemical variables to be consistent with the research objectives.

2.3 Data Preprocessing

The data was checked for consistency, accuracy and appropriateness for statistical analysis and machine learning prediction. A total of 149 complete observations were kept for the major predictive analysis. Potential inconsistencies and extreme values were checked for soil physicochemical measurements and descriptive statistics and graphical methods used for continuous variables. The relationships between the soil properties were assessed for possible multicollinearity and redundant predictors. Where appropriate numerical variables were standardized to ensure good performance of the model, such as Support Vector Regression (SVR). The training observations were standardized to obtain data standardization and the training data were then applied to the testing observations to avoid data leakage and ensure the reliability of the model evaluation.

2.4 Descriptive Statistical Analysis

The descriptive statistical analysis was employed to describe cacao yield and the soil selected physicochemical properties. Mean, standard deviation, minimum and maximum values were used to describe continuous variables. The distribution of soil pH, Organic carbon, N, P, K and cacao yield were analyzed to assess the variation of the plantations. Descriptive comparisons were also to be made between the three climatic zones, to see if there were any differences in soil properties and productivity between the regions. The distribution of the main soil variables and of the cacao yield were chosen to be represented by a histogram and a boxplot, respectively.

2.5 Correlation Analysis

A correlation analysis was planned to explore the correlation between soil physicochemical properties and cacao yield. Pearson and Spearman rank correlation coefficients were chosen for approximate linear relationships between continuous variables and for substantially skewed distributions and for monotonic nonlinear relationships, respectively. Direction and magnitude of correlation coefficients were used to interpret the correlation, and the significance of the correlation was determined at the 0.05 level. Before predictive modelling, strong correlations between independent variables were also explored, to determine if there was multicollinearity.

2.6 Machine Learning Model Development

The four supervised machine learning regressors namely Linear Regression (LR), Decision Tree Regression (DTR), Random Forest Regression (RFR), and Support Vector Regression (SVR) were chosen for cacao yield prediction based on soil physicochemical properties. As a base model, Linear Regression was used to find any linear relationship between the soil indicators and yield. To capture the nonlinear patterns, recursive partitioning was done using Decision Tree Regression and to ensure stability of the prediction and reduce over fitting, multiple decision trees were combined to form Random Forest Regression. Support Vector Regression was added to assess the nonlinear prediction with the Kernel based modelling. The same predictor variables and outcome measurements were entered into all models to allow for a consistent comparison between models.

2.7 Model Training and Validation

Approximately 119 (80%) of the complete-case observations were proposed to be used as training observations and 30 (20%) observations were proposed to be used as testing observations. To evaluate the stability of the model and to aid with hyperparameter optimization, five-fold cross validation was planned within the training subset. The number of trees, maximum tree depth and minimum samples per leaf were to be optimized for the Random Forest model. As well the depth of the decision tree and minimum leaf size were taken into account and the parameters of the SVR were regularization parameter, kernel type and epsilon. A fixed random seed was used to ensure the reproducibility of the results. Only the training and cross-validation set were used to select the model; independent testing set was used for final evaluation of the performance of the model only.

2.8 Model Performance Evaluation

The coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were used to assess the predictive performance. R^2 was used to assess the coverage of the model prediction, while RMSE and MAE were used to assess prediction errors in the original unit of yield. The values of R^2 and RMSE and MAE were higher the better the predictive performance was. Testing observations were the same for the four regression procedures. Scatterplots of observed and predicted cacao yield, and residual plots were suggested to analyse the performance in predicting yield, systematic errors, and possible model overfitting.

2.9 Feature Importance Analysis

The soil physicochemical property analysis was planned using feature importance analysis in order to know which soil physicochemical property was more significant in the prediction of cacao yield. Permutation feature importance was

chosen to determine the changes in predictive error for randomly permuting the value of each predictor. This way of interpretation was used to enable consistent interpretation over the selected regression models. The different soil indicators were ranked based on their contribution to the predictive performance, and a horizontal bar chart was used to display this. The importance of a feature was used to mean that it is useful for prediction, not that it has a direct impact on the cacao yield.

2.10 Ethical Considerations

Secondary agricultural data was sourced from publicly available databases, Mendeley Data, to analyze and study the data. There was no direct contact with human subjects or animals in the proposed analysis. The original data were made available to the repository under the CC0 1.0 licence.

3. Results

3.1 Descriptive Characteristics of Soil Physicochemical Properties and Cacao Yield

The soil physico-chemical characteristics and the production of cacao beans were described. The first dataset comprised 150 plantations of cacao trees in three climatic regions in Ghana. All farms had the yield of cacao while 149 observations had full measurements for the selected soil predictors. As presented in Table 1, the overall mean cacao yield was 637.62 ± 508.87 kg/ha/year, ranging from 46.86 to 2,482.95 kg/ha/year. Soil pH ranged from 4.46 to 7.61, with a mean of 6.35 ± 0.84. The mean and standard deviations of soil organic carbon and total nitrogen are 1.56 ± 0.49%, and 0.14 ± 0.04%, respectively. There was also some variation in calcium, magnesium, phosphorus, potassium, and soil texture measurements. These descriptive results suggest that there is a great variability in soil properties and cacao productivity in the plantations sampled.

Table 1. Descriptive Statistics of Soil Physicochemical Properties and Cacao Yield

Variable	n	Mean ± SD	Minimum	Maximum
Cacao yield (kg/ha/year)	150	637.62 ± 508.87	46.86	2482.95
Soil pH	149	6.35 ± 0.84	4.46	7.61
Organic carbon (%)	149	1.56 ± 0.49	0.55	3.07
Total nitrogen (%)	149	0.138 ± 0.040	0.03	0.26
Calcium (cmol/kg)	149	7.54 ± 5.65	1.02	41.12
Magnesium (cmol/kg)	149	2.35 ± 1.52	0.53	11.75
Potassium (cmol/kg)	149	0.210 ± 0.096	0.06	0.49
Sodium (cmol/kg)	149	0.106 ± 0.047	0.04	0.26
Phosphorus (reported units)	149	11.19 ± 6.58	3.76	53.47
Exchangeable acidity (cmol/kg)	149	0.235 ± 0.208	0.02	1.00
Sand (%)	149	43.99 ± 14.88	13.60	81.04
Clay (%)	149	12.49 ± 4.53	5.36	24.17

Note: SD = standard deviation. Descriptive statistics were computed for the provided Mendeley Data file. Phosphorus units are reported as they are since the column label in the workbook does not include a unit of measure.

The yield of cacao was also found to vary among the climatic zones. Mean yield was 251.80 ± 172.92 kg/ha/year in the dry zone, 546.72 ± 292.38 kg/ha/year in the mid zone, and 1,111.88 ± 533.31 kg/ha/year in the wet zone. This difference is represented in a graphic form in Figure 1 by the boxplots for the observed yields of cacao beans in each of the three climatic zones, where higher yields are observed in the wet climatic region.

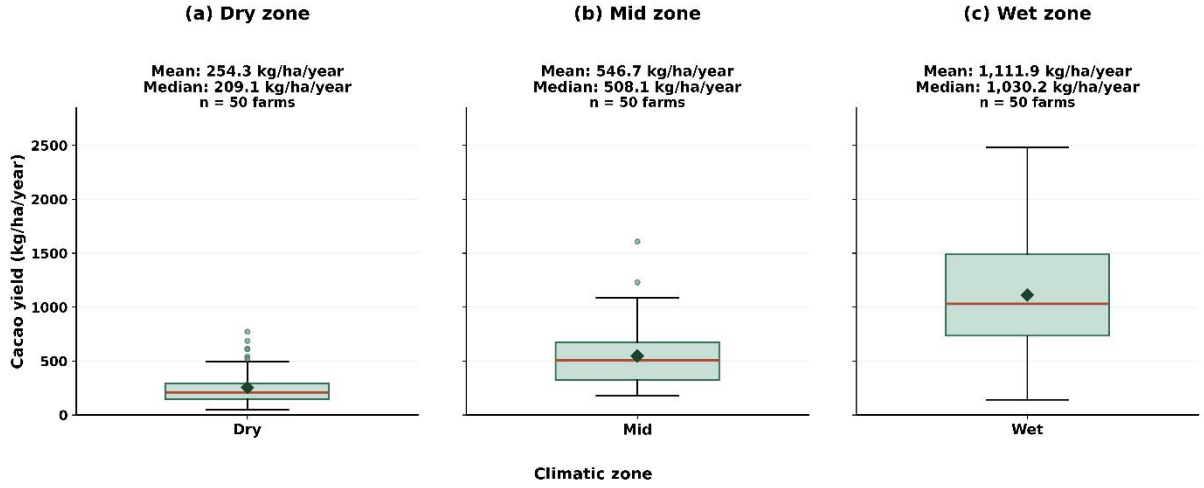


Figure 1. Distribution of Cacao Yield Across Climatic Zones Boxplots represent cacao yield distributions across (a) dry, (b) mid, and (c) wet climatic zones

3.2 Correlation Between Soil Physicochemical Properties and Cacao Yield

There is a correlation between soil physicochemical properties and the yield of cacao beans. Pearson correlation analysis revealed that several soil physicochemical properties were significantly related to the yield of cacao. The soil (exch. acidity) and soil (pH) data had the highest positive and negative correlation with yield, respectively (Table 2). Yield correlation was moderately, but negatively, related to sand content ($r = -0.504$, $p < 0.001$). Statistically significant negative correlations were also found between potassium, calcium and magnesium. The other properties (organic carbon, total nitrogen, P, Na, and clay) did not exhibit statistically significant linear relationships with cacao yield. These results reflect observations and not necessarily causal relationships between individual soil properties.

Table 2. Pearson Correlation Between Soil Physicochemical Properties and Cacao Yield

Soil property	Pearson r	p-value
Exchangeable acidity	0.601	<0.001
Soil pH	−0.600	<0.001
Sand (%)	−0.504	<0.001
Potassium	−0.259	0.001
Calcium	−0.238	0.004
Magnesium	−0.191	0.020
Total nitrogen	0.147	0.075
Phosphorus	−0.121	0.143
Organic carbon	0.111	0.178
Clay (%)	0.041	0.618
Sodium	−0.008	0.919

Note that p values less than 0.001 were not reported as 0.001, but as $p < 0.001$. Note that the number of observations used for calculations of the correlations is 149. Correlation coefs are not causal relationships.

The correlation heatmap in Figure 2 gives an overall picture of the relationship between the selected soil variables and cacao yield. Relationships between soil pH and exchangeable acidity with yield were opposite, while the other soil properties (organic carbon, phosphorus, sodium and clay) showed weaker relationships with yield. The figure also shows associations between soil variables, which are useful in the interpretation of potentially overlapping soil variables in machine learning models.

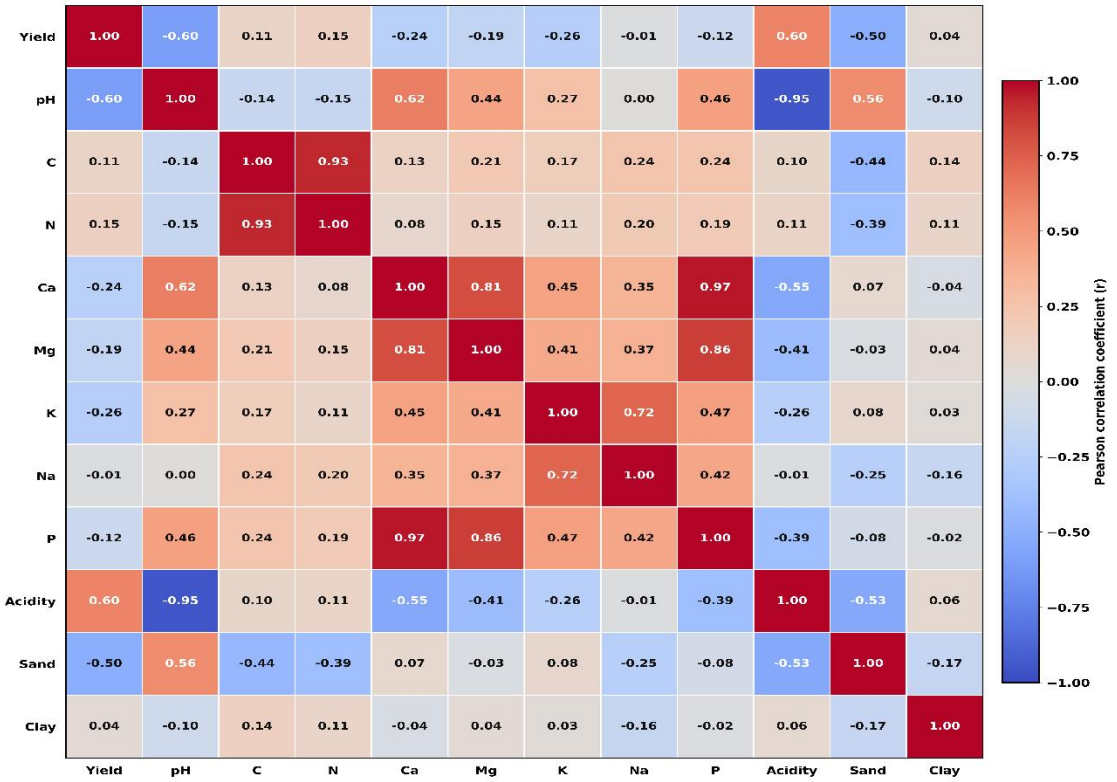


Figure 2. Pearson Correlation Heatmap of Soil Physicochemical Properties and Cacao Yield

3.3 Comparative Performance of Machine Learning Models

A total of 119 complete observations were used to train four regression models and an independent testing subset of 30 observations was used to evaluate the regression models. To select the model, five-fold cross-validation was used in the training subset. Table 3 shows that the highest value of the R^2 of testing data (0.487) and lowest value of RMSE (375.07 kg/ha/year) were obtained by the Random Forest Regression, which gives the maximum accuracy using these two performance measures. However, Decision Tree Regression had the highest R^2 on testing data (0.431) and the lowest MAE (274.23 kg/ha/year) which is slightly better than Random Forest in absolute prediction error. Linear Regression achieved a test R^2 of 0.391 while Support Vector Regression achieved the lowest test R^2 (0.312) and the highest prediction errors. Random Forest had the best overall performance in terms of explained variance and root mean square prediction error. Note: R^2 = Coefficient of determination, RMSE = Root mean square error, MAE = Mean absolute error. All tests were run on 30 withheld observations. Tuning of the models was done using five-fold cross-validation of the training subset, except for the baseline Linear Regression.

Table 3. Comparative Performance of Machine Learning Models for Cacao Yield Prediction

Model	Training R^2	Testing R^2	RMSE (kg/ha/year)	MAE (kg/ha/year)
Linear Regression	0.475	0.391	408.58	297.51
Decision Tree Regression	0.695	0.431	394.78	274.23
Random Forest Regression	0.616	0.487	375.07	275.73
Support Vector Regression	0.442	0.312	434.08	301.68

Note: R^2 = coefficient of determination; RMSE = root mean square error; MAE = mean absolute error. All testing results were obtained from the same 30 held-out observations. Models were tuned using five-fold cross-validation on the training subset, except baseline Linear Regression.

The observed and predicted cacao yield values are compared in Figure 3. The overall agreement between the different models and the observed values was not good, as observed in the prediction errors of the models. Random Forest Regression (RFR) predicted the values to be closer to the observed values than the other models, but there were still noticeable prediction errors. This difference between training and test performance was greater for Decision Tree Regression, suggesting less generalization ability when predicting unseen cases. The results agree with the choice of Random Forest since it has the highest overall testing accuracy.

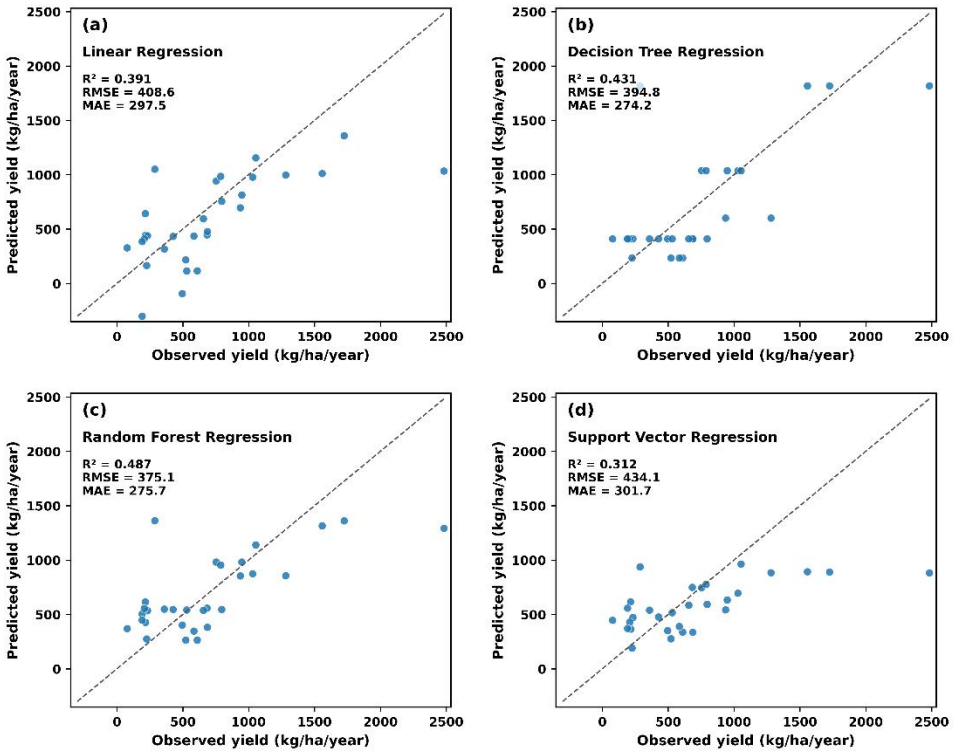


Figure 3. Comparison of Observed and Predicted Cacao Yield Across Machine Learning Models Scatterplots compare observed and predicted cacao yield for (a) Linear Regression, (b) Decision Tree Regression, (c) Random Forest Regression, and (d) Support Vector Regression.

The dashed diagonal represents perfect prediction, and the R^2 values indicate testing performance. Each panel uses the same independent testing observations for model comparison.

3.4 Identification of Important Soil Predictors

The selected RF model was used in the permutation feature importance analysis to determine the most important soil properties for predicting cacao yields. Soil pH resulted in the largest mean square prediction error for soil after permutation (MSPE) (91,327.81), followed by exchangeable acidity (25,912.94) and sodium (18,102.04), as presented in Table 4. Potassium had a lower positive importance score, and the contributions of sand, clay, calcium and organic carbon were relatively low. Magnesium and total nitrogen were negatively correlated with mean importance value and their permutation did not adversely affect predictive performance for this testing sample. These results indicate that the soil acidity related variables were significantly related in this model but the importance ranking may be influenced by the inter-correlation of the soil variables and the restricted sample of testing.

Table 4. Permutation Feature Importance of Soil Physicochemical Properties in Random Forest Regression

Soil property	Mean importance (MSE increase)	Rank
Soil pH	91,327.81	1
Exchangeable acidity	25,912.94	2
Sodium	18,102.04	3
Potassium	6,175.19	4
Sand	2,039.86	5
Clay	1,732.32	6
Calcium	594.12	7
Organic carbon	213.65	8
Phosphorus	63.33	9

Magnesium	−428.54	10
Total nitrogen	−2,288.22	11

Note: The importance is computed as the average of the change in mean squared error over 10 repetitions on the testing subset when the predictor is permuted for each repetition. The higher the positive value the more important it is to the fitted model. Negative values suggest that no improvement occurred in predicting the sample of this test when using that predictor, compared with using the constant.

The relative importance of the 11 soil predictors is shown in Figure 4. Soil pH was found to be the most important factor followed by exchangeable acidity and sodium. Several nutrient and texture variables have essentially lower importance values, suggesting that they are not able to provide much additional predictive power in the fitted Random Forest model.
Figure 4. Feature Importance of Soil Physicochemical Properties for Cacao Yield Prediction
Each set of error bars refers to variation in repeated permutations. The larger the positive scores, the more prediction error that was reduced after the permutation. Negative scores are included for predictors that do not contribute positively in any way on the test observations.

4. Discussion

The present study was designed to examine the relationship between the soil physicochemical properties with yield of cacao and compare the performance of four ML models for predicting yield. The results indicated that there was significant variation among the plantations in cacao productivity, with the soil pH and exchangeable acidity being the most significant factors that correlated well with productivity. There was positive association with exchangeable acidity ($r = 0.601$) and negative association with soil pH ($r = -0.600$). Random Forest Regression showed the highest R^2 (0.487) and lowest RMSE (375.07 kg/ha/year) among the models evaluated, which means these models had moderate predictive ability. The results indicate the significance of soil properties for assessing cacao productivity. In Ghana, previous studies have shown that the yield of cacao depends on the environment, agricultural management and the farm characteristics (Abdulai et al., 2018). Also, studies on cocoa fertilization have highlighted the significance of nutrient management for enhancing the productivity of cocoa farming (Vasquez-Zambrano et al., 2025).

The relationships between soil physicochemical properties and cacao production observed indicate that soil–crop interactions are complex. That there is a negative relationship between soil pH and yield should be taken to mean that the productivity of cacao does not increase with soil acidity. Rather, the relationship may be a result of variations in climate, soil development, and nutrient availability of the plantations sampled. A monoculture system and an agroforestry system have been studied and evaluated, and it has been shown that the productivity of the cocoa tree is influenced by several environmental and management variables (Ariza-Salamanca et al., 2024). Likewise, model-based research has also highlighted the need to adapt cacao production systems to the suitable climatic and environmental situation (Saputra et al., 2024). Organic carbon, organic nitrogen and organic phosphorus had relatively low individual correlations with the yield, but the nutrients are important in providing fertility and growth to the plants. Research on soil fertility interventions and their role in sustainable cocoa production has been identified as one of the actions needed for climate-smart soil management (Anokye et al., 2024). The results of research into the use of calcium nitrate fertilizer have also shown the importance of the management of nitrogen in cacao farming (Dogbatse et al., 2024). Furthermore, interaction among several physiological and environmental processes can potentially affect cacao productivity through better agricultural practices and nutrient management (Mendoza-Meneses et al., 2023).

The comparative machine learning results showed that there were differences between the predicting ability of the selected algorithms in predicting the yield of cacao. Random Forest Regression was better at representing relationship between soil properties and yield on testing basis, in terms of R^2 and RMSE than other three regression methods, namely Linear Regression, Decision Tree Regression, and Support Vector Regression, as a result, a beneficial outcome of the ensemble learning was obtained. But, the number of errors made by Decision Tree Regression were slightly less, which shows that the performance of the model was dependent on the evaluation metric. Machine learning algorithms have already been proved to be capable of estimating crop yield in previous agricultural studies (Iniyan et al., 2023). Similarly, machine learning models based on remote sensing data have demonstrated the potential in predicting wheat productivity (Cheng et al., 2022). Deep transfer learning has also been successfully implemented in yield prediction of corn and soybeans, which has a different set of modelling requirements than smaller soil-based studies (Khaki et al., 2021). Despite the moderate accuracy results obtained in this study, it is essential to conduct a thorough validation process, especially for predictions that are based on limited agricultural observations (Leukel et al., 2023).

The most important soil parameters were selected based on feature importance analysis in the Random Forest model, which revealed that soil pH, exchangeable acidity, and sodium were the most significant. This indicates that the measurements made for soil acidity may be beneficial for soil monitoring systems in cacao plantations. However, the permutation importance estimates are likely to be affected by correlated soil variables, and caution needs to be used when interpreting the importance of individual soil variables (Kaneko, 2022). Soil measurements combined with predictive modelling could be used to facilitate more focused agricultural assessment and, more broadly, machine learning has shown potential for enhancing regional scale crop yield prediction (Paudel et al., 2022).

The climatic factors, cacao varieties, pest occurrence and management practices were not entered as predictors, which might explain the low predictive ability. The value of the combination of remote sensing and environmental data to support agricultural yield estimation has been shown (de Freitas Cunha & Silva, 2020). Additionally, machine learning

studies have shown that there are environmental factors that can be used to enhance the robustness of the prediction process while only using a limited amount of information in the process (Xu et al., 2021). Larger, geographically expanded samples should be used in future studies, as should additional climatic and management indicators, and independent external validation to further enhance the reliability of cacao yield prediction.

5. Conclusion

The present study proved that machine learning methods can be used to predict the yield of cacao trees based on the physicochemical properties of the soils in the smallholder cacao plantations in Ghana. The results indicated a high variability among the cacao products and that some relationships existed between the yields and selected soil properties. Soil pH and exchangeable acidity proved to be significant parameters, with good negative and positive correlation with cacao yield, respectively. All four machine learning models were evaluated, and the overall predictive performance was the highest for Random Forest Regression, with a testing R^2 value of 0.487 and RMSE of 375.07 kg/ha/year. Performance was different in various evaluation measures, with Decision Tree Regression having the lowest mean absolute error. The soil parameters pH, exchangeable acidity and sodium were also identified as being influential in the cacao yield through feature importance analysis. The results demonstrate the potential of soil-based machine learning methods to explain yield variability and aid in agricultural decision-making. The relatively low predictive power, however, suggests that these soil physicochemical properties cannot account for all of the variations in cacao productivity. The findings are also not generalizable because of the small number of samples and geographic areas. Climatic conditions, agricultural management practices, crop varieties, and other environmental indicators should be added to future studies in order to enhance prediction accuracy. The use of soil assessment combined with machine learning is a promising strategy for the sustainable production of cacao, better soil management strategies, and for precision agriculture in the regions where cacao is grown.

References

1. Abdulai, I., Hoffmann, M. P., Jassogne, L., Asare, R., Graefe, S., Tao, H. H., ... & Rötter, R. P. (2020a). Variations in yield gaps of smallholder cocoa systems and the main determining factors along a climate gradient in Ghana. Agricultural Systems, 181, 102812.
2. Abdulai, I., Jassogne, L., Graefe, S., Asare, R., Van Asten, P., Läderach, P., & Vaast, P. (2018). Characterization of cocoa production, income diversification and shade tree management along a climate gradient in Ghana. PloS one, 13(4), e0195777.
3. Abdulai, I., Muilerman, S., Jassogne, L., Hoffmann, M. P., van Asten, P., Rötter, R. P., Tao, H. H., Vaast, P., Asare, R., Läderach, P., & Graefe, S. (2020b). Data for: Variations in yield gaps and soil fertility along a climatic gradient in smallholder cocoa systems of Ghana (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/bcrd9g226g.1
4. Adet, L., Rozendaal, D. M., Zuidema, P. A., Vaast, P., & Anten, N. P. (2024). Cocoa tree performance and yield are affected by seasonal rainfall reduction. Agricultural Water Management, 302, 108995.
5. Agbotui, D. K., Ingold, M., & Buerkert, A. (2024). Carbon and nutrient cycling in cocoa agroforests under organic and conventional management. Nutrient Cycling in Agroecosystems, 129(1), 7-20.
6. Amponsah-Doku, B., Daymond, A., Robinson, S., Atuah, L., & Sizmur, T. (2022). Improving soil health and closing the yield gap of cocoa production in Ghana–A review. Scientific African, 15, e01075.
7. Anokye, J., Abunyewa, A. A., Jørgensen, U., Kaba, J. S., Twum-Ampofo, K., Dawoe, E., ... & Ulzen, J. (2024). Mitigation of greenhouse gas emissions through shade systems and climate-smart soil fertility interventions in cocoa landscapes in the semi-deciduous ecological zone of Ghana. Soil Advances, 1, 100001.
8. Ariza-Salamanca, A. J., Navarro-Cerrillo, R. M., Crozier, J., Stirling, C., Mancini, A., Blaser-Hart, W., & González-Moreno, P. (2024). Drivers of cocoa yield and growth in young monoculture and agroforestry systems. Agricultural Systems, 219, 104044.
9. Asitoakor, B. K., Vaast, P., Ræbild, A., Ravn, H. P., Eziah, V. Y., Owusu, K., ... & Asare, R. (2022). Selected shade tree species improved cocoa yields in low-input agroforestry systems in Ghana. Agricultural Systems, 202, 103476.
10. Cheng, E., Zhang, B., Peng, D., Zhong, L., Yu, L., Liu, Y., ... & Yang, S. (2022). Wheat yield estimation using remote sensing data based on machine learning approaches. Frontiers in Plant Science, 13, 1090970.
11. de Freitas Cunha, R. L., & Silva, B. (2020, March). Estimating crop yields with remote sensing and deep learning. In 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS) (pp. 273-278). IEEE.
12. Doe, E. K., Attua, E. M., Obour, P. B., Quaye, A. K., & Fosu-Mensah, B. Y. (2023). Soil health and synergy of ecological determinants of green cocoa productivity in different soil ecotypes in Ghana. Frontiers in Sustainable Food Systems, 7, 1169015.
13. Dogbatse, J. A., Awudzi, G. K., Arthur, A., Owusu-Ansah, F., Amoako-Attah, I., Quaye, A. K., ... & Amaning, A. A. (2024). Assessment of calcium nitrate fertilizer as a suitable nitrogen source for cacao (Theobroma cacao L.) cultivation in Ghana. International Journal of Agronomy, 2024(1), 5578534.
14. Goudsmit, E., Rozendaal, D. M., Tosto, A., & Slingerland, M. (2023). Effects of fertilizer application on cacao pod development, pod nutrient content and yield. Scientia Horticulturae, 313, 111869.
15. Iniyan, S., Varma, V. A., & Naidu, C. T. (2023). Crop yield prediction using machine learning techniques. Advances in Engineering Software, 175, 103326.
16. Jabed, M. A., & Murad, M. A. A. (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).
17. Kaneko, H. (2022). Cross-validated permutation feature importance considering correlation between features. Analytical

Science Advances, 3(9-10), 278-287.
18. Khaki, S., Pham, H., & Wang, L. (2021). Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. Scientific Reports, 11(1), 11132.
19. Konaté, N. G., Ouattara, Y., Kouakou, A. K., & Barima, Y. S. (2024). Effects of traditional agroforestry practices on cocoa yields in Côte d'Ivoire. Sustainability, 16(22), 9927.
20. Leukel, J., Zimpel, T., & Stumpe, C. (2023). Machine learning technology for early prediction of grain yield at the field scale: A systematic review. Computers and Electronics in Agriculture, 207, 107721.
21. Mendoza-Meneses, C. J., Feregrino-Pérez, A. A., Guevara-González, R. G., & García-Trejo, J. F. (2023). Implementation of pre-harvest techniques in emerging agroforestry systems to increase the yield of cocoa tree (Theobroma cacao L.). Heliyon, 9(3).
22. Mwafulirwa, L., Sizmur, T., Daymond, A., Atuah, L., Quaye, A. K., Coole, S., ... & Hadley, P. (2024). Cocoa pod husk-derived organic soil amendments differentially affect soil fertility, nutrient leaching, and greenhouse gas emissions in cocoa soils. Journal of Cleaner Production, 479, 144065.
23. Oikonomidis, A., Catal, C., & Kassahun, A. (2023). Deep learning for crop yield prediction: a systematic literature review. New Zealand Journal of Crop and Horticultural Science, 51(1), 1-26.
24. Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., ... & Athanasiadis, I. N. (2022). Machine learning for regional crop yield forecasting in Europe. Field Crops Research, 276, 108377.
25. Saputra, D. D., Khasanah, N. M., Sari, R. R., & van Noordwijk, M. (2024). Avoidance of tree-site mismatching of modelled cacao production systems across climatic zones: Roots for multifunctionality. Agricultural Systems, 216, 103895.
26. Schmidt, J., Addo-Danso, S. D., Asare, R., Tettey, A., & Isaac, M. E. (2024). Soil quality reflects microbial resource availability and drives rhizosphere microbiome variation in Ghanaian cocoa farms. Applied Soil Ecology, 198, 105378.
27. Shawon, S. M., Ema, F. B., Mahi, A. K., Niha, F. L., & Zubair, H. T. (2025). Crop yield prediction using machine learning: An extensive and systematic literature review. Smart Agricultural Technology, 10, 100718.
28. Tripathi, A., Tiwari, R. K., & Tiwari, S. P. (2022). A deep learning multi-layer perceptron and remote sensing approach for soil health based crop yield estimation. International Journal of Applied Earth Observation and Geoinformation, 113, 102959.
29. 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.
30. Vasquez-Zambrano, E., Woittiez, L. S., van Heerwaarden, J., Rusinamhodzi, L., Hauser, S., & Giller, K. E. (2025). Deriving fertiliser recommendations for cocoa: An offtake model approach. European Journal of Agronomy, 164, 127463.
31. Xu, H., Zhang, X., Ye, Z., Jiang, L., Qiu, X., Tian, Y., ... & Cao, W. (2021). Machine learning approaches can reduce environmental data requirements for regional yield potential simulation. European Journal of Agronomy, 129, 126335.

EPH – International Journal of Mathematics and Statistics
https://eijms.com/index.php/ms/index



This work is licensed under a Creative Commons Attribution 4.0 International License. The authors retain ownership of the copyright for their article,You are free to: Share — copy and redistribute the material in any medium or format for any purpose, even commercially. To see the complete license contents,

please visit **http://creativecommons.org/licenses/by/4.0/.**