Nigerian Journal of Technology (NIJOTECH)
Vol. 44, No. 4, December, 2025, pp 779 - 790
www.nijotech.com

Print ISSN: 0331-8443
Electronic ISSN: 2467-8821
https://doi.org/10.4314/njt.2025.5528

YIELD PREDICTION OF NSUKKA YELLOW PEPPER USING SOME MACHINE LEARNING MODELS

AUTHORS:
O. J. Ogbu[1], A. O. Ani[1,2,4], K. O. Ugwueze[2], I.E. Ngwu[2*], N. Apanta[2], M.C. Odo[2], U. Ezechi[2], V.N. Amu[2], O.O. Oguanya[2], C. Anioke[3]

AFFILIATIONS:
[1] Department of Agricultural and Bioresources Engineering, University of Nigeria, Nsukka
[2]Department of Mechatronic Engineering, University of Nigeria, Nsukka
[3]Department of Computer and Electronic Engineering, University of Nigeria, Nsukka
[4]Department of Electrical and Smart Systems Engineering, University of South Africa

***CORRESPONDING AUTHOR:**
Email: Ifeanyi.ngwu@unn.edu.ng

ARTICLE HISTORY:
Received: August 28, 2025.
Revised: December 31, 2025.
Accepted: December 31, 2025.
Published: January 03, 2026.

KEYWORDS:
Nsukka Yellow Pepper, Crop Yield Prediction, Machine Learning, Experimental Investigation

ARTICLE INCLUDES:
Peer review

DATA AVAILABILITY:
On request from author(s)

EDITORS:
Chidozie Charles Nnaji

FUNDING:
None

Abstract

In this study, the accuracy and efficiency of four machine learning models, Random Forest, Decision Tree, Multiple Linear Regression and Support Vector Machine were evaluated for predicting the yield of Nsukka yellow pepper. This kind of predictive study will help farmers and other related workers in informed decision making with farm inputs, resource optimization, risk management and market planning. The study was carried out at the Research Farm, Department of Agricultural and Bioresources Engineering, University of Nigeria, Nsukka, Enugu State, Nigeria. Historical data for temperature, humidity and solar radiation were collected from the Meteorological Station of the National Centre for Energy Research and Development, University of Nigeria, Nsukka. The historical data covered a period of three months, 9th January to 9th April 2024, corresponding to the period when crop parameters such as plant height, leaf length, leaf width and crop yield were measured and collated for one hundred and seven plants of Nsukka yellow pepper. This period covers the time from transplanting to harvesting. The machine learning regression algorithms were trained, validated and tested using yield data from the pepper farm. The algorithms developed with Python codes were trained in Google Colab. Results revealed that the Multiple Linear Regression model had the best performance metrics with an MSE of 45.23, RMSE of 6.73 and R^2 of 0.67. This implies that the Multiple Linear Regression model is an effective tool for predicting the yield of Nsukka yellow pepper, and hence useful for decision support to farmers. Furthermore, to statistically confirm the performance differences among the models, one-way ANOVA and ANCOVA tests were conducted. ANOVA results ($F = 25.76$, $p < 0.001$) showed significant variation among the mean predicted yields for the four models, while the ANCOVA ($F = 18.92$, $p < 0.001$), controlling for actual yield, confirmed that the models differed significantly even after accounting for real yield effects. This confirms that the Multiple Linear Regression and Support Vector Regression models outperformed Random Forest and Decision Tree models in terms of predictive consistency.

1.0 INTRODUCTION

Many smallholder farmers in Africa rely on pepper growing as a source of income [1]–[6]. A distinctive pepper species that is extensively grown as a commercial vegetable crop in Southeast Nigeria is the Nsukka yellow pepper. In Nsukka, peasant women's primary, and even exclusive, agricultural endeavor is cultivating it [7]. Because it was mostly grown by women, this crop was once known as a "women's crop." However, because pepper is so profitable, many males are now involved in its production.

HOW TO CITE:
Ogbu, O. J., Ani, A. O., Ugwueze, K. O., Ngwu, I.E., Apanta, N., Odo, M.C., Ezechi, U., Amu, V.N., Oguanya, O.O. Anioke, C. "Yield Prediction of Nsukka Yellow Pepper Using Some Machine Learning Models", *Nigerian Journal of Technology*, 2025. 44(4), pp. pp 779 - 790. https://doi.org/10.4314/njt.2025.5528

The Nsukka yellow pepper is a common species with widespread acceptance and promise that has provided many farmer families in the Nsukka agricultural zone with a substantial source of revenue, despite the fact that Nigeria is home to a large variety of pepper varieties. Usually grown in Enugu State, the yellow pepper is native to Nsukka. Perhaps because other ecologies would not be able to support its development as well as those in Nsukka and the surrounding area, it is uncommon in other areas of the country. The fruits of this particular pepper species are fairly large and turn yellowish-green when fully ripe. It also has a unique scent. In the local rural and urban markets, the pepper is more popular than other pepper varieties due to its unique flavor, color, and aroma.

Maximizing pepper yield at the lowest cost while preserving a healthy ecological is one of the fundamental goals of agricultural production. Many crop management and financial choices depend on yield estimation, and early detection and resolution of pepper yield limit problems can help increase output and subsequent profit. In the food production industry, pepper yield prediction is crucial because it helps farmers and other relevant personnel make well-informed decisions about planting, fertilizer, irrigation, resource optimization to maximize productivity, risk management, and marketing. Policymakers depend on precise estimates to make prompt import and export choices in order to improve national food security [8], [9].

Even though the Nsukka yellow pepper has enormous economic potential, not much research has been done to anticipate the yield utilizing contemporary research approaches like machine learning prediction algorithms. Nsukka yellow pepper is not often grown in the majority of Nigerian states, according to observations. The production of premium peppers, like Nsukka yellow pepper, needs to be increased in other parts of Nigeria. Nonetheless, a method to forecast Nsukka yellow pepper yield in regions with distinct soil and environmental variables from Nsukka is crucial. Using information from crops, soil, climate, satellites, and other sources, machine learning has been utilized by several academics to estimate crop yields [10], [11].

Numerous factors, such as rainfall, temperature, and season, were considered in order to forecast the crop production. Since no further machine-learning methods were used on the datasets, there was no way to compare how well various algorithms performed. Computer simulation models based on MLR and RF were used by [12] to evaluate yield prediction and nitrate loss accounting. The most accurate and dependable approaches, according to a report from [13], are the average ensemble and the optimal weighted ensemble. However, the effectiveness of ensemble modeling is impacted by the selection of many machine learning models; therefore, further research is needed. The work by [14] extended conventional neural networks for regression. The conformal prediction (CF) framework was used to predict tomato output in a greenhouse by accounting for temperature, radiation, CO2, and vapor pressure deficit (VPD). This approach required the utilization of almost 60,000 records. Satellite imagery and meteorological data were used by [15] to anticipate soybean yield in real time using LSTM, OLS linear regression, and RF models. The LSTM performed remarkably well in comparison to RF and multivariate OLS regressions. The DNN model outperformed the SVR, RFR, and PLSR regression techniques as the number of input features rose. Research carried out by [16] discovered that soybean production could be precisely determined using a DNN technique that used RGB, multispectral, and temperature sensors in conjunction with UAV-based multimodal data fusion.

In a hybrid MLR-ANN model created by [17], the ANN's input layer weights and bias were initialized using the MLR's coefficients and bias rather than random weights and bias initialization. The suggested hybrid MLR-ANN model outperformed SVR, K-NN, and RF in terms of prediction accuracy on the same agricultural dataset. The use of SVR, OLS, and GWR models by [18] showed that the GWR model fared better than the SVR and OLS models since it assigned different weight values to geographical grids. In a comparison of the efficacy of SVR, RF, ANN, and K-NN, it was determined by [19] that the RF algorithm is better since it attains the best accuracy. In general, hybridized prediction algorithms outperform single prediction algorithms. A mathematical optimization method that predicts potato yield using the model's projected biomass was created by [20]. Predicting wheat production is another well-liked study in this area that has been conducted by numerous researchers.

According to [21], the spatial NN model was able to estimate wheat production for a particular planting region with excellent accuracy, in contrast to the temporal NN models. However, because they only included one component—the planting area—the

model is less reliable. A yield forecast model based on climate and satellite data was created by [22] using the LASSO, SVM, RF, and NN algorithms. Their increased precision led them to the conclusion that using satellite data to create an annual wheat map would be easier because it reduces errors. A modeling approach that included data from the soil, climate, and vegetation index to predict winter wheat output was investigated by [23]. SVM, GPR, and RF were the top three techniques for yield prediction, according to the researchers' analysis of eight popular machine learning models. In their research, a new activation function for MLP was proposed by [24], which also updated the random weights and bias values. The findings demonstrated that newly created activation functions worked better than the sigmoid, which is the NN's default activation function. The use of RGB and NDVI data separately with the CNN model by [25] led to the claim that RGB images performed better than NDVI. Maize yield at the county level in China was forecasted by [26] using the RF, LASSO, LSTM, and XGBoost algorithms in combination with thermal, optical, fluorescence, and environmental satellite data. The investigation's findings showed that SIF's poor signal-to-noise ratio and coarse spatial resolution allowed it to outperform EVI in terms of output quality. The forecast of potato yield utilizing NDVI and soil property datasets was performed by [27] using LR, SVR, and k-NN algorithms. When it came to forecasting potato yield, the SVR models fared better than the other three models. However, additional climatic, chemical, meteorological, and physical factors caused forecasts in related areas to differ in different years.

In light of irrigation season and water management, it was found by [28] that RF was the most effective way to forecast mango fruit yield. A strategy for utilizing MLR and RF to predict the yield of maize, potatoes, and wheat was proposed by [29]**.** The RF considerably outperformed MLR in terms of crop yield prediction. Across a variety of crop datasets, it was found by [30] that M5-Prime and K-NN approaches performed better than other ML and LR models for agricultural yield prediction.

Knowledge representation, model structure, training time, management of missing data, and implementation costs are some factors that affect the accuracy of the previously stated methodologies. The purpose of this project is to use machine learning models based on crop and weather data to forecast the yield of Nsukka yellow pepper. The work will specifically create a pepper farm dataset that will be used to train the models, build machine learning regression models using the dataset, and quantitatively assess and validate the models' accuracy by using them to forecast pepper yield and compare the yield that was predicted to the actual yield.

2.0 MATERIALS AND METHODS

2.1 Data Collection

Crop data were collected from an experimental pepper farm (Figs. 1 and 2) at the Department of Agricultural and Bioresources Engineering, University of Nigeria, Nsukka. One hundred and seven stands of pepper were used for the project. They were all tagged using paper tape. Steel meter rule was used to measure the crop height and the measurement was taken from the ground level. The leaf length and width were also measured. The longitudinal direction was taken as the length. The measurement was taken from the apex to the base of the leaf. The transverse direction was taken as the width, measurement was taken one-third distance from the leaf base. The leaf length and width measurements were taken randomly and the average was taken for each pepper stand. These measurements were taken weekly. Yield data were recorded in terms of the weight of the harvested peppers.

Figure 1: Pepper stands 6 weeks after transplanting with drip irrigation installations

Weather data were collected from Meteorological Station of the National Centre for Energy Research and Development, University of Nigeria, Nsukka for three (3) months starting from 9th January to 9th April 2024. This period covers the time from transplanting to harvesting. A weighing balance was used to weigh the pepper after harvesting per stand.

Figure 3 is a flowchart showing the general process involved in the prediction

Figure 2: A stand of pepper with ripe fruits 10 weeks after transplanting

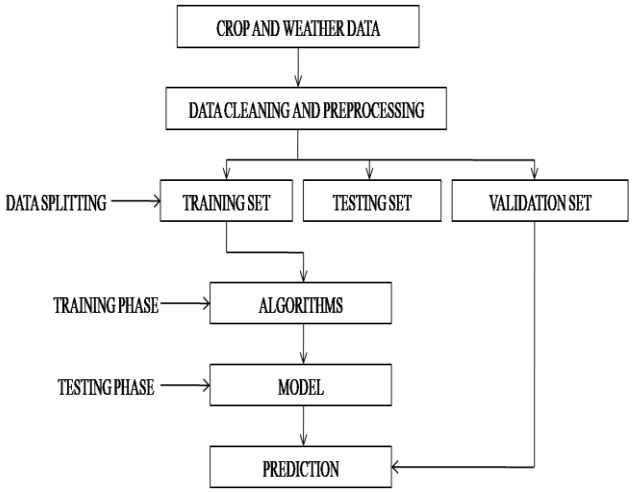

Figure 3: Flowchart of the methodology used for this study

2.2 Data Cleaning and Preprocessing

The data gathered from different data sources were tabulated, cleaned, and sorted using Microsoft Excel. Information on each of the features from the observations were collected separately to create the final dataset that was used to train the model. Resolving missing or null values was one of the pre-processing steps that each dataset went through. All of the rows with null values were eliminated since the yield data contained null values. 84 data points were used for this investigation after cleaning. Differentiating between independent and dependent variables in a dataset is crucial for machine learning. The dataset's independent variables include temperature, humidity, solar radiation, leaf length and width, leaf area, and final crop height, while the dependent variable is yield.

One of the most important aspects of data preparation is dividing the dataset into Train and Test sets, as this will improve the performance of the machine learning models. Therefore, developing a machine-learning model that performs well on both the training and test datasets is crucial. The following is how the dataset was divided: Following the consolidation of all feature variables into a single file, there were 80% training data and 20% testing data.

2.3 Feature Selection

According to an investigation by [31], choosing pertinent features is crucial since the feature subsets utilized for model training have a significant impact on a prediction model's accuracy.

A feature selection algorithm determines which subsets of the features are the best. The correlation matrix (CORR) is utilized in this investigation to determine which feature subsets offer the highest accuracy. Without cross-validation, the models are implemented using the set of features chosen using the mentioned technique. Pearson's correlation matrix, which is used to display correlations, shows coefficients among the variable groupings. There is a correlation between every randomly selected variable from the dataset and every other variable in the dataset. The correlation between the features is compared using a heat map and the P-value generated in the correlation matrix. For the given data, the Pearson's coefficient values of the variables corresponding to the yield variable are compared. Variables with comparable values between 0 and 0.9 are designated as characteristics for model training. Therefore, crop height, leaf area, temperature and solar radiations are the features that were chosen using the correlation matrix.

2.4 Experimental Models

Random Forest Regression, Decision Tree Regression, Multiple Linear Regression, and Support Vector Regression were all employed and examined in order to meet the goal of this study.

Google Colaboratory (Colab), a hosted Jupyter Notebook service was used to write the codes in Python Language for the four models.

One well-liked machine-learning technique for creating prediction models is Random Forest. Regression and classification trees, which are

https://doi.org/10.4314/njt.2025.5528

fundamental models that employ binary splits on predictor variables to forecast results, are a subset of Random Forests. The approach generates a large number of decision trees during data training so that an output can be obtained by simply calculating the average forecast of each decision tree separately. The over-fitting issue in decision trees is addressed by the Random Forest algorithm. For optimal results, a 100-tree Random Forest approach has been used in this work. A decision tree (DT) is a machine-learning technique that divides data into many segments according to a given parameter.

The decision tree has an if-then structure where each node is broken into two or more branches using only one independent variable. Whether the independent variable is continuous or categorical is irrelevant. A decision tree consists of leaf nodes that indicate final choices or forecasts, internal nodes representing characteristics or traits, and branches representing potential outcomes.

Experts frequently utilize the multiple linear regression (MLR) algorithm to forecast crop output. This model has been used by numerous academics to forecast yield in a variety of domains. The dependent variable in an MLR regression model has a linear relationship with several independent factors.

According to [32], the model equation is given as

$$Y = A + BX_1 + CX_2 + DX_3 + \cdots \quad (1)$$

Since there are numerous regression lines, A is the intercept, ε is the sum of the residual errors computed for each regression line, Y is the response, X1, X2, X3,…, and B, C, D,… are independent variables and their slopes, respectively.

In contrast to previous models, Support Vector Regression (SVR) is distinct. It uses the support vector machine (SVM) approach, commonly used to classify data samples, to make predictions about a continuous variable. In order to forecast the target variable using the input qualities or to split data points into discrete classes, it finds the optimal hyperplane. A hyperplane is a decision boundary that separates data points from different classifications. Support vector machines (SVM) are designed to find the optimal hyperplane to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class. SVM seeks to get the best possible class separation by optimizing this margin. Support vectors are the data points that are closest to the hyperplane or decision boundary. They are crucial to SVM because they set the decision boundary. The SVR requires a kernel function in order to perform linear regression on high-dimensional data.

A kernel is a function that aids in transforming a sample of lower-dimensional data into a higher-dimensional dataset. The SVR model in this work employs a linear kernel function.

2.5 Evaluation of the Models

Mean squared error (MSE), root mean squared error (RMSE), and r-squared (R2) were used to evaluate the models' accuracy and performance. Pepper yield was anticipated using the models, and Microsoft Excel was used to compare the expected and actual yield numbers.

The average squared difference between the expected and actual yields is measured by the MSE.

$$\text{Mathematically, MSE} = \frac{1}{N}\sum_1^N (Y_i - \check{Y}_i)^2 \quad (2)$$

Where.... Y_i is the actual yield,

$\check{Y}_i$ is the predicted yield,

and N is how many peppers stands there are. The square root of the MSE is the RMSE. RMSE is frequently used to compare models and can be interpreted in the same units as the expected yield.

$$\text{Mathematically, RMSE} = \sqrt{\text{MSE}} \quad (3)$$

The percentage of the dependent variable's variation that can be predicted from the independent variables is indicated by the coefficient of determination, or R^2 value.

$$\text{Mathematically, } R^2 = \left(1 - \frac{SSE}{TSS}\right), \quad (4)$$

Where, SSE is the Sum of squared Errors, Calculated as

$$\text{SSE} = \sum_1^N (\mathrm{Y_i} - \bar{\mathrm{Y}}_\mathrm{i})^2 \quad (5)$$

TSS is the Total Sum of squares, Calculated as

$$\text{TSS} = \sum_1^N (\mathrm{Y_i} - \overline{\mathrm{Y}}_\mathrm{i})^2 \quad (6)$$

$\bar{Y}_i$ is the mean of all the actual yield.

2.6 Statistical Analysis.

To strengthen the comparison between model performances, inferential statistical analyses were carried out-using Python's *pandas* and *stats models'* libraries. A one-way ANOVA was applied to determine whether significant differences existed in the mean predicted yields of the four models—Random Forest, Decision Tree, Support Vector Regression, and Multiple Linear Regression. Additionally, an Analysis of Covariance (ANCOVA) was performed with the Actual Yield as a covariate to control for differences due to natural yield variability.

The ANOVA returned an F-statistic of 25.76 ($p = 4.56\times10^{-15}$), indicating a highly significant difference among model predictions. The ANCOVA confirmed that, even after adjusting for actual yield, model type remained a significant factor ($F = 18.92$, $p = 1.34\times10^{-11}$), while the covariate itself (Actual Yield) was highly significant ($F = 169.35$, $p = 7.82\times10^{-30}$). These results statistically validate that not all models predict equally and that certain models consistently produce closer approximations to actual yields.

3.0 RESULTS AND DISCUSSION

3.1 Results

The four regression models: Random Forest, Decision Tree, Multiple Linear and Support Vector Regressions were successfully trained and tested to predict pepper yield. Predictions made by each model were compared to the actual yields. Table 1 and Figures 4 to 7 show the actual versus predicted yields for each model.

Table 1: Actual and predicted yields of the models

STAND NO	ACTUAL YIELD (grams)	Random Forest Prediction (grams)	Decision Tree Predictions (grams)	Support Vector Regression Predictions (grams)	Multiple Linear Regression Predictions (grams)
1	39	35.7	29	39.18	40.17
2	49	47.1	49	31.29	36.09
3	59	52.49	59	38.38	43.06
4	30	37.05	30	47.40	47.49
5	57	31.55	29	32.02	36.41
6	20	22.17	20	27.25	30.10
7	10	11.2	10	9.94	12.42
8	57	51.74	57	47.00	45.36
9	11	18.18	11	27.59	29.50
10	54	44.31	54	32.24	32.69
11	13	14.12	14	13.20	14.10
12	32	34.65	32	31.90	35.70
13	20	33.15	54	27.37	30.95
14	51	49.78	51	33.19	35.90
15	53	51.89	53	44.39	44.58
16	14	14.7	14	15.48	16.00
17	59	53.93	59	43.89	42.19
18	20	22.93	20	28.87	29.92
19	27	25.67	27	31.11	31.61
20	45	48.37	45	44.83	43.87
21	60	53.17	60	44.18	44.84
22	29	26.88	29	26.93	28.67
23	25	26.38	29	27.00	28.83
24	27	26.97	27	27.95	29.94
25	29	35.43	29	34.86	37.95

STAND NO	ACTUAL YIELD (grams)	Random Forest Prediction (grams)	Decision Tree Predictions (grams)	Support Vector Regression Predictions (grams)	Multiple Linear Regression Predictions (grams)
26	42	44.43	42	45.19	44.56
27	18	23.8	18	28.64	32.13
28	33	35.89	33	38.26	39.49
29	66	56.88	66	33.13	36.04
30	18	17.11	18	17.53	19.69
31	31	38.47	31	65.73	67.01
32	70	58.23	70	31.79	35.12
33	34	40.27	34	33.93	37.14
34	45	43.39	32	41.60	43.05
35	27	28.99	27	35.88	35.46
36	19	19.97	19	21.74	22.43
37	30	30.91	30	34.78	36.03
38	61	57.14	61	47.01	46.30
39	49	47.24	49	44.17	43.71
40	28	33.63	28	36.92	38.40
41	21	19.25	18	21.57	23.24
42	24	24.84	24	34.31	34.80
43	23	23.63	23	25.00	27.01
44	29	36.03	29	31.72	36.89
45	27	26.4	27	27.34	28.65
46	23	20.61	23	20.27	20.04
47	14	18.42	14	22.69	25.81
48	32	40.34	33	33.48	35.26
49	34	32.55	34	29.30	32.20
50	58	56.22	58	45.57	46.13
51	23	22.32	23	22.95	24.46
52	31	34	31	37.47	38.41
53	14	16.44	14	18.85	22.38
54	38	40.33	38	32.21	36.38
55	31	32.74	34	28.89	32.33
56	19	19.16	19	18.68	21.67
57	32	34.24	32	37.71	41.30
58	29	28.08	29	24.60	25.20
59	40	37.1	34	35.22	39.61
60	15	14.89	15	11.92	12.92
61	29	29.52	29	30.14	31.93
62	31	32.03	31	35.74	37.43
63	32	32.85	32	30.41	34.19
64	32	31.39	32	33.24	34.68
65	25	22.7	25	23.93	28.46
66	39	39.73	34	40.85	40.91
67	49	46.02	49	29.38	33.69

https://doi.org/10.4314/njt.2025.5528
http://creativecommons.org/licenses/by-nc-nd/4.0/

STAND NO	ACTUAL YIELD (grams)	Random Forest Prediction (grams)	Decision Tree Predictions (grams)	Support Vector Regression Predictions (grams)	Multiple Linear Regression Predictions (grams)
68	57	54.31	57	39.66	43.49
69	30	35.87	30	43.78	45.48
70	56	43.31	56	32.51	38.58
71	20	24.41	25	27.32	31.65
72	10	11.21	10	10.33	12.84
73	57	52.48	57	41.19	41.47
74	45	44.41	32	41.67	42.73
75	27	27.18	27	32.30	34.12
76	19	18.13	19	20.13	21.92
77	30	31.66	30	34.74	37.19
78	31	31.2	31	29.83	31.20
79	19	18.73	19	19.22	20.37
80	32	34.49	32	36.14	41.06
81	29	28.69	29	26.29	27.19
82	40	46.65	59	36.63	41.33
83	15	14.89	15	11.92	12.94
84	29	28.86	29	28.57	30.82

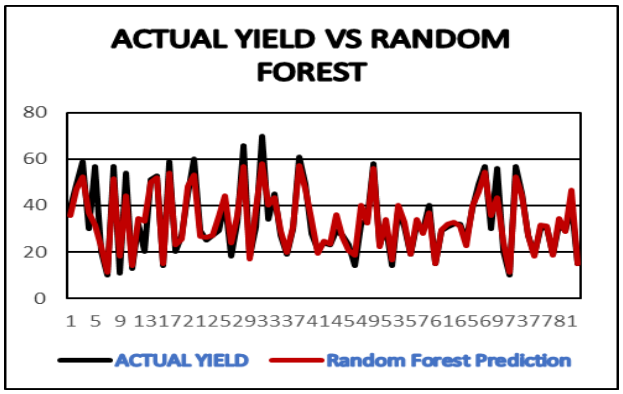

Figure 4: Actual yield vs random forest.

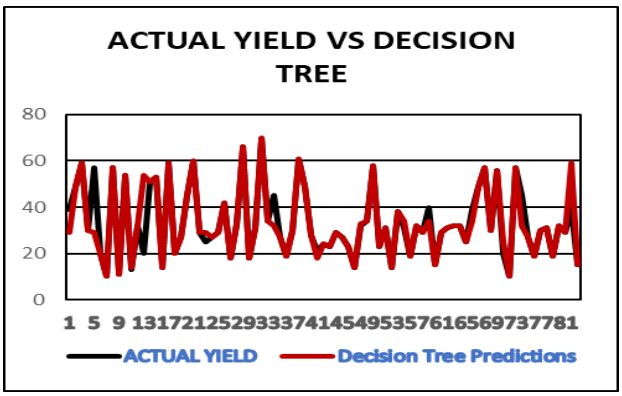

Figure 5: Actual yield vs Decision Tree

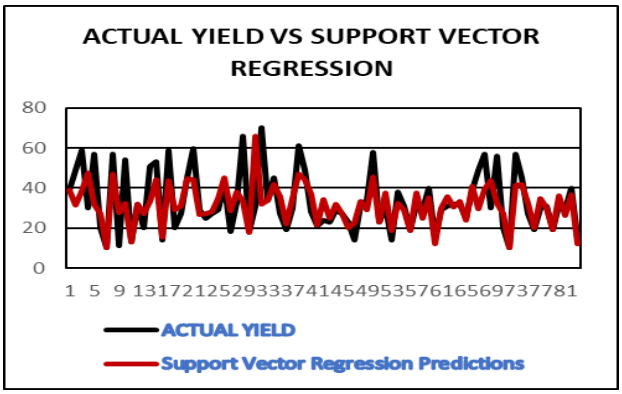

Figure 6: Actual yield vs support vector regression

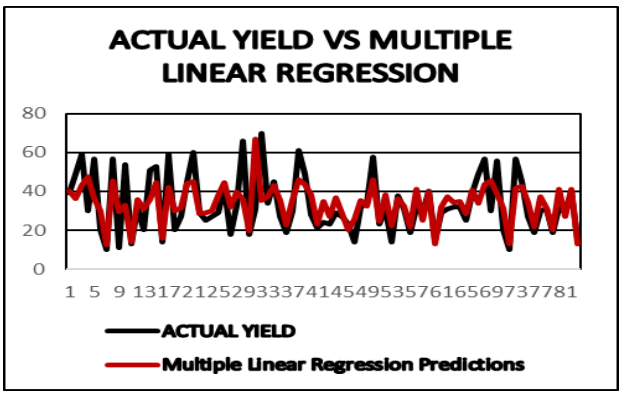

Figure 7: Actual yield vs multiple linear regression.

The outcomes of the four models are shown in Table 2. The Random Forest model yielded a mean squared

error (MSE) of 58.15, a root mean squared error (RMSE) of 7.62, and an R-squared value of 0.58.

The Decision Tree model, on the other hand, reported an R-squared value of -0.21, an RMSE of 12.97, and an MSE of 188.29. Finally, MSE values of 48.69 and 45.23, RMSE values of 6.98 and 6.73, and R-squared values of 0.50 and 0.67 were shown by the Support Vector Regression and Multiple Regression models, respectively.

Table 2: Regression models and Performance metrics scores

Model	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R-Squared	ANOVA F-statistic (p-value)	ANCOVA F-statistic (p-value)
Random Forest	58.15	7.62	0.58		
Decision Tree	168.29	12.97	-0.21		
Support Vector Regression	48.69	6.98	0.65		
Multiple Linear Regression	45.23	6.73	0.67		
Overall Model Comparison				25.76 (4.56e-15)	18.92 (1.34e-11)

3.2 Discussions

The models' varying efficacy is revealed by the performance evaluation based on MSE, RMSE, and R2. For this particular dataset, a linear approximation reasonably reflects the link between the independent variables (plant height, leaf dimensions, and climatic data) and yield, according to the better performance of the MLR model (R^2 = 0.67). The SVR model's capacity to simulate possible non-linear patterns is demonstrated by its comparable performance (R^2 = 0.65). While the catastrophic failure of the single DT (R^2 = -0.21) emphasizes its instability and unsuitability for this prediction, the modest performance of RF (R^2 = 0.58) may be explained by overfitting on a smaller dataset.

3.3 Statistical Validation and Comparison with Previous Studies

One-way ANOVA and ANCOVA were used to support the results. The statistical significance of the variations in mean expected yields between the models is confirmed by the significant ANOVA result ($F=25.76$, $p<0.001$). The trustworthiness of the metric-based ranking is further confirmed by the ANCOVA ($F=18.92$, $p<0.001$), which controls for real yield and confirms that the performance differences are intrinsic to the algorithms rather than data aberrations.

This result both follows and deviates from patterns found in earlier studies on crop yield prediction. Studies like [17], who discovered that a hybrid MLR-ANN model initialized with MLR coefficients beat numerous complicated methods, are consistent with the robust performance of simpler models like MLR. Similar to this, researchers frequently find that a single Decision Tree performs poorly, which is why they choose ensemble approaches like Random Forest [19]. The outcome, however, is in contrast to research such as [29], where RF significantly outperformed MLR for forecasting yields of maize, potatoes, and wheat, and [15], where LSTM networks dramatically outperformed multivariate regressions and RF for soybean yield.

This disparity highlights a crucial finding: crop type, data quality, feature set, and size all have a significant impact on the ideal model. As shown here for Nsukka yellow pepper, traditional linear methods can be remarkably effective and efficient for smaller, well-defined agronomic datasets, while ensemble and deep learning models frequently perform well with large, multi-source datasets (e.g., satellite imagery fused with climate data [16, 26]).

4.0 CONCLUSION AND FUTURE WORK

In conclusion, the Multiple Linear Regression model, which explains 67% of the yield variance, turned out to be the most reliable and accurate method for forecasting the yield of Nsukka yellow pepper based on the gathered crop and meteorological characteristics. Although this offers a useful baseline for decision-support, an R^2 of 0.67 shows that 33% of yield variability is still unaccounted for.

Future research should consider the following suggestions in order to increase predicted accuracy toward a goal of 0.90 R^2:

-Feature Development and Growth: More predictive variables could be added, such as soil nutrient content, specific management methods (e.g., irrigation schedules, fertilizer types and dosages), precise pest/disease incidence data, and more detailed microclimate data.

-Advanced Modeling Techniques: More complex algorithms that have demonstrated great potential in other agricultural research could be investigated, such as: -Gradient Boosting Machines: Known for their great accuracy with tabular data, such as XGBoost and LightGBM.

-Deep Neural Networks (DNNs) or Long Short-Term Memory (LSTM) Networks: Especially when there are more intricate, non-linear interactions or temporal data sequences.

-Hybrid or Ensemble Models: According to [13] and [17], integrating the advantages of many models (such as the recommended MLR-ANN hybrid or stacked ensembles) can frequently result in better performance.

-Data Enhancement: To improve model generalizability and robustness and enable the successful training of more complicated models, future work could expand the dataset size across several growth seasons, locales, and cultivars.

-Hyperparameter Tuning: The RF and SVR models' performance may be greatly enhanced with ideal parameter choices, possibly outperforming the existing MLR baseline. Therefore, carrying out thorough hyperparameter optimization for these models could improve the performance of the models.

-Expanded Geographical Application: Future work should apply the created framework to localized data from other places to find patterns in yellow pepper production that are both universal and region-specific. This will help Nigerian farmers create a precision agriculture tool that is scalable.

5.0 ACKNOWLEDGEMENT

We thank the Department of Agricultural and Bioresources Engineering, UNN and the National Centre for Energy research and Development, UNN for providing weather data for research purposes.

REFERENCES

[1] Alawode,O.O. Obegunde,V.O. "Economic Analysis of Pepper Marketing in Oyo State, Nigeria," *Applied Tropical Agriculture.*, 21(3), 116-121, 2016.

[2] D.C. Raoul-Fani,D.C. Ukpe,H.U. . Ngo,N.V. Mohamadou,S. Adedze,M. Pemunta,N.V. "Perceived effects of climate change on profit efficiency among small scale chili pepper marketers in Benue State, Nigeria, *GeoJournal*, 86, pages 1849–1862, 2021

[3] Gobie,W. "A seminar review on red pepper (Capsicum) production and marketing in Ethiopia", *Cogent Food & Agriculture*, 5:1, 1647593, 2019, DOI: 10.1080/23311932.2019.1647593.

[4] Ebe,F.E. Obike,K.C. Oti,G.O. Njoku,J.I.K. Abonyi,J.O. "Analysis of resources use, productivity and profitability among yellow pepper rural female farmers in Nsukka Agricultural Zone of Enugu State, Nigeria," *Faman Journal*, 21(1), 2021.

[5] Olutumise,A.I. "Determinants of market participation and preference for production of pepper (*Capsicum* spp.) among farmers in southwest, Nigeria, *Heliyon*, 8, 9e10585 2022.

[6] Maspaitell,M. Garnevska,E. Siddique,M.I. Sha,N. "Towards high value markets: a case study of smallholder vegetable farmers in Indonesia," *International Food and Agribusiness Management Review*, 21(1), pp. 73–88, 2023.

[7] Onwubuya,E.A. Okporie,E.O. Nenna,M.G. "Nsukka yellow pepper: Processing & preservation techniques among women farmers," *African Journal of Agricultural Research*, 4(9), pp. 859–863, 2009.

[8] van Klompenburg,T. Kassahun,A. Catal,C. "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture.*, 214, 108298, 2023.

[9] Saruk,B.S. Mokesh,R.G. "Optimizing rice yield prediction: A policy-oriented approach to agricultural sustainability," *Results in Engineering*, 28, 2025, 107844, https://doi.org/10.1016/j.rineng.2025.107844

[10] Liu,S.C. Jian,Q.Y. Wen,H.Y. Chung,C.H. "A Crop Harvest Time Prediction Model for Better Sustainability, Integrating Feature Selection and Artificial Intelligence Methods," *Sustainability*, 14, 14101, 2022.

[11] Zhang,L. Zhang,Z. Luo,Y. Cao,J. Tao,F. "Combining Optical, Fluorescence, Thermal Satellite, and Environmental Data to Predict County-Level Maize Yield in China Using Machine Learning Approaches," *Remote Sensing.*, 12(1), 21, 2020.

[12] Shah Hosseini,M. Martinez-Feria,R.A. Hu,G. Archontoulis,S.V. "Maize yield and nitrate loss prediction with machine learning algorithms," *Environmental Research Letters*, vol. 14, no. 12, 124026, 2019.

[13] Chang,Y. Latham,J. Licht,M. Wang,L. "A data-driven crop model for maize yield prediction," *Communications Biology*, 6(439), 2023.

[14] Qaddoum,K. Hines,E.L. "Reliable Yield Prediction with Regression Neural Networks," *WSEAS Trans. Syst. Control*, 7(1), pp. 1–10, 2012.

[15] Schwalbert,R.A. Amado,T. Corassa,G. Pott,L.P. Prasad,P. Ciampitti,I.A. "Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil," *Agricultural and Forest Meteorology*, 284, 107886, 2020.

[16] Maimaitijiang,M. Sagan,V. Sidike,P. Hartling,S. Esposito,F. Fritschi,F.B. , "Soybean yield prediction from UAV using multimodal data fusion and deep learning," *Remote Sens. Environ.*, 37, 111599, 2020.

[17] Gopal,P.M. Bhargavi,R. "A novel approach for efficient crop yield prediction," *Computers and Electronics in Agriculture*, 165, 104968, 2019.

[18] Shiu,Y.S. Chuang,T.C. "Yield Estimation of Paddy Rice Based on Satellite Imagery: Comparison of Global and Local Regression Models," *Remote Sensing*, 11(2), p. 111, Jan. 2019, doi: 10.3390/rs11020111.

[19] Patil, Y. Ramachandran, H. Sundararajan, S. Srideviponmalar, P. "Comparative Analysis of Machine Learning Models for Crop Yield Prediction Across Multiple Crop Types," *SN Computer Science, 6(1), p.64*, 2025. https://doi.org/10.1007/s42979-024-03602-w

[20] Awad,M. "Toward Precision in Crop Yield Estimation Using Remote Sensing and Optimization Techniques," *Agriculture*, 9(3), 54, 2019.

[21] Guo,W.W. Xue,H. "Crop Yield Forecasting Using Artificial Neural Networks: A Comparison between Spatial and Temporal Models," *Mathematical Problems Engineering*, 2014, pp. 1–7, 2014.

[22] Cai,Y. Guan,K. Lobell,D. Potgieter,A.B. Wang,S. Peng,J. Xu,T. Asseng,S. Zhang,Y. You,L. Peng,B. "Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches," *Agricultural and Forest Meteorology*,274, pp. 144–159, 2019.

[23] Han,X. Chen,L. Zhao,S. Li,J. "Improving winter wheat yield forecasting based on multi-source data and machine learning," *Agriculture*, (5), 571, 2020.

[24] Bhojani,S.H. Bhatt,N. "Performance Analysis of Activation Functions for Wheat Crop Yield Prediction," *IOP Conference Series.: Matererial Science and Engineering*, 1042, 012015, 2021.

[25] Nevavuori,P. Narra,N. Lipping,T. "Crop yield prediction with deep convolutional neural networks," *Computers Electronics in Agriculture*, 163, 104859, 2019.

[26] Zhang,L. Zhang,Z. Luo,Y. Cao, J. Tao,F. "Combining Optical, Fluorescence, Thermal Satellite, and Environmental Data to Predict County-Level Maize Yield in China Using Machine Learning Approaches," *Remote Sensing*, 12(1), 21, 2020.

[27] Abbas,F. Afzaal,H. Farooque,A.A. Tang,S. "Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms," *Agronomy*, 10(7), 1046, 2020.

[28] Fukuda,S. Spreer,W. Yasunaga,K.Y. Sardsud,V. Müller, J. "Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes," *Agricultural Water Management*, 116, pp. 142–150, 2013.

[29] Jeong,J.H. Resop,J.P. Mueller,N.D. Fleisher,D.H. Yun,K. Butler,E.E. Timlin,D.J. Shim,K.M. Gerber,J.S. Reddy,V.R. Kim,S.H. "Random Forests for Global and Regional Crop Yield Predictions," *PloS one, 11(6), p.e0156571*, 2016.

[30] Gonzalez-Sanchez,A. Frausto-Solis,J. Ojeda-Bustamante,W. "Predictive ability of machine learning methods for massive crop

yield prediction," *Spanish J. Agric. Res.*, 12(2), pp. 313–328, 2014.

[31] Hooda,B.K. Hooda,E. "Feature selection using matrix correlations and its applications in agriculture," *International Journal of Applied Mathematics & Statistical Sciences (IJAMSS).*, 10(1), pp. 13–20, 2021.

[32] Priya,P. Muthaiah,U. Balamurugan,M. "Predicting Yield of the Crop Using Machine Learning Algorithm," *International Journal of Engineering Sciences & Research Technology*, 7(1), pp.1-7, 2015.