

An Exploratory Data Analysis of Growth, Pest Infestation and Treatment Dataset Towards Improved Prediction of Cowpea Yield

R. S. Babatunde^{1*}, A. Musa², J. F. Ajao¹, R. M. Isiaka¹, J. A. Ojo³, L. O. Abdulrahman³ and A. K. Mohammed¹

¹Department of Computer Science. Kwara State University, Malete, Kwara State. Nigeria.

²Department of Electrical and Electronics Engineering. Kwara State University, Malete, Kwara State. Nigeria.

³Department of Agriculture and Crop Production, Kwara State University, Malete, Kwara State. Nigeria.

ARTICLE INFO

Received: February 2025

Accepted: April 2025

Keywords:

Machine Learning, Pattern Recognition, Agriculture, Optimal Production, Crop Management

Corresponding author:

ronke.babatunde@kwasu.edu.ng

ORCID ID:

0000-0002-9286-5110

Phone No: +2347065259708

DOI: 10.55518/fjpas.MURU5824

Abstract

The rising need for food security requires creative agricultural methods, especially in the farming of essential crops such as cowpea (*Vigna unguiculata*). This research introduces an improved automated forecasting system for cowpea production utilizing machine learning methods, drawing on an extensive dataset comprising factors like plant growth metrics, pest presence, and treatment methods. Utilizing regression models and sophisticated algorithms, the system examines past growth data to determine significant factors affecting yield results. Preliminary findings indicate a strong relationship between growth indicators like plant height at different developmental phases and the overall yield. The predictive model will not only enable prompt actions in crop management but also assists farmers in making refined choices to enhance production. Ultimately, this study aids in crafting precision agriculture techniques that improve cowpea production while tackling issues arising from pests and environmental factors. This automated method indicates a way to enhance agricultural output and sustainability in areas dependent on cowpea as an essential food source.

1.0 Introduction

Cowpea (*Vigna unguiculata*) is an important legume crop, and it grows mainly in the subtropical and tropical regions of the world, including sub-Saharan Africa where it is a major source of protein and other nutrients. The desiccation tolerance of the crop makes it an important part of sustainable agricultural systems. However, the provision of optimum cowpea yields is still a problem due to different biotic and abiotic factors such as pests, diseases and environmental factors. Nevertheless, the resistance of cowpea to various biotic stresses is quite high, the production of the crop is still affected by various biotic factors including pest infestations. Some pests that affect cowpea include aphids, pod borers and the cowpea weevil which are pests that are destructive to the crop resulting in reduced growth and yield. This is important in the view of the facts that the dynamics of pest infestation and its impacts on cowpea production vary with the agroecological zone and the cropping system [1]. In the face of increasing food demand, there is the need to improve crop production and productivity to meet the food needs of the increasing population. The advancement in data analytics and machine learning has provided a new approach to yield prediction through the use of historical data and essential factors that affect yield [2]. The current study is an exploratory analysis that seeks to predict cowpea yield using a dataset that comprises growth parameters, PGPR application and disease management, as well as environmental factors.

The significance of combining various datasets to increase prediction accuracy has been emphasized in earlier research [3], [4]. Comprehensive data on cowpea varieties

exposed to various treatments and the associated growth indicators over time are included in the dataset used in this study. This study looks for patterns in the data that are associated with yield results by using statistical analysis and machine learning approaches. Determining these connections can yield important information about the agronomic techniques that optimize cowpea yield. Moreover, understanding the factors that influence cowpea yield can significantly aid farmers in making informed decisions regarding crop management, resource allocation, and pest control measures. Infestations of pests have been found to cause direct harm by feeding on plant tissues, which lowers the amount of harvestable yield and compromises the general health of the plants. Aphid infestations, for example, can spread viral infections like Cowpea Mosaic Virus (CMV), making pest control more difficult and increasing yield losses [5]. Furthermore, the employment of chemical management methods is frequently required when pests are present, which can have an adverse effect on soil health and the environment. In order to reduce insect impacts on cowpea, recent research have emphasized the significance of integrated pest management (IPM) strategies that combine biological control techniques with cultural practices [6]. These strategies aim to reduce reliance on chemical pesticides while promoting sustainable agricultural practices. Furthermore, advancements in precision agriculture and data analytics provide opportunities for real-time monitoring of pest populations and their effects on crop health.

The purpose of this study is to investigate how cowpea growth characteristics, including plant height, leaf count, and total yield, relate to the degree of insect

infestation. The goal of this study is to determine the main variables affecting cowpea productivity in the presence of pests by examining data gathered from field tests that evaluate different pest management techniques. The results will help farmers and other agricultural stakeholders better understand how pest dynamics affect cowpea development and yield, which will ultimately lead to more informed management strategies. This study not only aims to contribute to the existing body of knowledge on cowpea cultivation but also aspires to support sustainable agricultural practices that enhance productivity while addressing challenges posed by climate change.

2. Related Work

Shuaibu *et. al* [7] addresses the challenges in predicting seasonal crop yields using traditional methods, which often struggle with accuracy and adaptability to changing conditions. It highlights the need for improved prediction techniques that can leverage modern technology and data analysis. The authors developed a prediction model that integrates machine learning techniques with software development to enhance the accuracy of crop yield predictions. They also explored the application of reinforcement learning for crop rotation planning, aiming to optimize agricultural decision-making. The new prediction model demonstrated significant improvements in accuracy, particularly in regions where irrigation is prevalent, indicating better crop performance during dry seasons. The study also found that the proposed methods could effectively support agricultural practices and decision-making. This work contributes to the field of agricultural technology by bridging the gap

between manual and automated prediction techniques, encouraging further research into machine learning applications in agriculture, and providing a framework for future studies on crop management and optimization. Mayank *et. al.* [8] focuses on crop yield prediction using machine learning and how climate change negatively impacts agricultural crop performance in India. An interactive prediction system prototype is developed for farmers. Random forest algorithm was utilized for accurate yield predictions. The system provides predictions based on climatic data inputs. The developed user-friendly web interface enhances accessibility for farmers. Bera *et. al.* [9] developed a system that Enhance crop yield by early detection of plant nutrition deficiencies and diseases using deep learning. The work introduces PND-Net, integrating Graph Convolutional Networks (GCN) with Convolutional Neural Networks (CNN) for improved feature representation. The research achieved state-of-the-art accuracy on multiple datasets, including 90.00% for banana and 96.18% for potato diseases. The implication of the research is a demonstration of state-of-the-art performance in agricultural growth. Morales and Villalobos [1] proposed a system focused on predicting crop yields for sunflower and wheat using various machine learning algorithms, particularly assessing their effectiveness and accuracy. The authors used two machine learning models, including Random Forest (RF) and Artificial Neural Networks (ANN), and compared their performance using metrics like Root Mean Square Error (RMSE) and Coefficient of Determination (R^2). Data partitioning strategies were also analyzed to enhance predictive accuracy. The results obtained

shows that Random Forest demonstrated superior performance with RMSE values between 35-38%, while ANN and linear models yielded higher RMSE. The study highlighted the importance of comparing ML models against simple baseline algorithms. The major insight from the research is the recommendation of the effectiveness of machine learning for crop yield prediction, emphasizing the reliability of RF and the need for careful data handling and model comparison to improve agricultural forecasting.

Islam *et. al.* [2] develop an AI-enabled leaf disease detection system using deep learning models to assist farmers in identifying plant diseases through image analysis. The work aims to reduce economic losses in agriculture by enabling timely disease detection and treatment recommendations. The study employed multiple convolutional neural networks (CNNs), including VGG-16, VGG-19, and ResNet-50, with ResNet-50 achieving the highest accuracy of 98.98%. A web application was developed to allow users to upload images for disease detection and treatment recommendations. The results demonstrated that the proposed model significantly improves the accuracy of disease detection, aiding in better diagnosis and agricultural productivity. The research contributes to the field by providing a practical tool for farmers and advancing machine learning techniques in agriculture. Nimbale and Mareeswari [10] conducted a comprehensive review of state-of-the-art ML techniques for crop yield prediction, highlighting recent advances and future directions. The study synthesized findings from various studies, focusing on supervised learning techniques (e.g., regression analysis, decision trees, neural networks) and data

sources (e.g., meteorological, soil, remote sensing). The result of the review reveals that ML models, particularly deep learning, demonstrated promising results across different crops and regions, with performance influenced by data quality and model architecture. The research provides valuable insights for researchers, practitioners, and policymakers, promoting improved agricultural productivity, sustainability, and the empowerment of small-scale farmers through actionable ML models. Kavitha *et. al.* [11] developed a system for identifying medicinal plants using deep learning. The work used MobileNet for automatic identification of medicinal leaves. Images of six medicinal plants were collected and augmented. The MobileNet model achieves 98.3% accuracy in plant identification. The model was incorporated into a mobile app which enables real-time identification of medicinal leaves. The study addresses challenges in manual plant identification. Ahmed *et. al.* [12] developed a system for crop yield prediction in Nigeria, utilizing machine learning techniques. The study adopted three (3) machine learning approaches, including a decision tree classifier, random forest, and support vector machine, to model data from different zones and make predictions. Decision tree classifier showed the best performance in rainy season. The work revealed that Southeast Nigeria has the highest yield accuracy. The work recommended the consideration of weather and soil texture in the prediction of crop yield. Sunkara and Beulet [13] proposes the use of machine learning applications in agriculture. The work developed a model for predicting crop yield using KNN algorithm. Environmental parameters like soil moisture and temperature are analyzed. KNN shows

high accuracy in crop yield predictions. The work recommended the inclusion of more influencing factors for building robust yield prediction models. Karan and Farhana [14] predicts crop yield using Artificial Neural Networks (ANN). The work seeks to addresses farmers' challenges by considering soil quality and environmental factors as they influence crop selection. The proposed system takes into consideration the data related to environmental factor, soil factor, weather, fertility of soil and past year production and suggests the best profitable crops that can be cultivated in the given environment condition [15]. The work concluded that ANN is promising, playing a major role in future evaluation of crop yields based on the present/past data.

3. Methodology

The experimental investigation was carried out on a desktop computer with an Intel Core i9-7900X Skylake processor, an NVIDIA GTX 1080 Ti 11GB graphics card, and 64GB of RAM utilizing the Keras framework (version 2.3) and TensorFlow (version 1.14). Exploratory Data Analysis (EDA), the first step in the process, was carried out using sophisticated statistical techniques from the scikit-learn Python library. In this first step, the dataset was carefully examined and analyzed with an emphasis on determining the connections among crop production, crop treatments, and soil characteristics.

The significance of mean differences between different groups within the data was

assessed using statistical tests like ANOVA and t-tests in addition to the EDA. A variety of graphical representations, such as box plots, violin plots, histograms, and bar charts, were used to display the results. These visualizations effectively conveyed the results and insights derived from the analysis.

3.1 Description of Dataset.

The dataset used in this research was collected from a farmland in Kwara State University, Nigeria. It is structured into four main segments:

- i. Treatment: This segment details the different crop treatments applied.
- ii. Yield Data: This includes measurements of grain yield weight.
- iii. Plant Growth Data: This encompasses metrics such as pod length and the number of pods.
- iv. Pest Infestation Data: This covers the presence of specific pests, namely **Megalurothrips sjostedti** and **Aphis craccivora**.

Table 1 summarizes the crop yield dataset, outlining key parameters and variables crucial for subsequent analysis and modeling. To enhance clarity and facilitate interpretation, abbreviated codes were used for longer variable names. Additionally, the dataset was supplemented with meteorological data obtained from the local weather station's website for the farm region.

.....

Table 1: Description of Dataset Variables

Category	Variable Name	Code_ABB	Data Type
Treatment	A	Treatment	object
	B		
	C		
	D		
	E		
POD	Pod length at 65 days after planting (cm)	POD_LENGTH_65	float64
	Number of pod per plant at 60 days after planting	POD_PLANT_60	float64
	Pod load scoring	POD_SCORE	int64
Observations/Infestations	Maruca vitrata % of flower infestation and damage	MVF_INF	int64
	Megahurothrips sjostedti observation 1st observation	MSJ_OBS_1ST	float64
	Megahurothrips sjostedti observation 2nd observation	MSJ_OBS_2ND	float64
	Megahurothrips sjostedti observation 3rd observation	MSJ_OBS_3RD	float64
	Megahurothrips sjostedti observation 4th observation	MSJ_OBS_4TH	float64
	Megahurothrips sjostedti observation 12th observation	MSJ_OBS_12TH	float64
	Aphis craccivora infestation observation 1st	ACINF_OBS_1ST	float64
	Aphis craccivora infestation 2nd observation	ACINF_OBS_2ND	float64
	Aphis craccivora infestation 3rd observation	ACINF_OBS_3RD	float64
	Aphis craccivora infestation 4th observation	ACINF_OBS_4TH	float64
	Aphis craccivora infestation 5th observation	ACINF_OBS_5TH	float64
	Aphis craccivora infestation 6th observation	ACINF_OBS_6TH	float64
Grain Yield	Weight of grain yield per hill of 20 plants (g)	GY_PER_20	float64
	Net plot grain weight (g)	NP_Grain (g)	int64

3.2 Experimental Pre-processing of Data

Preprocessing is essential for ensuring data quality and preventing errors during analysis and model training. The process began by merging the soil nutrient data with the soil particle data using the `SampleID` as a key. During data cleaning, it was discovered that the soil nutrient data had missing values. These were addressed using the method proposed by [16], which involved replacing missing values with the average value from each column. Additionally, the cowpea yield dataset had 10 missing data points, which were removed to prevent anomalies and bias in model deployment. The dataset includes several categorical variables, such as Treatment, Soil Classes, and Sample Specimen, as outlined in Table 1. To prepare these categorical variables for analysis and machine learning training, numerical values were assigned using the `LabelEncoder` class

from scikit-learn. This step facilitates the encoding of categorical data for further processing. Key features that significantly influence yield outcomes were identified from the merged and cleaned data. Spearman's rank correlation coefficient (Eq. 1) was used to ensure data consistency during feature selection. Features with absolute correlation coefficients greater than 0.35 were selected as crucial for model training.

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (M_i - \bar{M})^2}} \quad (1)$$

r symbolizes the coefficient outcomes. ' y_i ' represents the actual occupancy values pertinent to the prediction tasks, while ' M_i ' corresponds to the values of each feature. Further clarification stems from $\bar{y}$ and $\bar{M}$, signifying the average values of ' y_i ' and ' M_i ', respectively. It is noteworthy that the proximity of the absolute value of r to 1 or -

1 indicates a heightened degree of correlation. To facilitate analysis and comparison, the dataset was segmented by treatments and replications. Data normalization was then performed to improve data quality and consistency. The 'MinMaxScaler' function from scikit-learn was employed for normalization, scaling values to a range between 0 and 1. This step is critical for accurate data evaluation and interpretation, as detailed in Eq. (2).

$$A_{normalized} = \frac{A - A_{minimum}}{A_{maximum} - A_{minimum}} \quad (2)$$

Where A is the attribute data considered, and the respective minimum and maximum value for each attribute data considered.

3.3 Data Interpretation and Exploratory Data Analysis

The yield statistic based on the treatment categories is presented in Table 2. The yield statistics indicate that Treatment C produced the highest average yield (417.5 units), although with significant variability (Std = 112.27 units), suggesting a trade-off between high yield and consistency. Treatment B, while yielding lower on average (327.5 units), exhibited the least variability (Std = 49.76 units), making it the most consistent and reliable option. Treatment D showed relatively high yields (Mean = 356.0 units) but with considerable variability and a skew towards lower yields. Treatment A presented moderate yields with low variability, while Treatment E had the lowest yield (136.5 units) and moderate variability, indicating it was the least effective among the treatments in terms of yield output. The mean yield per treatment on each observation is shown in Figure 1.

Table 2: Yield Statistics by Treatment

Treatments	Mean	Median	Std	Min	Max
A	296.5	304.0	56.7	224	354
B	327.5	342.0	49.76	260	366
C	417.5	426.0	112.27	278	540
D	356.0	393.0	130.61	176	462
E	136.5	119.0	41.29	110	198

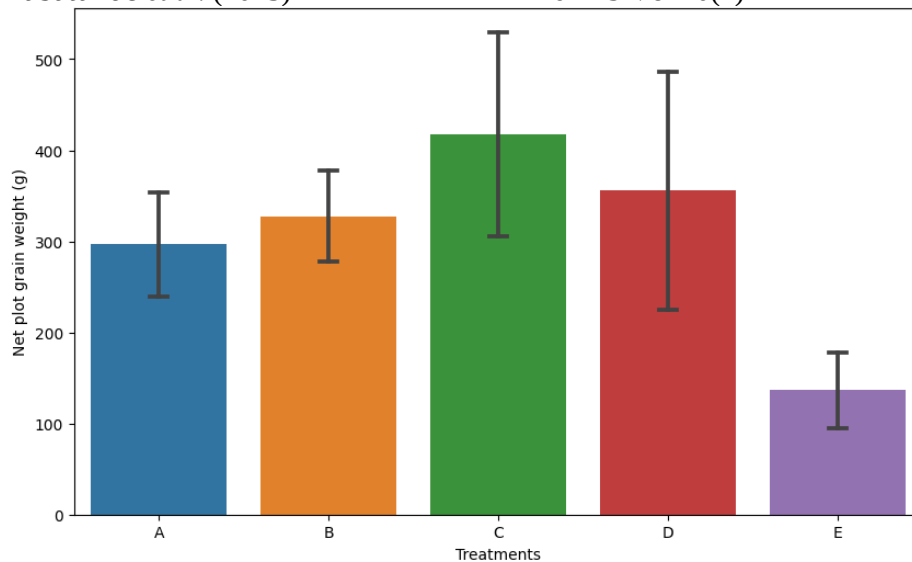


Figure 1: Mean yield per Treatments class on each observation and net grain

The soil class distribution presented in Table 3 shows that most of the samples are Loamy Sand (13 counts), followed by Sand (4 counts) and Sandy Loam (3 counts). The mixed class of Sand/Loamy Sand and Sandy Clay Loam were less common, each with 3 and 2 counts respectively. This distribution indicates a predominance of loamy sand, with fewer instances of pure sand and other sandy soil types.

Table 3: Soil class distribution

Soil Class	count
LOAMY SAND	13
SAND	4
SANDY LOAM	3
SAND/LOAMY SAND	3
SANDY CLAY LOAM	2

The one-way ANOVA results presented in Table 4 indicate a statistically significant difference between the treatment group means, as evidenced by an F-statistic of 5.9653 and a p-value of 0.0044. Since the p-value is below the conventional threshold of 0.05, we can reject the null hypothesis, suggesting that at least one group's mean is significantly different from the others.

Table 4: One-way ANOVA results

Metrics	Value
F-statistic	5.9653
p-value	0.0044

The pairwise t-test results presented in Table 5 reveal that most comparisons between treatments do not show statistically significant differences, as indicated by p-values greater than 0.05. Specifically,

comparisons between Treatments A and B, A and C, A and D, B and C, B and D, and C and D all have p-values above 0.05, suggesting no significant differences in mean yields between these pairs. However, significant differences were found in the comparisons of Treatments A vs. E ($p = 0.0038$), B vs. E ($p = 0.001$), C vs. E ($p = 0.0033$), and D vs. E (p

$= 0.0185$), indicating that Treatment E's mean yield is significantly different from all other treatments. This suggests that Treatment E is statistically distinct in its effect on yield compared to the other treatments.

Table 5: Pairwise t-test results

Treatment 1	Treatment 2	t-statistic	p-value
A	B	-0.8219	0.4425
A	C	-1.9242	0.1027
A	D	-0.8358	0.4353
A	E	4.5624	0.0038
B	C	-1.4658	0.1931
B	D	-0.4078	0.6976
B	E	5.908	0.001
C	D	0.7142	0.5019
C	E	4.6983	0.0033
D	E	3.2048	0.0185

4. Results and Discussion

The result of the exploratory analysis of the data in this research is presented in this section. The inference from the experiment and feature importance of the variables in the dataset are also discussed.

4.1. Result of Analysis of Plant Growth and Yield Dataset

The Spearman's rank correlation heatmap for all variables, identifying the strength and

direction of the relationships between them is presented in Figure 2. Notable correlations include a strong positive relationship between "GY_PER_20" and "NP_Grain (g)" (0.59), "POD_LENGTH_65" and "POD_PLANT_60" (0.67), as well as a negative correlation between "POD_SCORE" and both "NP_Grain (g)" (-0.53) and "POD_LENGTH_65" (-0.60). The matrix highlights crucial features with absolute correlation coefficients greater than 0.35, which are critical for model training.

For instance, "MVF_INF" and "POD_SCORE" (0.78) also show a significant positive correlation.

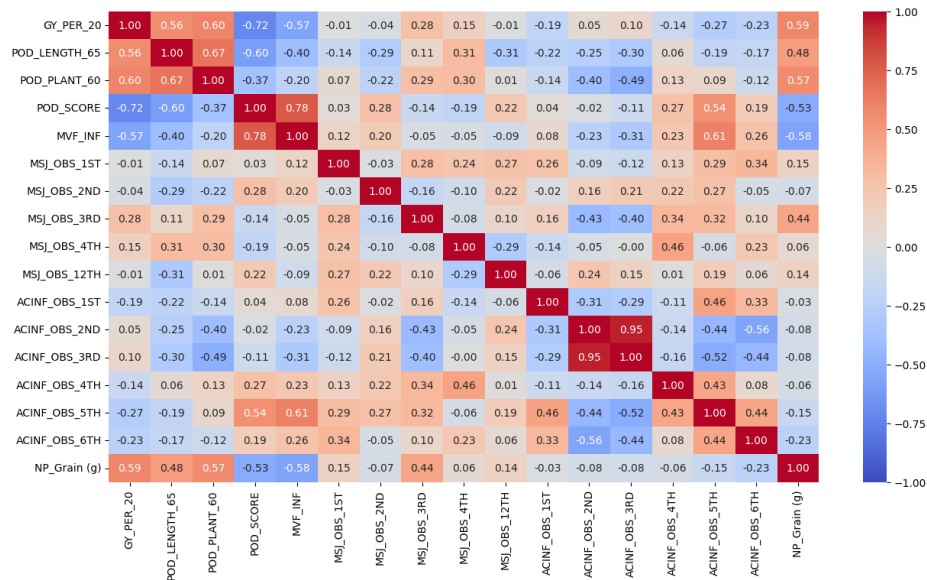


Figure 2: Correlation heatmap of all variables

The correlations among the POD variables is presented in Figure 3. "POD_LENGTH_65" and "POD_PLANT_60" are positively correlated (0.67), indicating that plants with longer pods tend to have more pods per plant. Conversely, "POD_SCORE" is negatively correlated with "POD_LENGTH_65" (-0.60), "POD_PLANT_60" (-0.37), and "NP_Grain (g)" (-0.53), suggesting that a higher pod score might be associated with shorter pod length and fewer pods per plant, as well as lower grain weight. These correlations, especially those with absolute values greater than 0.35, highlight important variables for predictive modeling in agricultural studies.

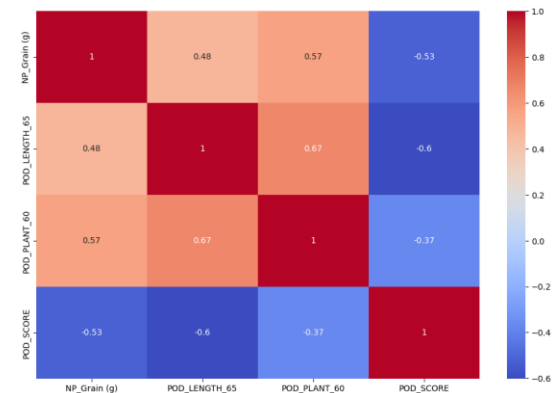


Figure 3: Correlation score of POD variables

The boxplot illustrated in Figure 4 shows the distribution of the various crop yield variable, highlighting differences in central tendencies and variability. Metrics like "GY_PER_20" and "ACINF_OBS_6TH" show higher median values, suggesting higher central values, while others, such as "POD_LENGTH_65" and "ACINF_OBS_2ND," display outliers, indicating variability or anomalies. The implication is that the POD_LENGTH and

ACINF_OBS_2ND are variables which might not have direct effect on crop yield as also observed in [14] and [15]. The varying box widths reflect different levels of data dispersion, with "ACINF_OBS_1ST" showing less variability and metrics like "POD_SCORE" and "MSL_OBS_1ST" exhibiting greater spread. Overall, the boxplot effectively summarizes the

distribution and variability across these metrics.

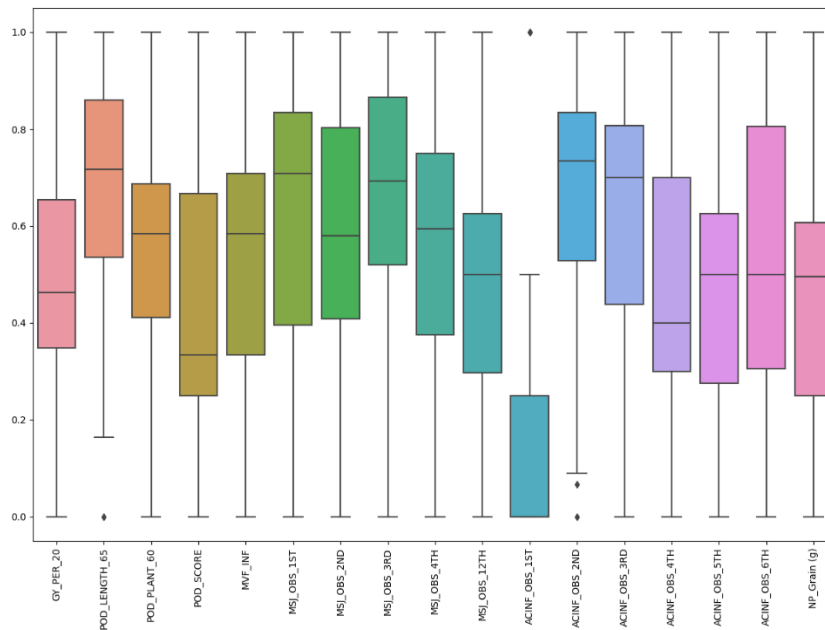


Figure 4: Relationship between crop yield variables

The histograms illustrated in Figure 5 shows the distribution of the crop-related metrics. Each subplot represents a specific metric, showing the frequency of data points across different ranges. The x-axis represents the metric values, while the y-axis indicates the count of observations within each bin. The distributions vary considerably among the metrics: some, like "POD_PLANT_60" and "ACINF_OBS_1ST," exhibit a skewed distribution with most values clustered

towards one end, while others, such as "GY_PER_20" and "MVF_INF," display a more uniform or bimodal distribution. The variation in shapes and spreads across the histograms suggests diverse patterns and characteristics in the data, reflecting different aspects of the crop's attributes and their measurements.

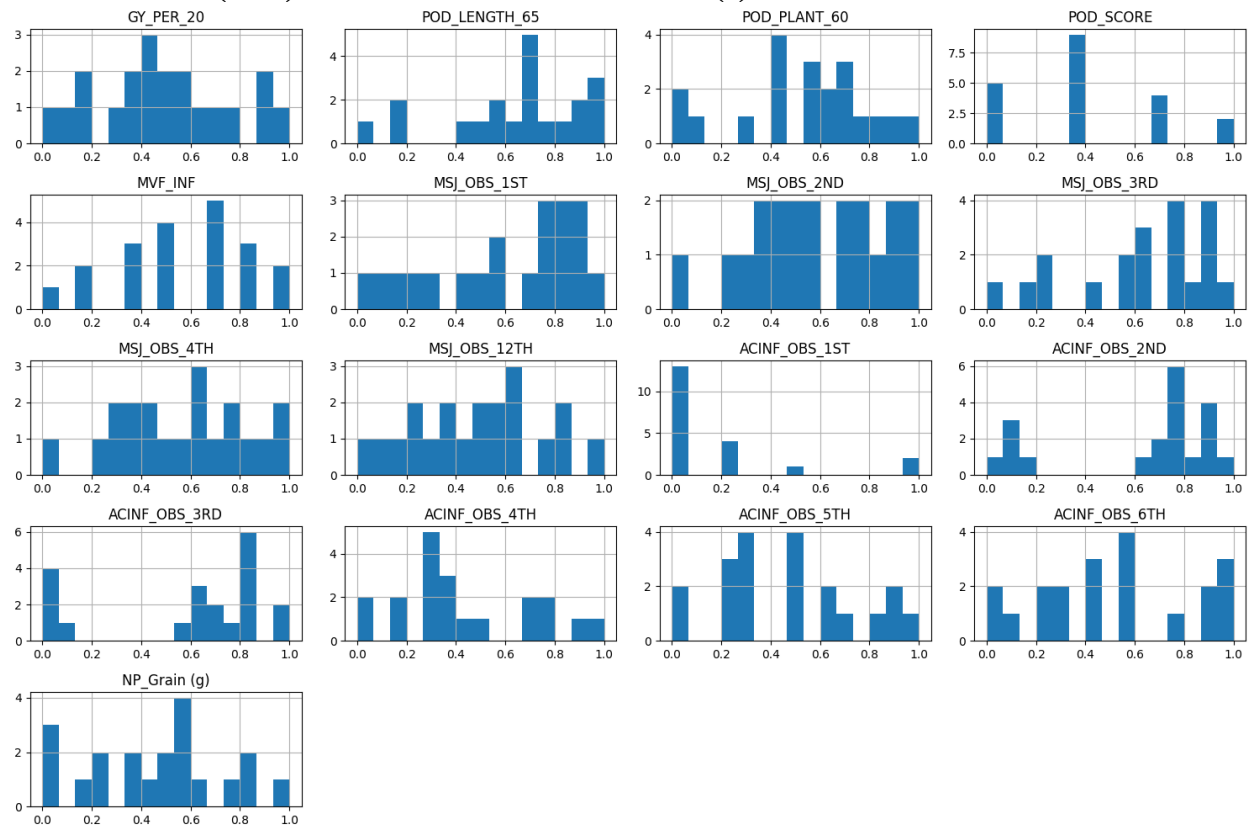


Figure. 5: Histogram plot of crop metrics

The boxplots in Figure 6 illustrates the influence of different treatments (A, B, C, D, and E) on various observations related to plant growth and pest infestation. For the yield-related metrics like GY_PER_20, POD_LENGTH_65, POD_PLANT_60, and MVF_INF, Treatment E consistently shows distinct outcomes, often with lower

variability and different median values compared to other treatments, indicating that Treatment E may have a unique impact on these parameters. Conversely, Treatments A, B, C, and D tend to show more overlap, particularly in their effects on pod length and plant metrics, suggesting similar influences on these aspects of plant growth.

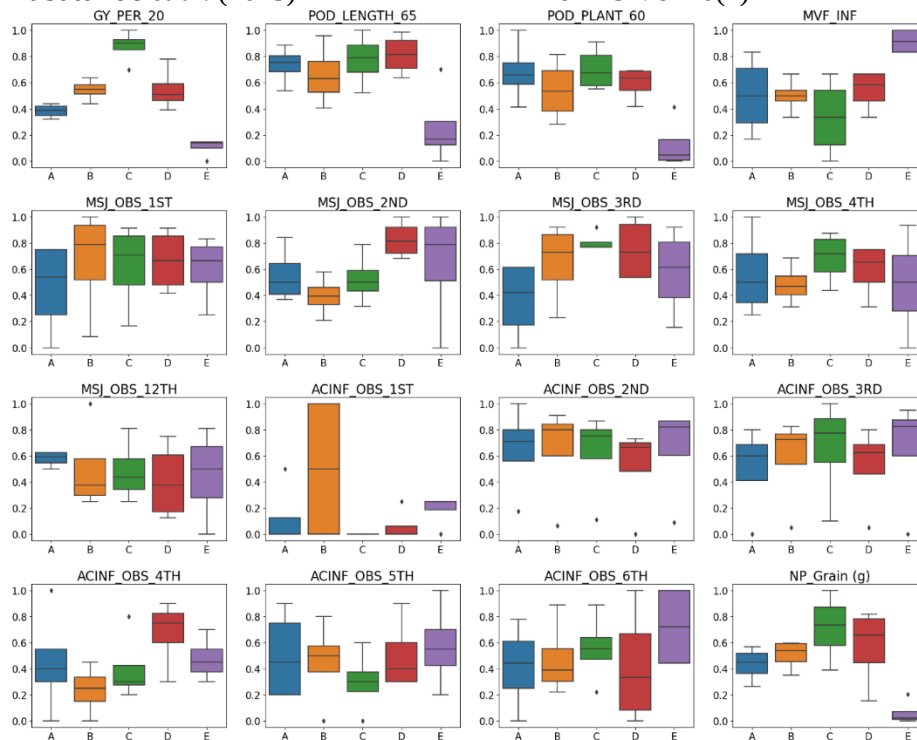


Figure 6: Influence of Treatments class on each observation

It can be seen that regarding pest infestation observations (MSJ_OBS and ACINF_OBS series), the variability across treatments is more pronounced. For example, the first infestation observations (ACINF_OBS_1ST) show stark differences, with Treatment B having a notably high median and little variability, while others like Treatment E show much lower infestation levels. Over time, as seen in later observations like ACINF_OBS_4TH to 6TH, the differences between treatments appear lesser, with most treatments converging towards similar infestation levels, though some treatments still show distinct patterns, indicating a varying degree of effectiveness in pest management across treatments.

4.2. Result of Analysis of Pest Infestation and Treatment Dataset

The progression of *Aphis craccivora* infestation over time across five treatments is illustrated in Figure 7a. Initially, all treatments show a similar increase in infestation, but over time, differences emerge. Treatments A, B, C, and D exhibit a steady rise in infestation levels, with Treatment E showing a slightly steeper increase initially, then gradually stabilize. The shaded regions indicate variability, with some overlap among treatments, but overall, infestation tends to rise across all treatments, with no clear outlier in terms of effectiveness in reducing infestation.

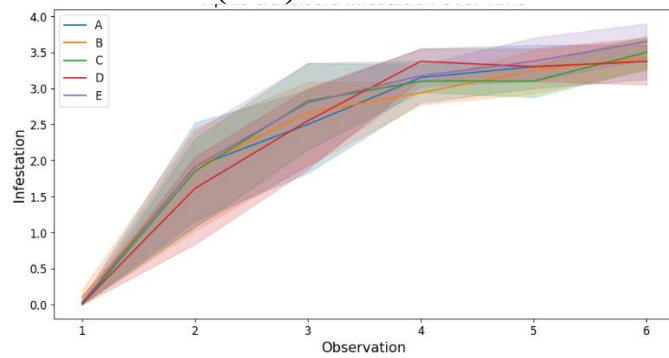


Figure 7a: *Aphis craccivora* Infestation observation over time based on treatment

Figure 7b presents *Megalurothrips sjostedti* infestation over a longer observation period. Unlike the first graph, this one shows more fluctuation in the early observations, particularly with Treatment D, which experiences a sharp initial increase followed by a decrease, then stabilization. Most treatments show a more stable trend over time, with slight increases or relatively flat

patterns after the initial fluctuation. The variability among treatments appears greater, especially in early observations, but over time, the differences between treatments diminish, indicating similar levels of infestation in the later stages of observation, which is similarly identified in Mayank *et al.* [8].

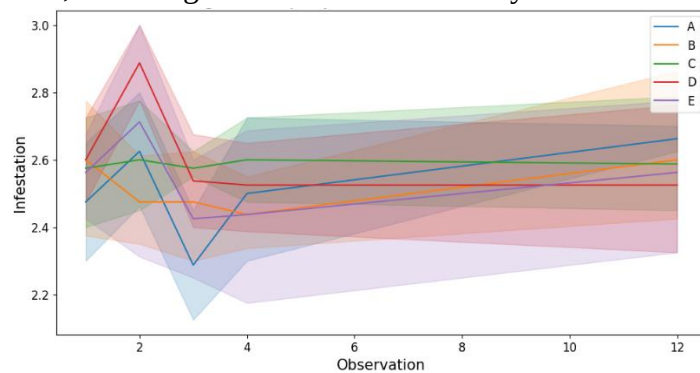


Figure 7b: *Megalurothrips sjostedti* Infestation observation over time based on treatment

5. Conclusion and Future Work

This research carried out an exploratory data analysis of cowpea yield and pest infestation dataset. The result of the experiment has provided valuable insights into the relationships between cowpea yield and presence of pest. The patterns and trends in the empirical result establishes that pest infestation significantly impacts crop yield, which calls for the need for effective pest management strategies. The analysis also

revealed seasonal variations in pest infestation, suggesting that timely interventions can mitigate yield losses. Furthermore, the correlation between environmental factors and pest activity underscores the importance of integrating ecological considerations into agricultural practices. This study largely emphasizes the critical role of data-driven decision-making in enhancing agricultural productivity and sustainability, paving the way for future research to explore advanced predictive

modeling and integrated pest management solutions. The future work in this direction will also incorporate additional datasets, with more input parameters which include soil types, nutrient contents, cowpea seed varieties and effect of application of bio-control agents on cowpea yield.

Declarations

Ethics Approval and Consent to Participate

Not Applicable

Consent for Publication

Not Applicable

Availability of Data and Materials

The Dataset used in this experiment was obtained from a researcher in the Department of Agriculture (Crop production) Kwara State University, Malete. Nigeria. Correspondence and requests for data should be addressed to R. S. Babatunde.

Competing Interests

The authors declare that they have no competing interests

Funding

This research is funded by the Federal Ministry of Communication Innovation and Digital Economy (FMCIDE) Nigeria, in collaboration with National Information Technology Development Agency (NITDA), in the 2023 intervention of NAIRS research grant.

Authors' Contributions

RSB: Conceptualization, Methodology, Data curation, Implementation, Visualization, Investigation, original draft preparation. AM: Conceptualization, Methodology, Supervision, Validation, Reviewing and Editing. JFA: Conceptualization, Supervision, Validation, Investigation. LOA: Investigation, Data curation. JO: Data Collection and curation. AMI: Guidance, Visualization, Supervision. AKM: Draft preparation and visualization.

Acknowledgments

Our appreciation goes to Dr. James Adebayo Ojo, an Entomologist of the Department of Agriculture and Crop Production, Kwara State University for his assistance towards the collection of Cowpea data for this research.

References

- [1] Morales, A. and Villalobos, F. J. (2023). Using machine learning for crop yield prediction in the past or the future. *Front. Plant Sci.* 14:1128388. doi: 10.3389/fpls.2023.1128388
- [2] Islam, M., Abdul, A., Talukder, A., Khabir, U., Ashraf, U., Kamran, H., Selina, S., Mahbubur, R. and Sumon, K. D. (2023). DeepCrop: Deep learning-based crop disease prediction with web application. *Journal of Agriculture and Food Research.* 14: 100764. <https://doi.org/10.1016/j.jafr.2023.100764>.
- [3] Khan, M. A., Ahmad, S. and Ali, A. (2022). "Machine Learning Approaches for Predicting Crop Yields: A Review." *Agricultural Systems.* 192: 103191.

- [4] Ojo, J. A., Adeyemo, A. J. and Olaniyan, O. J. (2023). "Predictive Modeling of Crop Yield Using Machine Learning Techniques: A Case Study of Cowpea." *Journal of Agricultural Science*. 15(1): 45-59.
- [5] Khan, M.A., Ahmad, S. and Ali, A. (2021). "Impact of Insect Pests on Cowpea Yield: A Review." *Journal of Agricultural Science*. 13(4): 45-56.
- [6] Ojo, J.A., Adeyemo, A. J. and Olaniyan, O.J. (2022). "Integrated Pest Management Strategies for Cowpea Production: Current Trends and Future Directions." *Agricultural Systems*. 192: 103200.
- [7] Shuaibu, N., Obunadike, G. N. and Jamilu, B. A. (2024). Crop yield prediction using selected machine learning algorithms. *FUDMA journal of Sciences*. 8(1):61-68. <https://doi.org/10.330>
- [8] Mayank, C., Darpan, C., Chaitanya, C. and Mansing, R. (2020). Crop Yield Prediction Using Machine Learning. *International Journal of Science and Research*. 9(4): 645-648. Doi <http://dx.doi.org/10.21275/SR20402185927>
- [9] Bera, A., Bhattacharjee, D. and Krejcar, O. (2024). PND-Net: plant nutrition deficiency and disease classification using graph convolutional network. *Sci Rep*. 14: 15537. <https://doi.org/10.1038/s41598-024-66543-7>
- [10] Nimbal, N. B. and Mareeswari, V. (2024). Machine Learning Techniques for Crop Yield Prediction: A Survey on Techniques, Applications, and Future Directions. *TWIST*. 19(4): 395-402. <https://twistjournal.net/twist/article/view/564>
- [11] Kavitha, S., Kumar, S. T., Naresh, E., Kalmani, V. H., Bamane, K. D., and Pareek, P. K. (2024). Medicinal Plant Identification in Real-Time Using Deep Learning Model. *SN COMPUT. SCI*. 5, 73. <https://doi.org/10.1007/s42979-023-02398-5>
- [12] Ahmed, A., Adewumi, S. E. and Yemi-peters, V. (2023). Crop Yield Prediction in Nigeria Using Machine Learning Techniques: (A Case Study of Southern Part of Nigeria). *UMYU Scientifica*, 2(4): 31 – 38. <https://doi.org/10.56919/usci.2324.004>
- [13] Sunkara, P. and Beulet, P. S. (2019). Heuristic Prediction of Crop Yield using Machine Learning Technique. *In International Journal of Engineering and Advanced Technology*. 9(1s3):135–138.
- [14] Karan, C. and Farhana, K. (2020). Prediction of Crop Yield Using Machine Learning. *International Journal of Engineering Applied Sciences and Technology*. 4(9):153-156. IJEAST (<http://www.ijeast.com>) doi:[10.33564/IJEAST.2020.v04i09.019](https://doi.org/10.33564/IJEAST.2020.v04i09.019)
- [15] Champaneri, M., Chachpara, D., Chandvidkar, C. and Rathod, M. (2020). Crop Yield Prediction Using Machine Learning. *International Journal of Science and Research*. 9(4): 645-648. Doi <http://dx.doi.org/10.21275/SR20402185927>
- [16] Newgard, C. D. and Lewis, R. J. (2015). Missing Data: How to Best Account for What Is Not Known. *JAMA*. 314(9):940-1. doi:10.1001/jama.2015.10516.

<https://fjpas.fuoye.edu.ng>