International Journal of Contemporary Computer Research (IJCCR), Vol.1 Issue.1 (June, 2026)
ISSN: 2600-9048

COMPARATIVE ANALYSIS AND BAYESIAN HYPERPARAMETER OPTIMIZATION OF MACHINE LEARNING MODELS FOR MAIZE YIELD PREDICTION USING A LARGE-SCALE SYNTHETIC DATASET

Sumaira Imtiaz[1]
Faculty of Computer and Information
Al-Madinah International University
Kuala Lumpur, Malaysia
CH226@lms.mediu.edu.my

Yazeed Al Moaiad[2]
Faculty of Computer and Information
Al-Madinah International University
Kuala Lumpur, Malaysia
yazeed.alsayed@mediu.edu.my

***Abstract*— Accurate yield prediction of crops is critical in enhancing crop planning and ensuring food security. This work describes a comparative assessment of regression-based machine learning models for maize yield prediction using a large-scale synthetic agricultural data with 166,824 maize data points from the structured data of a million observations. The data set contains agronomic and environmental variables such as rainfall, temperature, soil type, irrigation use, use of fertilizer, and the days to harvest.**

Four regression models were tested: Linear Regression, Random Forest, XGBoost, and Bayesian-optimized XGBoost. Performance was evaluated in terms of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R2). Linear Regression performed the best prediction (RMSE = 0.4993; R2 = 0.9143). Random Forest gave slightly lower results (R2= 0.9067), and the default XGBoost achieved competitive results (R^2= 0.9134). Bayesian optimization slightly improved the performance of the XGBoost (RMSE = 0.4999; R2 = 0.9141).

A paired t-test showed statistical significance ($p < 0.001$), but effect size analysis (Cohen's's d = -0.0226) showed a trivial practical difference. Feature importance analysis proved that the use of fertilizer, rainfall, and irrigation, together, explained most of the predictive power. The results imply that under structured dataset conditions, simpler linear models might be able to perform in the competition with complex ensemble approaches.

Keywords— maize yield prediction; machine learning; XGBoost; random forest; Bayesian optimization; regression analysis

I. INTRODUCTION

The agricultural productivity remains a key factor of world food systems, especially in areas where crop production has a direct impact on economic stability and food supply. Amongst cereal crops, maize (Zea mays L.) is one of the most commonly cultivated and economically vital crops. For maize production, which is affected by environmental unpredictability and management decisions, accurate prediction of yield has turned out to be an area of research that is important in agricultural analytics. Accurate predictions can help with irrigation management, fertilizer distribution, and other overall climate adaptation strategies [1].

Agricultural productivity is one of the important factors of the world food systems, particularly those where the production of crops has a direct bearing on the economic stability and the food availability. Amongst cereal crops, maize (Zea mays L.) is one of the most commonly grown and most economically important crops. For maize production, which is affected by environmental unpredictability and management decisions, accurate prediction of yield has turned out to be an area of research that is important in agricultural analytics. Accurate predictions can help in the irrigation management process and distribution of fertilizers and other general climate adaptation measures [2]. However, agricultural systems do not usually operate under perfectly linear conditions. Interactions between rainfall, soil characteristics, nutrient availability, and temperature commonly add to the complexity that may not be fully modeled in simple parametric models.

With developments in computational methods, Machine learning algorithms have been increasingly applied to agricultural data sets, with an attempt to more effectively model the nonlinear relationships [3]. Ensemble approaches like Random Forest [4], and boosting-based algorithms like XGBoost [5] can often be seen to beat traditional regression methods on regression (predictive) tasks. Additionally, Bayesian optimization has become a practical way to tune the

hyperparameters of models in a systematic and efficient way [6,7].

Notwithstanding all these developments, it is not always clear whether increased algorithmic complexity implies meaningful improvements under all conditions of data. In particular, when datasets have structured and proportionate relationships between predictors and target variables, simpler models may be able to compete. Moreover, although many studies have shown statistically significant improvements, fewer studies have considered whether these improvements are of practical significance.

For these reasons, the present study aims to make a critical reflection on model performance under structured synthetic agricultural conditions. Rather than assuming that advanced ensemble methods will necessarily be more effective than linear approaches, this research examines the relative behaviour of the two approaches while also taking into account how great the differences in performance are using effect size analysis.

A. Research Contributions

The primary contributions of this study can be summarized as follows:

1) ***Comparative evaluation of regression models:*** This study provides a systematic comparison of several regression-based machine learning models, including Linear Regression, Random Forest, XGBoost, and Bayesian-optimized XGBoost, for maize yield prediction.

2) ***Assessment under structured dataset conditions:*** Unlike many previous studies that rely on heterogeneous real-world datasets, this research evaluates model behavior using a large-scale synthetic dataset designed to represent structured agricultural relationships.

3) ***Integration of statistical and practical performance analysis:***
In addition to standard performance metrics such as RMSE, MAE, and R^2, this study incorporates statistical hypothesis testing and effect size analysis to distinguish between statistically significant and practically meaningful differences in model performance.

4) ***Interpretation of feature importance for agricultural factors:***
The study analyzes feature importance values to identify the most influential variables affecting maize yield prediction, highlighting the roles of fertilizer usage, rainfall, and irrigation.

5) ***Insights into model complexity versus dataset structure:***
The findings demonstrate that increased algorithmic complexity does not necessarily produce substantial predictive improvements when the underlying dataset exhibits predominantly linear relationships.

These contributions provide insights into the practical selection of predictive models for agricultural yield forecasting and emphasize the importance of aligning modeling strategies with dataset characteristics.

II. RELATED WORK

The prediction of crop yield has helped to engage the fields of agricultural informatics and applied machine learning to the greatest extent. The initial methods mainly used statistical modeling methods, including multiple linear regression and autoregressive. These techniques gave estimations of parameters that could be interpreted and were generally applied when any particular relationship between environmental factors and crop yield was assumed to be approximately linear. However, such models usually find it difficult to represent the complexity of interactions between climatic, soil, and management variables. As the computational power and access to bigger data sets grow, machine learning techniques have become noticeable in agricultural forecasting exercises. Random Forest (RF), which was proposed by Breiman [4], has found many adherents because it is resistant to overfitting and can be used to estimate nonlinear relationships. RF has proven to be more predictive and accurate compared to classical regression models in agricultural applications, especially in circumstances where heterogeneous environmental variables are concerned. It has an ensemble structure, which is built on bootstrap aggregation and random feature selection, thus allowing us to model the effects of interactions without having to specify them. Gradient boosting models have also improved the performance of predictive models. XGBoost is a framework suggested by Chen and Guestrin [5] and combines gradient boosting with regularization, effective tree building, and parallelization. These attributes have made it gain popularity in many predictive analytics applications, such as in agriculture. It has been reported in several studies that XGBoost performance can be improved when using crop yield data that has a high variance, and the predictors interact in a nonlinear manner.

Another important advancement in the machine learning processes is hyperparameter optimization. However, grid search or random search, which are more traditional methods, may be time-consuming or inefficient in a high-dimensional parameter space. Bergstra and Bengio [6] and Snoek et al. [7] describe Bayesian optimization methods as a way to explore the hyperparameter search space in a probabilistic manner, to do so efficiently. Through the performance modeling, Bayesian optimization is able to find promising configurations with a minimal number of iterations in contrast to exhaustive search strategies. Optimized boosting model supports better results than default settings in the context of agricultural prediction, but again, the extent of this improvement differs with the structural makeup of data.

Irrespective of these developments, some methodological issues are under-researched in the literature. One, most comparative studies focus on improvements in performance measures (statistically) without necessarily assessing practical significance. The large datasets may also deliver statistically significant differences to the small improvements in predictive models. The conclusion on model superiority can be exaggerated without the analysis of the complementary effects [8] [9].

Second, there is a relative lack of research that directly analyses the effects of the structural properties of datasets on the model performance. The agricultural data in the real world is usually characterised by spatial heterogeneity,

measurement noise, and time variation that can have an advantage for the non-linear ensemble methods. Conversely, the patterns of variance in structured or synthetic datasets, in which predictor response relationships are produced under controlled assumptions, can have a preponderantly linear character. In this case, the marginal cost of complex ensemble models can be small [10] [11].

Lastly, the interpretability and computational efficiency are also becoming important in agricultural decision-support systems. Ensemble methods can be highly predictive and can demand increased computational resources, but can be less interpretable than linear models. Some of these models might be simpler and be beneficial in real deployment scenarios, particularly when predictive performance differences are small [12] [13].

Considering these reasons, the current work paper can be added to the current collection of studies because it compares linear regression, Random Forest, and XGBoost models in a systematic way under the conditions of a structured synthetic agriculture. In contrast to most earlier reports, the given research involves statistical hypothesis testing as well as the analysis of the effect size to differentiate between the statistical detectability and actual relevance. In this way, it will contribute to a more detailed explanation of the situations in which a more complex model can make sense in tasks related to the prediction of crop yields [14].

III. Dataset Description

The data available in this research is a large-scale synthetic dataset of agriculture records of one million records, which are from various crop production environments. Based on this data, the observations that are specific to the maize farming process were generated to be analyzed. Upon filtering, 166,824 records of maize were saved to be used in modeling and evaluation.

The examples of the variables included in the dataset are environmental variables and agronomic management variables, which are often related to the crop yield variability. The environmental factors are rainfall and temperature, and the management-related variables are the farming activities (irrigation and use of fertilizers). Moreover, such categorical attributes as soil type and weather conditions would help in giving contextual information about the cultivation environment [12].

The predictor variables used in this study include:

- Rainfall (mm): The quantity of rainfall throughout the crop growth period. The rainfall is very important in maize productivity, as it affects the soil moisture and the uptake of nutrients.
- Temperature (o C): Conditions of temperature influence the rate of growth of the plant, the rate of photosynthesis, and the grain development. Extremely low and extreme high temperatures can have a detrimental effect on crop yield.
- Fertilizer Usage: A binary variable that shows the existence of fertilizer when cultivating. Fertilizer helps in providing nutrients, especially nitrogen, phosphorus, and potassium, that are needed in maize growth.
- Irrigation Usage: This will be a binary variable that will take the form of irrigation support. Irrigation complements the rainfall and ensures the release of moisture on the soil when there is dry weather.
- Soil Type: A nominal variable that describes the various soil compositions, like sandy soil, clay soil, loam, or silt soil. The structure of soil determines the ability to retain water, nutrient concentration, and root penetration.
- Weather Condition: This is a categorical variable that refers to the general climatic conditions, such as sunny, cloudy, or rainy, at the time of growing.
- Days to Harvest: The time taken between planting and harvest. The period of crop maturity can affect biomass and ultimate grain yield.

The target variable in the dataset is yield measured in tons per hectare, which represents the productivity of maize cultivation under the given environmental and management conditions.

Since the data is synthetic, the interactions among the variables were created with regulated assumptions to recreate natural agricultural conditions whilst preserving organized relations among the predictors. The synthetic datasets are commonly applied in the research of machine learning to test the behavior of models under a known structure of variance without the confounding factors of missing data, measurement errors, or incomplete data.

Before model training, the dataset was preprocessed, such as making a series of maize records, one-hot encoding continuous categorical variables, and binary numerical representations of the Boolean variables. Following the stage of preprocessing, the data set was split into an 80% training 80% and a testing 20% set so that the effectiveness of the models could be tested with unseen data.

This is due to the fact that the size of the dataset used is relatively large, which offers adequate statistical power to assess predictive models and enables a small change in the performance to be detected via hypothesis testing. Simultaneously, the organized character of the dataset allows studying how various regression models perform in the case when the relations between the predictors and responses are more systematic than extremely irregular.

IV. Methodology

A. Linear Regression Model

Linear regression assumes a linear relationship between predictors and response:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon$$

where y is yield, x_i are predictor variables, β_i are coefficients, and ϵrepresents error.

B. Random Forest

Random Forest constructs multiple decision trees using bootstrap sampling and feature randomness. The final prediction is obtained by averaging the outputs of individual trees.

C. XGBoost

XGBoost is a gradient boosting algorithm that sequentially minimizes an objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k)$$

where l represents loss and Ω regularization term controlling model complexity.

D. Evaluation Metrics

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

E. Statistical Analysis

A paired t-test was performed on absolute prediction errors:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

Cohen's d effect size was calculated as:

$$d = \frac{\bar{d}}{s_d}$$

V. Experimental Setup

To evaluate the performance of the regression models, a structured experimental framework was implemented using Python-based machine learning libraries. The experiments were conducted using the **Scikit-learn** library for Linear Regression and Random Forest models, while the **XGBoost** library was used for gradient boosting implementation. All experiments were executed within a Python environment to ensure reproducibility and consistent parameter configuration.

A. Data Splitting Strategy

The dataset was divided into training and testing subsets using an 80:20 ratio. The training subset was used to train the models and learn the underlying relationships between predictor variables and maize yield, while the testing subset was reserved for evaluating predictive performance on unseen data.

This approach ensures that model evaluation reflects generalization capability rather than memorization of training data. A fixed random seed was used during the data splitting process to ensure consistent experimental results.

B. Feature Encoding and Transformation

Several preprocessing operations were performed before model training. Categorical variables such as soil type and weather condition were transformed using one-hot encoding, converting each categorical level into a binary indicator variable. This transformation allows machine learning algorithms to process categorical attributes without introducing ordinal bias.

Binary variables representing fertilizer usage and irrigation usage were converted into numerical values (0 and 1) to ensure compatibility with regression algorithms. Continuous variables such as rainfall and temperature were retained in their original numeric form.

C. Model Training Configuration

Four regression models were trained and evaluated in this study:

- Linear Regression
- Random Forest Regressor
- XGBoost Regressor
- Bayesian-optimized XGBoost

The Linear Regression model was implemented as a baseline approach due to its simplicity and interpretability.

The Random Forest model used multiple decision trees constructed through bootstrap sampling. Predictions were obtained by averaging the outputs of all individual trees, which helps reduce variance and mitigate overfitting.

The XGBoost model employed gradient boosting techniques, where each new tree is trained to minimize the prediction errors of the previous ensemble.

D. Hyperparameter Optimization

To improve the predictive performance of the boosting model, Bayesian optimization was applied to search for optimal hyperparameter configurations. The optimization process was conducted over 50 trials, allowing the algorithm to iteratively evaluate candidate parameter combinations.

The main hyperparameters considered during optimization included:

- Number of estimators
- Maximum tree depth
- Learning rate
- Subsample ratio
- Column sampling ratio

Bayesian optimization constructs a probabilistic model of the objective function and identifies promising hyperparameter regions through sequential exploration.

E. Evaluation Procedure

Model performance was evaluated using three commonly used regression metrics:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2)

These metrics provide complementary perspectives on predictive performance. RMSE emphasizes larger prediction errors, MAE provides a direct measure of average deviation, and R^2 quantifies the proportion of variance explained by the model.

To further examine differences between model predictions, a paired t-test was conducted on absolute prediction errors. In addition, Cohen's d effect size was computed to assess the magnitude of performance differences beyond statistical significance.

VI. RESULTS

A. Model Performance Comparison

The predictive performance of the evaluated regression models is summarized in **Table 1**. Three evaluation metrics were used to assess model accuracy: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). These metrics provide complementary perspectives on model performance, allowing both error magnitude and explained variance to be examined.

Table 1 Predictive performance comparison of regression models

Model	RMSE	MAE	R^2
Linear Regression	0.4993	0.3980	0.9143
Random Forest	0.5211	0.4161	0.9067
XGBoost	0.5021	0.4006	0.9134
Optimized XGBoost	0.4999	0.3984	0.9141

The results show that **Linear Regression achieved the highest coefficient of determination (R^2 = 0.9143)** and the lowest RMSE among the evaluated models. This indicates that the linear model was able to explain more than 91% of the variance in maize yield within the dataset.

Random Forest produced slightly lower predictive accuracy with an R^2 value of 0.9067, suggesting that ensemble tree structures did not provide a clear advantage under the structured dataset conditions used in this study.

The default XGBoost model achieved competitive performance (R^2 = 0.9134), demonstrating the effectiveness of gradient boosting in modeling complex relationships between environmental variables and crop yield. After Bayesian hyperparameter optimization, XGBoost achieved a marginal improvement (R^2 = 0.9141), approaching the performance of the linear baseline model.

Overall, the differences between the top-performing models were relatively small, indicating that the dataset may contain predominantly linear relationships between predictor variables and the target yield variable.

B. Actual vs. Predicted Yield Comparison

To visually evaluate model predictions, the relationship between actual yield values and predicted outputs was examined **Figure 1** presents the scatter distribution of actual versus predicted maize yields for the Linear Regression and optimized XGBoost models.

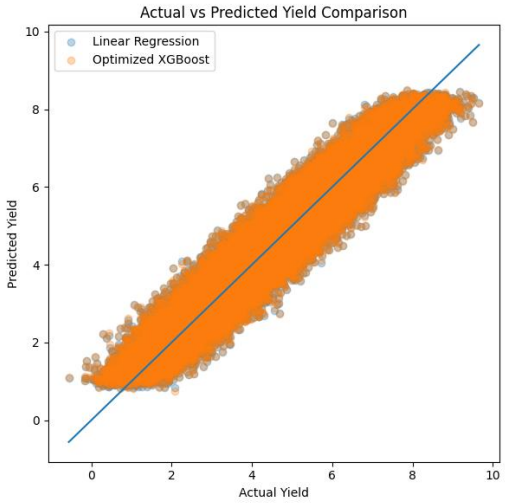

Figure 1 Actual versus predicted maize yield for Linear Regression and optimized XGBoost.

The data points in Fig. 1 are distributed closely around the diagonal reference line, indicating strong agreement between predicted and actual yield values. Both models exhibit similar prediction patterns, further confirming that their overall predictive performance is nearly equivalent.

The absence of systematic deviation from the diagonal line suggests that both models provide unbiased predictions across the yield range.

C. Residual Distribution Analysis

Residual analysis provides additional insight into the quality of model predictions Figure 2 illustrates the distribution of residuals for the Linear Regression and optimized XGBoost models.

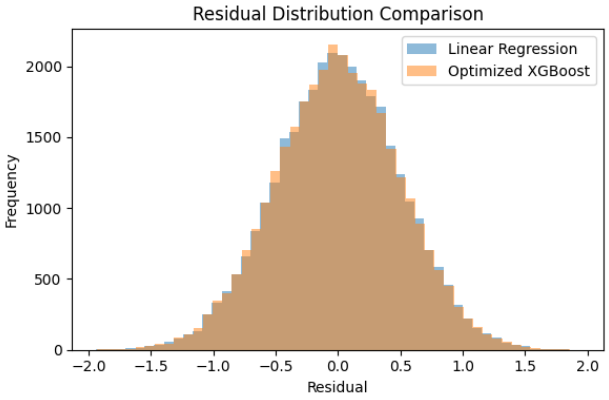

Figure 2 Residual distribution comparison between Linear Regression and optimized XGBoost.

The residual values are concentrated around zero for both models, indicating that prediction errors are relatively balanced and do not exhibit strong systematic bias. A symmetrical residual distribution suggests that the models do not consistently overestimate or underestimate maize yield.

The similarity between residual distributions further supports the observation that both models perform comparably under the current dataset conditions.

D. Feature Importance Analysis

Understanding which variables contribute most to yield prediction is important for interpreting model behavior.

Feature importance values derived from the optimized XGBoost model are presented in Figure 3.

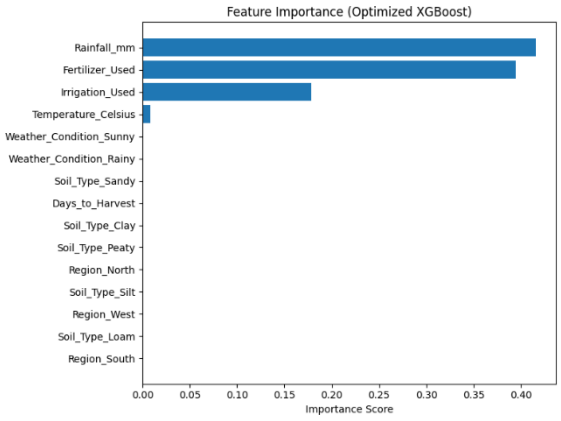

Figure 3 Feature importance ranking derived from the optimized XGBoost model.

The analysis shows that **fertilizer usage (49.9%)**, **rainfall (29.9%)**, and **irrigation usage (19.4%)** account for the majority of predictive influence. These variables collectively explain nearly all of the model's predictive capacity.

Other variables, including soil type, temperature, and weather conditions, contributed comparatively smaller importance values. This distribution aligns with agronomic knowledge that nutrient availability and water supply are primary determinants of crop productivity.

E. Statistical Significance Testing

To evaluate whether performance differences between models were statistically meaningful, a paired t-test was conducted on the absolute prediction errors of Linear Regression and optimized XGBoost.

The test produced a **t-statistic of −4.128 with a p-value below 0.001**, indicating statistical significance at conventional confidence levels. However, statistical significance alone does not necessarily imply practical improvement.

To assess practical relevance, **Cohen's d effect size** was calculated. The resulting value of **d = −0.0226** indicates an extremely small effect size, suggesting that the difference in prediction error between the two models is negligible in practical terms.

This finding highlights the importance of interpreting statistical significance alongside effect size, particularly when large datasets can make small performance differences statistically detectable.

VII. Discussion

The findings achieved within the current study allow for multiple conclusions regarding the correlation between the structure of the datasets and the performance of the machine learning models in predicting the crop yields. Even though random forest and XGBoost can be considered as some of the most efficient ensemble algorithms, the results show that they are not similarly effective in all the circumstances in which their usage may be considered. In the study, the conventional Linear Regression model performed as well as or better than complicated ensemble models.

The structural properties of the data set employed in the analysis can be one explanation for such a result. The agricultural data used in this research were artificially created under a controlled environment, and it is likely that the relationships between predictor variables and the target yield variable were rather systematic. Linear regression can be used to model predictor-response relationships that are, to a large proportion, linear or proportionate, as most of the explainable variance generally is well-modelled without elaborate nonlinear transformations.

Conversely, ensemble models like the Random Forest and the XGBoost are mostly useful when data sets exhibit intense nonlinearities, skewed distributions, or excessive noise. In this case, decision-tree-based techniques can model interactions between variables with intricate dependencies that can be hard to measure with other traditional regression techniques. Nonetheless, in cases where the underlying data structure is reasonably structured and foreseeable, it is not possible that the extra complexity of ensemble models can provide a significant predictive advantage.

The other notable point is related to the usage of statistical significance interpretation in large data sets. The paired t-test carried out in this analysis obtained a statistically significant difference between Linear Regression and optimized XGBoost predictions. Nevertheless, the obtained Cohen's d effect size was extremely insignificant (d = -0.0226), which means that the level of the difference in the performance is virtually insignificant. This demonstrates a classic problem of statistical analysis, that when samples are large, even minor variations in the error of prediction can be statistically significant.

Practically, the distinction between the two models might be insignificant in real-life agricultural decision-making. Even in practice-relevant settings (crop management or yield prediction), such a small difference in predictive accuracy is not likely to affect the operations decision dramatically. Thus, assessing the effect size and statistical significance can give a more moderate consideration of model performance.

The feature importance analysis is also useful in providing information about the factors that influence the prediction of maize yield in the dataset. The most significant predictors were found to be fertilizer usage, rainfall, and irrigation. These results are consistent with agronomic information, where the supply of water and nutrients is also a well-established factor affecting crop yield. These variables dominate, hence implying that such variation of yields in the dataset is largely attributed to a few key environmental and management factors.

Meanwhile, the comparatively minor role of the other factors, i.e., soil type and weather conditions, suggests that the simulated dataset may not be as representative of the complexity of the real agricultural system as it is. In the natural setting, the yield results tend to be affected by other factors such as pest pressure, variability of nutrition in the soil, climatic changes, and management. Such considerations can add non-linear relationships that can augment the relative advantage of developed machine learning models.

The other feature that should be considered is the trade-off between predictive accuracy, interpretability, and computational efficiency. The interpretability of Linear Regression models is very explicit since the effect of an individual predictor variable may be directly investigated

using regression coefficients. Conversely, the ensemble-based models like the Random Forest and the XGBoost tend to operate as black-box models, and thus, it is harder to explain how specific features are used to make predictions.

Computationally, simpler models are also beneficial. Linear Regression does not need as many training and computation resources as ensemble algorithms, which demand the creation and optimization of various decision trees. Linearity and transparency of linear means of performance may be desirable where the difference between models is minimal.

Even with the enlightenment given by this paper, there are a few limitations that must be noted. To start with, the data employed in the analysis is synthetic, and it implies that the connection among variables has been created under controlled assumptions and not observed in real agricultural systems. Synthetic datasets enable more controlled experimentation and large sample sizes, but could not be very useful in terms of the variability and uncertainty found in the actual crop production settings.

Second, it was only regression-based machine learning that was analyzed. As more complicated agricultural data sets are considered, other more modern methods, such as deep learning models and hybrid ensemble frameworks, could also provide more predictive power. Such models may be studied in future studies in conjunction with spatial/temporal agricultural data.

Lastly, further environmental variables such as soil moisture measurements, indicators of climate variability or remote sensing features could be integrated into future research. The combination of these data sources can offer a better idea of the factors that affect the crop yield and can point to the situations when complex machine learning models can more significantly benefit.

On the whole, the findings of this paper underline that the choice of models cannot be made only based on the sophistication of the algorithms but also based on the characteristics of the datasets. The structure and variability of the data is the key to understanding the appropriate predictive modeling approach.

VIII. Conclusion

In this research, a comparative analysis of some of the regression-based machine learning models of maize yield prediction on a large-scale synthetic agricultural dataset were presented. The models that were explored were Linear Regression, Random Forest, XGBoost and Bayesian-optimized XGBoost. The goal of the analysis was to find out whether more complex models based on ensembles have relevant predictive benefits when fit using structured agricultural data.

In the experiment, it was shown that the performance of the Linear Regression model in terms of its predictive performance was similar to more complex ensemble methods. Specifically, the best coefficient of determination was obtained with Linear Regression ($R^2 = 0.9143$) and almost the same result was obtained with the optimized XGBoost model ($R^2 = 0.9141$). The pair statistical test showed that there were significant differences in the model predictions, but the effect size calculated was very small. The above finding underscores the need to consider both statistical significance and practical significance in predictive model evaluation.

Another point brought out by the results is the impact of dataset structure on model performance. Since the dataset applied in this research has a relatively systematic association between the predictor variables and yield results, a simple regression model was able to explain a majority of the explanatory variance. Conversely, the ensemble methods, including Random Forest and XGBoost tend to offer more benefits when underlying relationships between variables are highly nonlinear or where the data have much noise.

The analysis of important features further indicated that the use of fertilizers, rainfall, and irrigation ranked as the most potent predictors of maize yield. These findings can be correlated with much-established agronomic information on the significance of water supply and nutrient regulation on crop yield. The existence of such dominance in these variables implies that the variation in the yields of the dataset can be attributed mostly to a few major environmental and managerial factors.

The other impactful implication of this work is connected to the selection of a model in practice in agriculture. Although an advanced machine learning model is supposed to be more effective than a traditional statistical model, the results have shown that the increase in the complexity of the algorithm does not necessarily result in a significant improvement in the predictive performance. Even in cases where datasets are characterized by structured or largely linear relationships, more complex models might not add value in terms of predictive accuracy, but also have the benefit of being easier to interpret and compute.

Though this research has contributed to the body of knowledge, there are various limitations that ought to be noted. The data to be used in the analysis is artificial and was created under regulated assumptions. Even though this method allows large-scale experimentation and comparative controls between models, it might not fully reflect the variability and complexity of the actual agricultural environment. Other nonlinearities that may affect the performance of the models include climate variability, soil heterogeneity, pest pressure, and management practices.

The present research can be further developed by future research by using the same comparative modeling framework on actual agricultural data derived by field experiments or remote sensing observations. The integration of spatial and time-based factors, e.g. satellite images, soil moisture sensor, or weather indicators, might give more useful information concerning crop yield dynamics. Besides, it is possible to consider sophisticated machine learning approaches including deep learning architecture or ensemble-based approaches, which may be more effective in enhancing predictive outcomes in more complicated agricultural settings.

To conclude, the results of this paper prove that a trade-off between the complexity of the model and the specifics of the analyzed dataset must be weighed with great care. Although ensemble machine learning model is still a potent predictive analytics tool, its benefits might be minimal in the application to structured datasets that are dominated by linear relationships. The nature of the data is thus important in

determining the most suitable modeling technique of agricultural yield prediction.

IX. REFERENCES

[1] D. B. Lobell and M. B. Burke, "On the use of statistical models to predict crop yield responses to climate change," *Agricultural and Forest Meteorology*, vol. 150, no. 11, pp. 1443–1452, 2010.

[2] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ, USA: Wiley, 2012.

[3] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, and S. H. Kim, "Random forests for global and regional crop yield predictions," *PLoS ONE*, vol. 11, no. 6, 2016.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[6] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[7] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 2951–2959.

[8] Q. Feng, J. Liu, and J. Gong, "Urban flood mapping based on remote sensing and XGBoost," *Remote Sensing*, vol. 11, no. 9, 2019.

[9] D. K. Ray, N. D. Mueller, P. C. West, and J. A. Foley, "Yield trends are insufficient to double global crop production by 2050," *PLoS ONE*, vol. 8, no. 6, 2013.

[10] Al Moaiad, Y., D., Alkhateeb, M., & Alokla, M. (2024). Enhancing Cybersecurity Practices in Nigerian Government Institutions an Analysis and Framework. J. Electrical Systems, 20(3), 5964-5967.

[11] Al Moaiad, Y., D., Alkhateeb, M., & Alokla, M. (2024). Deep Analysis of Iris Print and Fingerprint for Detecting Drug Addicts. J. Electrical Systems, 20(3), 5639-5644.

[12] Al Moaiad, Y., Alobed, M., & Alsakhnini, M. (2024). Python Solutions to Address Natural Language Challenges. International Journal, 10(3), 3594-3603.

[13] Al Moaiad, Y., Alobed, M., Alsakhnini, M., & Momani, A. M. (2024). Challenges in natural Arabic language processing. Edelweiss Applied Science and Technology, 8(6), 4700-4705.

[14] Al Moaiad, Y., Alobed, M., Alsakhnini, M., & Momani, A. M. (2024). Challenges in natural Arabic language processing. Edelweiss Applied Science and Technology, 8(6), 4700-4705.