© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

Development of Artificial Intelligence Based Model to Predict Crop Yields of Biofertilizers

ANAS MUSA SHEHU[1], SURAJUDEEN ABDULSALAM[2], MURTALA AMINU BABA[3]
[1, 2, 3]*Center for Embedded Artificial Intelligence and Smart Energy Systems, Abubakar Tafawa Balewa University, Bauchi, Nigeria.*
** Corresponding author: Anas Musa Shehu*

Abstract*— The research paper proposes the development and the analysis of an artificial intelligence model that can be used for predicting the crop yield under various biofertilizer treatment, types of soils and environmental conditions. The problem statement is the necessity of the prediction of the efficiency of biofertilizers taking into account the interaction of microbial strains, soil properties, inoculation rate, nutrients, crops, and climatic zones. In this research, there is a use of the dataset of 5,500 observations, including 500 real observations obtained during the published materials with biofertilizers and 5,000 artificial observations created by conditional tabular generative adversarial network (CTGAN) in python. The dataset contains data about maize and mixed crops, including such features as soil pH, nitrogen, phosphorus, potassium, soil moisture, biofertilizer type, inoculation dose, colony-forming units of microbes, soil type, climatic zone, and crop yield. There was used data pre-processing, including cleaning, encoding, scaling, and splitting the dataset into the training, validation, and test sets. In this research, such algorithms as linear regression, random forest, gradient boosting, and XG-boost were applied. The choice of models was based on the metrics R^2, RMSE, MAE, MAPE, and Nash-Sutcliffe Efficiency (NSE). The optimal algorithm for the maize dataset was XGBoost with the following parameters: $R^2 = 0.9479$, RMSE = 300.06 kg/ha, and MAE = 225.85 kg/ha. The most accurate prediction of the crop yield for the dataset of mixed crops was the result of the Random Forest algorithm: $R^2 = 0.9990$, RMSE = 234.91 kg/ha, and MAE = 113.73 kg/ha. The findings suggest that the interaction of Azotobacter and Azospirillum led to an increase of 28.7% in maize production in comparison with the control group. Variables that have the greatest influence on the process were found to be biofertilizer application, soil nitrogen, inoculation level, soil, and climate zone. Further, the best-performing model underwent optimization through feature elimination, quantization, and reduced precision, making it embeddable on the device. The optimization resulted in R^2 value equal to 0.9108, or about 96% of the original model's goodness-of-fit measure. It can be seen that tree-based machine learning models can learn intricate relationships between biofertilizer application and yield. However, more testing with independent data is needed before the use of such models for practice.

Keywords*—Biofertilizers, crop yield prediction, machine learning, XGBoost, Random Forest, explainable AI, embedded AI, sustainable agriculture.

I. INTRODUCTION

Globally, agriculture plays an essential role in supporting the sustenance of life in rural areas and food security, which is particularly evident in developing countries like Nigeria. Apart from factors such as climate change, environmental degradation, and decreasing availability of arable land, the industry faces increasing challenges with regard to the rising population estimated to be 9.7 billion people by 2050 (FAO, 2023). Sustainable intensification is critical for meeting the growing demands of the future without further exploitation of natural resources.

The Green Revolution from 1960 to 2020 increased wheat yields by 50 to 100%, mostly fuelled by chemical fertilizers. But it came with a hidden cost. Soil acidification, nutrient imbalance, groundwater contamination, and nitrous oxide emissions have all been connected to extended and frequently excessive fertilizer use (FAO, 2023). A significant portion of sprayed nitrogen is never absorbed by the crop and is instead lost to the environment in many developing locations due to low nitrogen-use efficiency (Ajibade *et al.,* 2023). This is the type of issue that the current study addresses: farmers have little assurance about how a biofertilizer will function on their own soil because current fertilizer management techniques are not well matched to the biological, slow-release behaviour of microbial inputs. The most popular methods for predicting yield response to chemical fertilizer regimes are still process-based simulation platforms like the Agricultural Production Systems Simulator (APSIM) and the Decision Support System for Agrotechnology Transfer (DSSAT). By simulating nutrient dynamics under various

management scenarios, both achieve respectable accuracy (R^2 of about 0.70–0.92) as shown in table 1.

Table 1: Comparison between the Existing Approaches.

Model Type	Core Approach	Handles Fertilizer Explicitly?	Typical Accuracy (R^2)	Scalability to Real-Time/Embedded	Main Limitation
1) DSSAT	Mechanistic, daily	Yes (application events)	0.75–0.92	Low	Requires extensive calibration
2) APSIM	Modular, process-based	Yes (management rules)	0.70–0.90	Low	Complex module interactions

However, they require extensive site-specific calibration and were not intended to represent the microbial processes that control biofertilizer performance (Kashyap et al., 2024). Eutrophication, nitrate poisoning of water bodies, and a progressive deterioration in soil health have all been linked to continued reliance on chemical fertilizers (Samantaray *et al*., 2024; Zhang *et al*., 2024). Biofertilizer formulations developed around living microorganisms such as Rhizobium (nitrogen fixers), bacillus and pseudomonas species (phosphate solubilizers), potassium mobilizers, and helpful fungi offer a substantially greener pathway (Bhattacharyya & Jha, 2022). Bhattacharyya and Jha (2022) indicate that, by boosting nutrient cycling, root development, hormone signalling, stress tolerance, and microbial diversity in the soil (Singh *et al.,* 2023), biodegradable microbial inputs might minimize dependence on synthetic fertilizers. But they have not been readily accepted by commercial farmers, mainly because of their inconsistent performance, which depends on the particular strain of micro-organism, the carrier, the soil, the crop, the climate and the management. On the production side, the nonlinear biology of strain selection, viability, contamination control, stability, and scale-up means that trial-and-error approaches struggle to keep up (Malusá *et al*., 2022; Kumar *et al*., 2025). While variability in field conditions makes it difficult to design protocols that work everywhere.

This is the reason why artificial intelligence based approach has been chosen for the present study. Artificial intelligence, including machine learning, deep learning, predictive analytics and soft sensing, provides an approach to understanding exactly this type of complexity by identifying patterns in noisy high-dimensional data (Kamilaris & Prenafeta-Boldú, 2018). Machine learning has already been used in the production of biofertilizer to optimize fermentation parameters including temperature, pH and aeration for maximizing microbial biomass with minimum energy consumption (Butean *et al*., 2025). Similarly, Vessey (2023) states that plant growth-promoting rhizobacteria respond to soil and management conditions in ways that are difficult to capture with fixed, rule-based models but are well suited to data-driven learning. Biofertilizers as a group do show genuine promise as a more sustainable complement or alternative to chemical fertilizers, but inconsistency of production, formulation stability and field level variability remain limitations to their impact. Artificial intelligence, applied throughout the value chain of biofertilizers, offers a realistic way to achieve the precision required to convert this promise into dependable gains in food security (Gunasekaran & Sreevardhan, 2025). To achieve this, the present study builds supervised machine learning regression models trained on combined real and synthetically augmented biofertilizer trial data to predict crop yield under different biofertilizer treatments, and then compresses the best-performing model so that it can eventually run on low-power embedded hardware in the field. The particular process employed to do this is spelled out in the introduction will be discussed thoroughly in three (3).

Despite rising acceptance of biofertilizers as a sustainable complement to chemical fertilizers, uptake remains constrained by persistent technical and practical constraints. Artificial intelligence and machine learning have shown clear potential for

optimizing complex biological and agricultural systems, yet few empirical studies apply these techniques across the full biofertilizer value chain from fermentation and formulation design through to field application and yield prediction (Sharma *et al*., 2023; Butean *et al*., 2025). In the absence of effective, data-enabled decision-making tools for biofertilizer specific, several innovative techniques associated with biofertilizers have remained restricted to lab-scale studies and have not delivered reliable and scalable gains to farmers from the present situation where biofertilizers are uncertain and adopted via hit-or-miss efforts to the future scenario where reliable predictions for farmers will be achieved. Additionally, the production process of biofertilizers is quite prone to various parameters including pH, temperature, aeration, and nutrient content. All these factors contribute towards microbial variability among different batches (Malusá *et al*., 2022; Kumar *et al.*, 2025). The farmer's hesitation in using biofertilizers is due to the fact that there are models to predict crop yield through the use of chemical fertilizers which have been designed without considering the biological fertilizer behaviour. Biological fertilizer works on the principle of gradual nutrient release by microbial activity and hence it is difficult to predict its return from the crops.

II. RELATED WORKS

Increasing empirical research is linking the use of biofertilizers to benefits in crop production, produce quality and soil quality in a range of Agro-ecosystems. A meta-analysis of 1,818 observations from 107 field experiments in China found that biofertilizers increased the yields of 21 of 23 crops studied, including millet and various legumes, which had increases of over 50% compared to untreated controls (Pei *et al*., 2025). These yield benefits were coupled with improvements in nutritional quality, higher protein and vitamin C content and lower nitrate levels, showing that biofertilizers can sustain both productivity and produce quality at the same time (Pei *et al*., 2025). The mechanisms for these benefits relate to changes in soil health and the nitrogen cycle. The same meta-analysis found that biofertilizer application raised soil organic matter and the activity of beneficial enzymes such as urease and phosphatase, encouraged beneficial bacterial and fungal populations, and suppressed pathogenic organisms, all of which improve nutrient availability and root development (Pei *et al*., 2025). Increases in total soil nitrogen and accessible phosphorus appeared as particularly strong indicators of yield increase, indicating to biofertilizers' capacity to maximize nutrient cycling under field conditions. On the modelling side, work by Kumar *et al*., (2023) and Botero-Valencia *et al*., (2025) suggests that AI can improve nutrient scheduling, soil monitoring, and decision support, albeit these applications have largely addressed inorganic fertilizer management rather than microbial inputs. Studies of soil microbiomes similarly highlight AI's potential for evaluating microbial interactions and influencing soil health models, but seldom tie these discoveries back to industrial biofertilizer manufacturing or field deployment. More broadly, machine learning has transformed crop yield forecasting, pushing it from basic statistical regression toward data-driven ensembles that include soil, meteorological, and remote-sensing data. In a systematic review of 50 studies, Van Klompenburg *et al.,* (2020) identified artificial neural networks as the most frequently used algorithm, typically built on temperature, rainfall, and soil-type features, achieving R^2 values up to 0.92, while also flagging a continuing need for hybrid modelling approaches. These findings are supported by field-level studies. For example, studies of maize–soybean intercropping systems revealed that mycorrhizal biofertilizers combined with organic amendments markedly enhanced nitrogen availability, soil health and yield in marginal land (Xue *et al*., 2025), demonstrating biofertilizer performance can be optimized when microbial inputs are integrated with complementary approaches adapted to local soil conditions.

Apart from fermentation optimization, AI and data-driven approaches are widely used in R&D of biofertilizers to address systemic challenges including strain selection, formulation development, environmental adaptability, and performance prediction. Machine learning methods may combine different inputs, soil properties, weather, and plant growth responses, as well as microbial genetic information to surface patterns and decision rules that drive biofertilizer development (Kumar *et al*. 2025). Supervised learning methods have been applied to classify microbial strains based on functional attribute and compatibility with various soils and crops, allowing for more tailored formulations with a higher chance of success in the field. Furthermore, AI-aided design frameworks have improved

formulation stability as well. Methods of optimization such as genetic algorithms and Bayesian optimization have been used to find carrier compositions that optimize microbial survival during storage and retain activity at the point of application (Patel *et al*., 2025), reducing the need for empirical trial-and-error and decreasing development times. AI models are being used in the field to predict the performance of biofertilizers under varying climatic conditions. With historical yield and soil data, these models can predict how a crop will respond to a particular biofertilizer under a given climate, allowing agronomists to develop site-specific treatment protocols. Evidence suggests that AI-augmented decision support systems can outperform traditional models on yield prediction and nutrient recommendation accuracy (Jha & Singh, 2024; Liu *et al*., 2025), reinforcing the broader case that AI and other data-driven methods can accelerate discovery, sharpen predictive capacity, and support the practical deployment of biofertilizers in agriculture.

To enhance the accuracy of large-scale rice yield forecasting in China, Cao *et al*. (2021) proposed a data-fusion approach instead of single-source predictors. The main objective was to evaluate if a stronger yield model at the national scale could be developed by integrating multiple independent data streams, such as satellite-derived vegetation indices, in-situ meteorological data and maps of soil properties, compared to approaches based on a single source. Methodologically, the authors collected a multi-source dataset from different rice growing regions across China, and trained a set of machine learning and deep learning models on the combined feature set, and compared their performance with models trained on individual data sources in isolation. Results showed that the fusion of satellite, weather and soil information consistently improved prediction accuracy compared to single source baselines. This confirmed that rice yield is shaped by the interaction of climatic, edaphic and vegetative factors, rather than a single dominant driver.

This finding reinforces the broader argument, central to the present paper, that multivariate, multi-source modelling outperforms narrower approaches when the underlying agronomic system is complex. However, the study's reliance on satellite and national meteorological infrastructure means its approach is difficult to replicate in data-sparse smallholder contexts, and, more importantly for this paper, it did not incorporate any biofertilizer or microbial treatment variables, leaving open the question of how such data-fusion techniques would perform when biological inputs are added to the feature set. Fashoto *et al*., (2021) aimed to demonstrate the practical value of machine learning over conventional statistical estimation for predicting maize crop yields, targeting a problem familiar to many Sub-Saharan African contexts: how to forecast yield reliably using routinely collected agronomic variables. Their method centred on multiple linear regressions (MLR) combined with a backward elimination procedure, which iteratively removed statistically insignificant predictors from an initial pool of candidate variables until only the most influential factors remained. The stepwise approach enabled authors to develop a working predictive model and identify which agronomic variables had the highest explanatory power for maize yield outcomes. This model was able to predict maize yields with a usable degree of accuracy and identify statistically significant input variables, giving the study some practical value as an accessible, computationally inexpensive benchmark for yield estimation. However, the model is based on ordinary least squares regression and this is a major limitation of the model as it makes a key assumption of linear, additive relationships between predictors and yield. It is known that in agronomic systems, especially those that include living microbial inputs such as biofertilizers, there are threshold effects, saturation and interaction effects that cannot be accounted for by a linear specification.

This makes the Fashoto *et al*., approach a useful baseline but an inadequate final solution for the present study, which instead adopts nonlinear, ensemble-based learners precisely because they can model the more complex soil–microbe–yield relationships that a biofertilizer-focused prediction task requires. Xie *et al*., (2021) investigated whether ensemble machine learning could outperform traditional linear statistics in estimating soil biological indicators, specifically extracellular enzyme activity, in a reclaimed coastal saline agricultural landscape where soil salinity and heterogeneity make estimation unusually difficult. The objective of this study was to be more focused than a full yield-prediction exercise but still methodologically relevant: to compare directly a Random Forest regressor with multiple linear regressions on the same soil enzyme dataset, employing identical evaluation criteria so that any

performance gap could be attributed to the modeling approach itself and not to differences in the preparation of data. Random Forest models trained on soil physicochemical and biological variables, and compared with the linear alternative. The results showed that Random Forest consistently outperformed multiple linear regressions in the estimation of soil enzyme activity, which supported the broader consensus in the literature that tree-based ensembles are better suited for nonlinear soil-biological relationships than linear models. This finding is of direct relevance for the methodological choices made in the present paper, since it provides independent corroboration, outside the crop-yield literature in particular, that Random Forest is well suited to noisy, nonlinear soil-biological data of the kind generated by biofertilizer trials. The main limitation of the study, from this paper' perspective, is that it looked at a soil biological property rather than crop yield directly, and it was done in a single, ecologically unusual context (reclaimed coastal saline land), so its findings on model comparison, while methodologically instructive, cannot be assumed to generalize to yield prediction under more typical arable soil conditions. Nayak *et al*., (2022) tackled a practical and widely reported issue in precision agriculture: why can there be considerable variation in yields from farm to farm, for the same wheat variety, grown under nominally similar high-productivity management regimes. The objective of this study was to go beyond simple yield averages to explain this on-farm variability with interpretable machine learning methods that could attribute yield gaps to specific, actionable agronomic factors rather than treating yield as an unexplained black box. Methodologically, the authors combined farm-survey data with agronomic records from wheat fields in Northwest India, training interpretable machine learning algorithms capable of predicting yield while revealing which input variables were driving the predictions, an approach conceptually similar to the SHAP-based explainability adopted later in the present paper. The results pin-pointed specific, farm-level agronomic drivers that led to the observed yield gaps providing actionable insight that pure black-box prediction would not have provided and demonstrating that interpretability and predictive accuracy need not be traded off against each other. This is directly relevant to the choice of the present study to adopt ensemble regression models in conjunction with SHAP analysis rather than simply prediction accuracy.

The main limitation is in scope: the analysis was restricted to wheat within one Agro-ecological region of India and did not consider biofertilizer or other microbial treatments as candidate explanatory variables, leaving the specific question this paper investigates, how biological, rather than purely agronomic-management, inputs shape yield outcomes, unaddressed. Salvador *et al*., (2022) had a purely agronomic goal, namely to find out whether rhizobacteria and arbuscular mycorrhizal fungi (AMF) when applied separately or together have different effects on soybean growth in combination with organic compost which is a common practice in integrated soil-fertility management. Rather than a modeling exercise, their approach was an experimental inoculation trial with soybean growth outcomes compared among various treatment groups: untreated control, rhizobacteria only, AMF only, and a combined rhizobacteria-plus-AMF treatment, both with and without organic compost amendment. Growth and physiological parameters were measured and statistically compared across treatment groups. The results showed that rhizobacteria and AMF had different and treatment-specific effects on soybean growth, suggesting that these two classes of biofertilizer organisms are not interchangeable and operate through at least partially independent mechanisms, an insight that supports the present paper's decision to treat biofertilizer type as a distinct categorical feature rather than a single undifferentiated "treated versus untreated" variable. The study is not a predictive modeling study, but a controlled agronomic trial, and so brings no machine learning methodology of its own. Its contribution to this paper is simply to provide empirical, mechanistic evidence that different biofertilizer categories do indeed produce different plant-growth outcomes. This is the biological premise that supports the inclusion of biofertilizer type as an important predictive feature in the yield-prediction models constructed later in this work. This study was limited to soybean and the specific compost-amended conditions that were investigated. Srivastava *et al.,* (2022) employed deep learning to forecast winter wheat yield directly from environmental and phenological data, with the goal of capturing the temporal dynamics of crop development that are often missed in static, snapshot-based models.

Methodologically, the study trained a convolutional neural network (CNN) on time-series environmental data (such as temperature records) together with

phenological observations tracking the crop's developmental stages across the growing season, allowing the model to learn how the timing and sequencing of environmental conditions, not just their average values, relate to final yield outcomes. The trained CNN achieved accurate wheat yield predictions, and subsequent feature-attribution analysis highlighted maximum temperature as one of the most influential predictors of yield outcomes, a finding broadly consistent with agronomic understanding of heat stress effects on grain filling. This result is useful context for the present paper because it confirms that temperature-related variables carry substantial predictive weight even in deep-learning architectures quite different from the tree-based ensembles used here, lending independent support to the general practice of including climate-zone and temperature-related features in yield models. The main limitations, from the perspective of this paper, are twofold: the study is specific to winter wheat in one climatic context, and convolutional neural networks are computationally heavier and less straightforward to compress and deploy on low-power embedded hardware than the tree-based models (Random Forest and XGBoost) that this paper ultimately selects for its embedded-deployment objective. Huber *et al*., (2023) tackled a specific methodological limitation of SHAP-based explainability: that standard SHAP output can become difficult to interpret when a model uses many correlated or conceptually related input features, since importance is then diluted across numerous individually small SHAP values. Their aim was to develop and test a technique for grouping Shapley value feature importances so that conceptually related variables could be assessed collectively rather than only individually, without sacrificing the underlying accuracy of the Random Forest model being explained. Methodologically, the authors applied Random Forest regression to yield-prediction data and then computed grouped SHAP values, aggregating the individual Shapley contributions of related features into combined importance scores, and compared the interpretability and consistency of this grouped output against conventional, ungrouped SHAP analysis. The results showed that grouped SHAP produced more interpretable and more aggregated feature-importance rankings, making it easier to draw practical, high-level conclusions about which categories of variables mattered most, all without reducing the predictive accuracy of the underlying Random Forest model. This is directly relevant to the present paper's use of SHAP to interpret feature importance across many related soil and treatment variables.

The principal limitation noted is the additional computational overhead required to compute grouped SHAP values relative to standard SHAP, and the fact that the grouping technique was validated on a limited number of crops and datasets, leaving its generalizability to biofertilizer-specific yield data untested prior to the present work. Ma *et al.,* (2024) aimed to build an explainable machine learning framework for maize yield prediction that relied primarily on satellite-derived data, addressing the practical challenge of forecasting yield across large areas where ground-truth agronomic measurements are unavailable or too costly to collect at scale. Their method combined machine learning regression models with SHAP-based interpretation, applying both to remote-sensing inputs such as vegetation indices and climate-derived variables extracted from satellite imagery, rather than relying on field-collected soil or plant measurements. The results demonstrated that the models could achieve accurate, interpretable yield predictions purely from remotely sensed data, with SHAP analysis consistently identifying vegetation indices and climate variables as the leading predictors of maize yield outcomes. This confirms that explainable machine learning can be extended successfully to remote-sensing-based agricultural applications, broadening the evidence base, alongside Nayak *et al*., (2022) and Huber *et al*., (2023), that SHAP is a robust and versatile explainability tool across very different types of agricultural input data, from field records to satellite imagery. For the present paper, the key limitation is that the approach depends entirely on satellite data resolution, coverage, and availability, which are not always accessible to smallholder farmers in resource-limited settings, and, more specifically, the study excluded soil-microbial and biofertilizer treatment variables entirely from its feature set, meaning it offers no direct evidence on how such biological inputs would interact with, or be captured by, a satellite-driven yield-prediction pipeline. Mohan *et al*., (2024) pursued a conceptual and integrative aim: to articulate how artificial intelligence and explainable AI (XAI) techniques together could be combined into a coherent framework for precision crop yield prediction, addressing the broader question of how the agricultural AI field should evolve beyond

isolated point solutions toward integrated, trustworthy decision-support systems.

Their approach combined a review-style synpaper of existing AI techniques with an experimental component integrating machine learning models with XAI tools such as SHAP and LIME (Local Interpretable Model-Agnostic Explanations), examining how these tools could be combined to produce yield forecasts that are not just accurate but also interpretable and, therefore, more trustworthy to end users such as farmers and agronomists. The results of this integration demonstrated that XAI meaningfully improves user trust and provides more actionable insight into AI-based yield forecasts than accuracy metrics alone can offer, supporting the wider argument, echoed throughout the present paper, that explainability is not merely a diagnostic add-on but a core requirement for any agricultural AI system intended for real-world adoption. This aligns closely with the present study's own decision to pair predictive modelling with SHAP analysis rather than treating accuracy as the sole success criterion. The main limitation is that the work is framework-oriented and synpaperes existing techniques rather than empirically validating a single new model across a wide range of crops, treatments, or geographic contexts, leaving open questions about how the proposed integration performs in practice. Anuradha *et al*., (2024) aimed to address a persistent bottleneck in agricultural yield forecasting: the scarcity of large, labelled datasets needed to train accurate deep learning models, particularly in the domain of satellite-based yield estimation. Their approach used a Deep Convolutional Generative Adversarial Network (DCGAN) to generate synthetic satellite imagery resembling real agricultural scenes, which was then used to augment the training data available to a convolutional neural network (CNN) tasked with yield estimation, effectively expanding the diversity and volume of training examples beyond what was originally collected. The results showed that this DCGAN-based augmentation strategy significantly improved the CNN model's agricultural yield-forecasting accuracy relative to training on the original, smaller dataset alone, providing clear evidence that generative adversarial approaches can meaningfully compensate for limited real-world agricultural data. This finding is particularly relevant to the present paper, which similarly confronts a small real-world biofertilizer dataset and adopts a comparable strategy of using GAN-generated synthetic samples to enlarge the training pool, though applied to tabular rather than image data. The principal limitations, from this paper's perspective, are that the approach is computationally demanding, requiring both GAN training and CNN training in sequence, that it depends on access to high-resolution satellite imagery infrastructure not always available in developing-country contexts; and that the resulting model was never evaluated for feasibility on low-power embedded hardware, leaving the deployment question entirely unaddressed. Abekoon *et al.,* (2025) aimed to predict soil nitrogen, phosphorus, and potassium (N, P, K) nutrient content in cabbage cultivation using an explainable machine learning approach, addressing the practical need for farmers and agronomists to understand not just what nutrient levels a model predicts, but why.

Methodologically, the study combined machine learning regression models with two complementary explainability techniques, SHAP and LIME, applying both globally, across the full dataset, and locally, to individual predictions, allowing users to interrogate both overall feature importance and the specific reasoning behind any single soil-nutrient forecast. The results demonstrated that the combined ML-plus-XAI approach produced accurate nutrient-content predictions alongside genuinely interpretable, feature-level explanations, reinforcing the growing consensus in the literature, also reflected in Huber et al. (2023) and Ma et al. (2024), that SHAP and LIME can be deployed together productively rather than as competing alternatives. This dual-explainability strategy offers a useful methodological reference point for the present paper's own use of SHAP to interrogate feature importance in its yield-prediction models. The main limitation relevant to this paper is one of scope: the study addressed soil nutrient content rather than final crop yield as its target variable, and it was validated on a single crop, cabbage, meaning its findings on model architecture and explainability technique cannot be assumed, without further testing, to transfer directly to a multi-crop, yield-focused, biofertilizer-driven prediction task of the kind undertaken here. Hernandez-Hidalgo *et al*., (2025) aimed to develop a fully offline, low-power precision-irrigation system capable of operating in resource-constrained environments without reliance on cloud connectivity, directly addressing the practical needs of smallholder farmers who often lack stable internet access. Their method involved training three candidate models, Linear

Regression, Random Forest, and a Neural Network, on vegetation and environmental data including the Normalized Difference Vegetation Index (NDVI), then quantizing the best-performing models from 32-bit floating-point precision down to 8-bit integer precision using TensorFlow Lite, before deploying the compressed models onto microcontroller-class hardware for real-time, on-device inference.

The results were striking from an embedded-deployment perspective: quantization achieved up to approximately 82% reduction in model size with only minimal loss of predictive accuracy, and among the three models compared, Random Forest delivered the best performance, with a root mean squared error of 0.062. This is one of the most directly relevant studies to the present paper, since it provides a concrete, quantified precedent for exactly the kind of pruning-and-quantization pipelines this paper applies to its own XGBoost yield-prediction model, and it independently confirms that Random Forest-family models compress well without catastrophic accuracy loss. Its principal limitation, from this paper's perspective, is that the application domain was irrigation and NDVI anomaly detection rather than yield prediction, and the work was validated in a single-crop context, leaving open whether comparable compression ratios and accuracy retention would hold for a biofertilizer-treatment yield-regression task. Onyeke *et al*., (2026) aimed to improve cowpea yield prediction under climate variability specifically within the North Central Nigerian Agro-ecological zone, a context directly relevant to the present paper given its shared Nigerian setting. Their method combined a traditional process-based crop simulation model, of the kind long used in agronomic forecasting, with a neural network post-processing stage designed to correct systematic biases and residual errors left uncorrected by the process-based model alone, creating a hybrid pipeline that leveraged the mechanistic grounding of process-based simulation together with the pattern-recognition strength of machine learning. The results showed that this hybrid process-based-plus-neural-network approach improved yield-prediction accuracy over the process-based model used in isolation, demonstrating the practical value of combining mechanistic and data-driven methods within a single Nigerian agricultural forecasting pipeline, and offering local, regionally grounded evidence that machine learning-based correction techniques can meaningfully improve on established simulation approaches under Nigerian climatic conditions. This is valuable corroborating evidence for the present paper, which is similarly situated within a Nigerian institutional and Agro-ecological context. The main limitation is that the study centred on climate variability as its primary driver of interest rather than biofertilizer application, and it focused on a single crop, cowpea, within a single Nigerian region, meaning its findings on hybrid-model design, while encouraging, do not directly address the biofertilizer-yield relationship or extend to model compression and embedded deployment.

Kwaghtyo *et al.,* (2026) aimed to address data scarcity and class imbalance in tabular crop-recommendation datasets, a challenge structurally similar to the small, imbalanced biofertilizer trial data confronted by the present paper. Their method involved designing a class-conditioned Generative Adversarial Network, termed CropGAN, purpose-built to generate realistic synthetic tabular agricultural data while explicitly respecting the categorical class structure of the original dataset, and benchmarking this approach against two established alternatives, the Synthetic Minority Oversampling Technique (SMOTE) and Variational Autoencoders (VAEs), on the same underlying crop-recommendation data. The results showed that CropGAN produced synthetic samples that improved downstream crop-recommendation model performance more effectively than either SMOTE or VAE baselines, providing strong evidence that GAN-based approaches, when properly adapted to tabular and categorical agricultural data rather than simply repurposed from image-generation contexts, can outperform more conventional data-augmentation techniques. This finding is highly relevant to the present paper's own use of GAN generated synthetic samples to enlarge its real, relatively small biofertilizer dataset, lending methodological support to that design choice from a directly comparable Nigerian agricultural-data context. Its main limitation is that the technique was developed and validated for a classification-style crop-recommendation task rather than continuous yield regression, and testing was confined to one regional dataset, leaving open the question, which the present paper addresses, of whether a similar conditional-GAN augmentation strategy transfers effectively to a continuous, biofertilizer-treatment yield-prediction problem.

Jahan and Nassiri-Mahallati (2026) conducted the single most directly relevant study reviewed in this paper, aiming to compare eight machine learning algorithms for predicting corn grain yield specifically under four distinct biofertilizer treatments: an untreated Control, arbuscular mycorrhizal fungi (AMF) alone, plant growth-promoting rhizobacteria (PGPR) alone, and a combined AMF-plus-PGPR treatment. Their method drew on a genuinely substantial feature set, ninety-six samples characterized by seventy-three features, comprising thirty-two primary agronomic and physiological measurements plus forty-one engineered interaction features, collected over a two-year field trial, with ANFIS, Transformer, Artificial Neural Network, Support Vector Regression, LightGBM, XGBoost, an Enhanced Deep Neural Network, and Support Vector Machine models all trained, tuned, and compared, and SHAP used throughout to interpret feature contributions. The best models were ANFIS ($R^2 = 0.555$), Transformer ($R^2 = 0.545$), and Artificial Neural Network ($R^2 = 0.518$) while XGBoost demonstrated a modest R^2 of 0.339 on this dataset. SHAP analysis consistently identified canopy temperature and the interaction between root colonization and nitrogen-related variables as the dominant predictors of biofertilizer-treated yield outcomes. This is the only reviewed study to model biofertilizer treatment type directly as a yield predictor with machine learning, and is the closest existing precedent to the present paper, although it did not go as far as any model compression or embedded-deployment work. The authors themselves concede that the reported model rankings and accuracy figures are limited in terms of generalizability by the relatively small sample size of ninety-six observations and the trial's focus on a single region and hybrid.

Bracho-Mujica *et al*., (2024) aimed to assess the combined influence of anticipated shifts in climatic means and variability, along with the utilization of various Agro-technologies, on future wheat and maize yields at ten geographically distinct locations worldwide. Their research tackled the broader question of the relationship between climate change and technological adaptation over the long-term, rather than the yield outcome of an individual season. Their approach was based on crop simulation modeling, a process-based not machine-learning approach, used consistently across all ten sites under a set of future climate change scenarios and associated Agro-technology (AT) adoption pathways, allowing the authors to disentangle and compare the relative contributions of climatic change versus technological change to projected yield outcomes at each site. Results show that the combined effects of Climatic and Agro-technological change significantly modified projected wheat and maize yields, but the direction and magnitude of these changes varied greatly from site to site, highlighting the significance of location-specific modeling rather than assuming uniform climate-change effects across different agro-ecological regions. This global multi-site view is a good complement to the more local and biofertilizer specific present paper. It highlights the general principle that yield-relevant relationships are highly context-dependent. From the perspective of this paper, the most obvious limitation is methodological. This study relies entirely on process-based crop simulation rather than machine learning, and biofertilizer or other biologically mediated inputs are not included as a separate technology pathway in the Agro-technology scenarios.

III. MATERIALS AND METHOD

3.1 Introduction

In this section, an outline of the study idea was presented, resources used and the methodology employed in developing and assessing the AI-based model for predicting crop yield under the influence of biofertilizer. It combines computational modeling with a mindful consideration of the data behind it, based on reproducibility and aligned with sustainable agriculture objectives.

3.2 Materials

3.2.1 Software and Programming Environment

The study was conducted on the following software stack:

a. Programming language: Python 3.10+
b. Core libraries:
   i. Data handling - Pandas, NumPy
   ii. Visualization - Matplotlib, Seaborn
   iii. Machine learning - Scikit-learn, XGBoost, LightGBM
   iv. Deep learning (for future extension) - TensorFlow/Keras
   v. Model optimization and embedded deployment - TensorFlow Lite, ONNX, Scikit-learn pruning utilities

vi. Explainability - SHAP (SHapley Additive exPlanations)

Development environment: Google Colab/Jupyter Notebook for model training, and for embedded-deployment simulation.

3.2.2 Hardware Requirements

Training phase: a standard laptop/desktop (Intel i7/i9 or equivalent, 16–32 GB RAM, with an NVIDIA GPU to accelerate XGBoost training).

Embedded optimization and deployment phase (optional):

Raspberry Pi 4 Model B (4 GB RAM) - the primary target device.

ESP32 DevKit - used to simulate IoT sensor integration.

Sensors (for future field validation): soil moisture, pH, temperature, and NPK sensors (e.g., DFRobot SEN0321).

3.2.3 Datasets Used

Two complementary datasets, totalling 5,500 records, were used; combining data from published biofertilizer field trials with CTGAN-generated synthetic samples (500 real observations and 5,000 synthetic ones) all included in two folders:

nfb_maize_yield.csv - a maize-specific dataset covering soil pH, nitrogen, phosphorus, potassium, moisture, inoculum dose, encoded NFB treatment, microbial CFU count, climate zone, and yield (kg/ha).

mixedcrops_yield_analysis.xlsx - a multi-crop dataset covering crop type, biofertilizer type, percentage yield increase, baseline yield, nutrient-use efficiency, and soil parameters.

3.2.4 Microbial Strains Represented in the Data

Although the study's focus is on modelling rather than laboratory work, the source data reference the following strains:

Nitrogen-fixers: *Azotobacter chroococcum*, *Rhizobium leguminosarum*

Phosphate-solubilizers: *Bacillus megaterium*, *Pseudomonas fluorescens*

Consortia: mixed PGPR formulations

3.3 Methods

3.3.1 Data Acquisition and Pre-Processing

Datasets were imported using Pandas.

Missing values were handled by imputing numerical features with the median and dropping rows with more than 30% missing data.

Categorical variables - biofertilizer type, soil type, climate zone, and crop type, were label-encoded.

All numerical features (pH, N, P, K, moisture, inoculum dose, CFU) were standardized using a StandardScaler.

Feature engineering included creating interaction terms (e.g., NFB treatment × soil nitrogen) and, where needed, one-hot encoding of multi-category variables.

The target variable was crop yield (kg/ha), or percentage yield increase where applicable.

3.3.2 Data Splitting

The combined dataset was divided into training (70%), validation (10%) and test (20%) subsets. Each model was trained on the training set to learn the relationship between input features and crop yield. The validation set was used to tune hyperparameters and help guard against overfitting. The test set, held out from training entirely, was used to provide an unbiased measure of how well each model generalizes to unseen data.

3.3.3 Model Selection and Training

Several regression algorithms were trained and compared:

a. Random Forest Regressor (an ensemble of decision trees)
b. XGBoost Regressor (gradient boosting - the primary model of interest)
c. Benchmark models: Linear Regression, Ridge, Lasso, Decision Tree and XG-Boosting.

3.3.3.1 Training Procedure

a. Define the feature matrix (X) and target vector (y).
b. Initialize each model with default hyperparameters.
c. Tune hyperparameters (e.g., n_estimators, max_depth, learning_rate for XGBoost).
d. Fit each model on the training data.

3.3.4 Model Evaluation

Model performance was assessed using:

a. Coefficient of determination (R^2)
b. Root Mean Squared Error (RMSE)
c. Mean Absolute Error (MAE)
c. Mean Absolute Percentage Error (MAPE)
d. Nash–Sutcliffe Efficiency (NSE)

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

These metrics are defined in Eq. (2), Eq. (3), Eq. (4), Eq. (5), Eq. (6) and Eq. (7):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \quad SS_{res} = \sum (y_i - \hat{y}_i)^2, \quad SS_{tot} = \sum (y_i - \bar{y})^2 \quad \ldots (2)$$

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1} \quad \ldots (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2} \quad \ldots (4)$$

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i| \quad \ldots (5)$$

$$RMSLE = \sqrt{\frac{1}{n}\sum (\log(1 + y_i) - \log(1 + \hat{y}_i))^2} \quad \ldots (6)$$

$$MAPE = \frac{100}{n}\sum \frac{|y_i - \hat{y}_i|}{y_i} \quad \ldots (7)$$

Where y_i is the observed yield, $\hat{y}_i$ the predicted yield, $\bar{y}$ the mean observed yield, and n the number of observations.

3.3.5 Biofertilizer Efficacy and Feature Importance Analysis

a. Treatment comparison: data were grouped by biofertilizer type to compute mean yield improvement relative to the untreated control.
b. Feature importance: assessed using the built-in importance scores from the tree-based models (feature_importances) and, for deeper interpretability, SHAP values at both the global and local level.

The SHAP value used for interpretability is given in Eq. (8):

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \ldots (8)$$

where φi is the SHAP value for feature i, F is the full feature set, and S is a subset of features excluding i.

3.3.6 Model Optimization for Embedded Deployment

The best-performing model (XGBoost) was optimized for deployment on low-power hardware through the following steps:

a. Pruning: low-importance features and weights were removed using cost-complexity pruning.
b. Quantization: 32-bit floating-point weights were converted to 8-bit integers using TensorFlow Lite (dynamic or full-integer quantization).
c. Knowledge distillation: a smaller "student" model was trained to reproduce the behaviour of the full XGBoost "teacher" model.
d. Conversion to TFLite: the final model was exported as a tflite file.
e. Inference testing: the model was deployed on a Raspberry Pi 4 and evaluated for model size (MB → KB), inference latency (ms), memory usage (RAM), and estimated power consumption (via multimeter measurement).

The quantization relationship is summarized in Eq. (9):

$$Q(x) = \text{round}\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}} \times (2^b - 1)\right) \ldots$$

where x is the original floating-point value, x_min and x_max define its range, and b is the target bit-width.

3.4 Proposed Framework for this Study

The study is presented as an end-to-end pipeline of data collection, machine learning modeling, optimization for embedded AI deployment and evaluation. The aim is to fill a gap in biofertilizer-

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

specific modeling, by explicitly including biological, soil and environmental components, compared to typical chemical-fertilizer models, which only focus on nutrient inputs. Finally, the architecture is amenable to the possibility of iterative refinement: results from any one stage fermentation, formulation, or field experiment- can be fed back into adjustments at any other stage. This loop is key to the design of biofertilizer formulations that work well in a range of soils, climates and cropping systems and to bridge lab research to field-level application. The corresponding flowchart is shown in Figure 2.

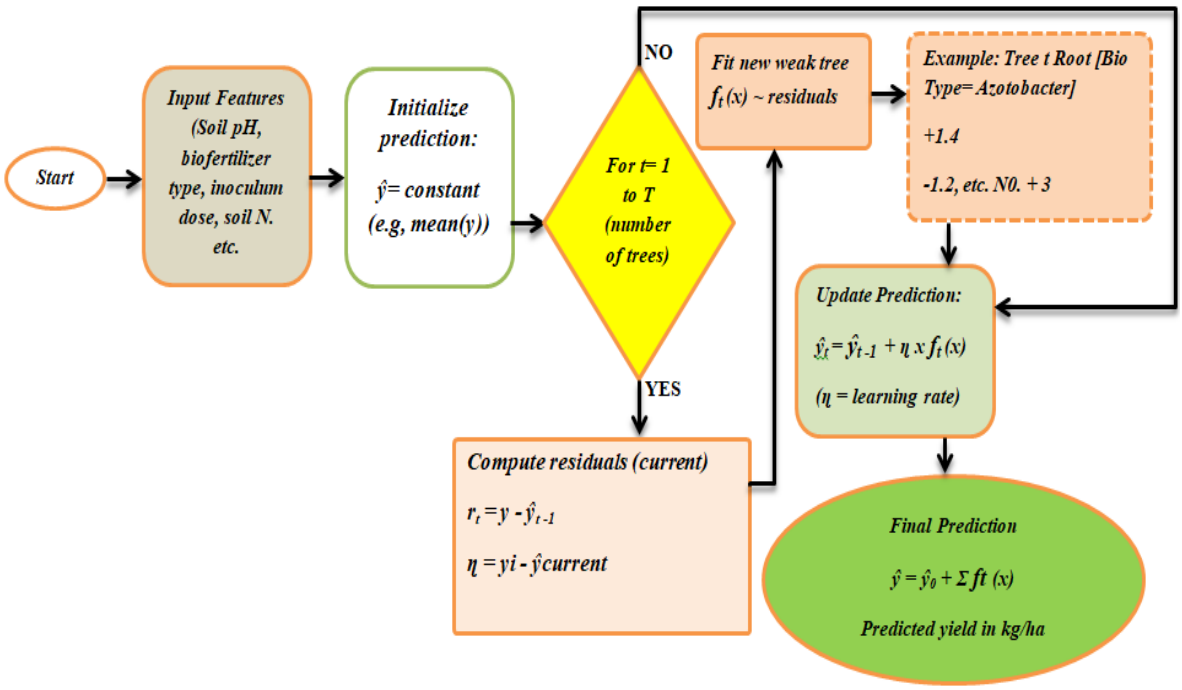

Figure 1: *Proposed Biofertilizer Tree-Ensemble Modelling Flowchart*

This graphic a suggested workflow of the proposed study that can be used to predict crop production based on agricultural and soil-related variables. The model creates a sequence of weak choice trees where each subsequent tree attempts to correct the errors of the previous trees. The process starts with the input features that are variables like soil pH, type of biofertilizer, inoculum dose, soil nitrogen content (soil N) and other agronomic aspects. These features are the predictors to predict the crop yield. A first prediction, y ˆ 0, is then made, usually using a constant value such as the mean of the target variable. Then the algorithm enters an iterative boosting procedure that lasts for T trees. At each iteration the residuals (prediction errors) are calculated as:

$$r_t = y - \hat{y}_{t-1} \quad ...(10)$$

where y is the actual yield and y ˆ$_{(t-1)}$ is the prediction from the previous iteration. These residuals show the amount of the goal value that remains unexplained by the existing model. A new weak decision tree is subsequently trained to predict the residuals.

The example given in the figure displays a tree split based on the feature Bio Type = Azotobacter, with terminal node values signifying revisions to the prediction. Rather of estimating the goal yield directly, this tree learns the residual faults from the prior model.

The projections are later adjusted according to:

$$\hat{y}_t = \hat{y}_{t-1} + \eta f_t(x) \quad ... (11)$$

where y is the actual yield and $\hat{y}_{t-1}$ is the prediction from the previous iteration. These residuals represent the portion of the target value that remains unexplained by the current model and η is the learning rate. The learning rate controls how much each tree contributes to the overall model and helps prevent overfitting by making gradual improvements.

This process is repeated until all T trees have been constructed. The final prediction is obtained by combining the initial prediction with the contributions from all weak learners:

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

$$\hat{y} = \hat{y}_0 + \sum_{t=1}^{T} f_t(x) \qquad \ldots (12)$$

The resulting output is the expected crop yield in (kg/ha). Overall, the figure demonstrates how gradient boosting incrementally improves prediction accuracy by fitting successive decision trees to the residual errors of previous trees, enabling the model to capture complex nonlinear relationships between soil characteristics, biofertilizer treatments, and crop yield.

IV. RESULS AND DISCUSSION

4.1 Introduction

This section presents the main results from model building and analysis. The combined visualizations and results suggest the feasibility and effectiveness of the proposed AI-based strategy in predicting crop output under various biofertilizer treatments.

4.2. Maize dataset distribution

Figure 1 presents an in-depth analysis of the distribution of crop production across several treatment groups using a histogram, boxplots, violin plots and a Q-Q plot. The distribution of the yield values is skewed to the higher end with an average of 8,074 kg/ha and a median of 8,281 kg/ha (Figure 1). The mean and median comparison shows that the mean is much lower than the median, indicating a slightly negatively skewed distribution and some lower yield observations. The boxplots indicate significant differences among treatments. The Azotobacter + Azospirillum combination has the highest median yield and least variability, suggesting better and more consistent performance. The control treatment had the lowest median output and the greatest dispersion, indicating poorer productivity and larger variability. Similar patterns are also visible in the violin plots, which represent the yield values distribution and density across treatment groups. Microbial inoculant treatments tend to have distributions skewed toward higher yields, while the control is skewed toward lower yields. Moreover, the Q-Q plot shows considerable departures from the reference line, especially at the tails of the distribution, implying that the yield data do not strictly follow a normal distribution. The figure shows that, overall, biofertilizer treatments significantly increase crop production compared to the untreated control. The combination of Azotobacter and Azospirillum provides the most effective and consistent increase in yield.

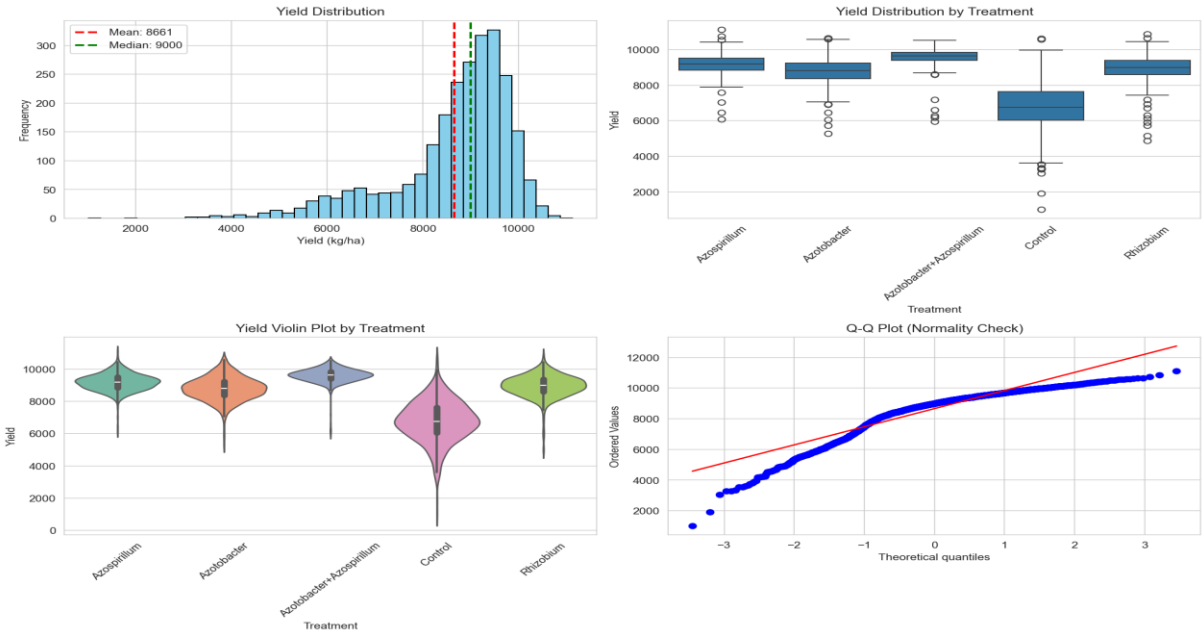

Figure 1: *Maize Dataset Distribution Analysis.*

4.2.1 Mixed-Crops Dataset Distribution.

The detailed exploratory analysis of crop output across five different biofertilizer treatment groups namely, AMF, Combined N+P, N-Fixers, P-Solubilizers and other treatments is represented in Figure 2. The histogram (top left) shows a strongly bimodal distribution, with one group of yield values clustered between approximately 2,000 and 6,000 kg/ha and the other between 20,000 and 25,000 kg/ha. The mean yield (8,378 kg/ha) is significantly

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

greater than the median yield (4,274 kg/ha), suggesting that the data is positively skewed with some very high yield values. This suggests considerable variation in treatment performance and likely differences in crop response under experimental conditions. The boxplots (top right) compare the distributions of the yield in the treatment groups. The Other treatment category has the highest median yield and the most variability, indicating both high yield potential and uneven performance. Conversely, the AMF, Combined N+P, N-Fixers, and P-Solubilizers treatments generally show lower median yields clustered around 2000 to 5000 kg/ha, but all groups contain many high-yield outliers reaching approximately 18,000 to 25,000 kg/ha. These outliers show that in some cases biofertilizer treatments can significantly increase crop yield. The high interquartile range reported for the other treatment also indicates more variability in yield outcomes than the remaining groups. The violin plots (bottom left) provide further information on the yield density and distribution patterns. Two distinct density peaks are observed for most treatment groups, one at lower yield levels and one at higher yield levels, consistent with the bimodal pattern observed in the histogram.

The broader parts of the violins reflect locations where observations are more concentrated, whereas the narrow regions represent less frequent yield values. The Other treatment displays a larger distribution extending into the higher yield range, whereas the remaining treatments are centered predominantly at lower yields with smaller high-yield clusters. The Q-Q plot (bottom right) was used to assess the normality of the yield data. The large departure of the measured quantiles from the reference line suggests that the data do not follow a normal distribution. The typical S-shaped curve and flattening at the upper tail suggest the presence of skewness, multiple distribution modes, and extreme values. Consequently, the assumption of normality is violated, signalling that non-parametric statistical approaches or data manipulations may be more appropriate for following investigations. Overall, Figure 2 reveals high diversity in crop output among treatment groups, with indications of bimodality, skewness, and multiple outliers. Although all biofertilizer treatments indicate instances of high productivity, the other treatment category records the highest median yields and largest variability, while the yield data as a whole vary greatly from a normal distribution.

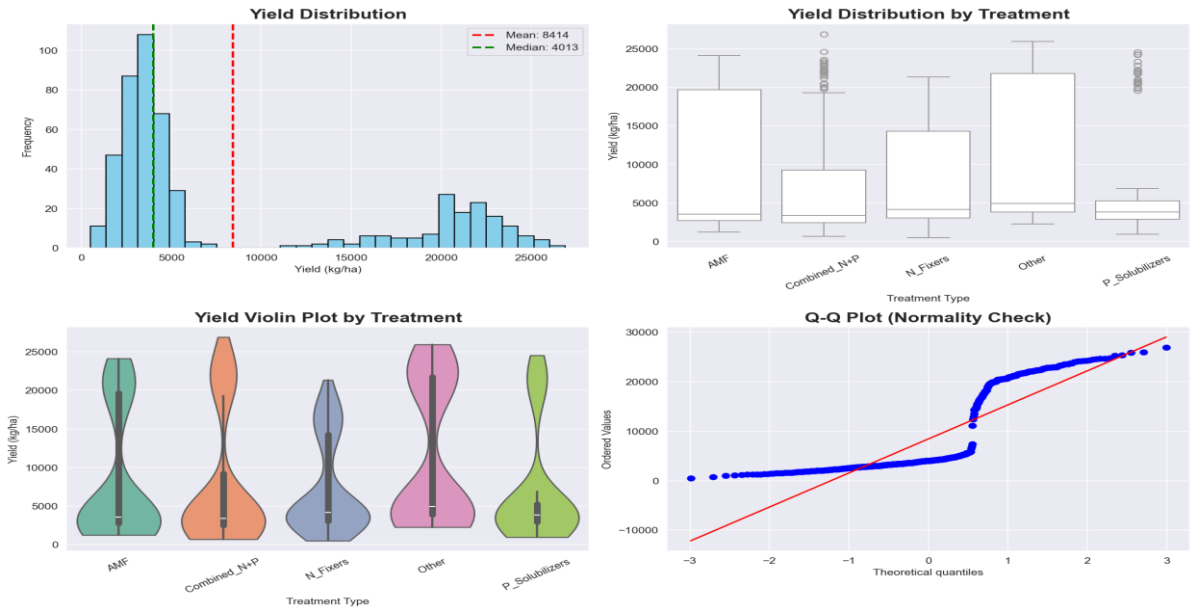

Figure 2: *Mixed-Crops Dataset Distribution Analysis.*

4.2.2 Correlation Structure

Figure 3. Correlation heatmap showing the linear correlations between soil parameters, microbiological features, treatment factors and maize yield. Correlation coefficients range from -1 to 1. The closer the value is to 1, the stronger the positive relationship. The closer the value is to -1, the stronger the negative relationship. Values near 0 indicate little or no linear relationship. The heatmap shows that soil nitrogen (soil_N) has the strongest positive

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

association with crop production (crop_yield_kg_ha) with a correlation coefficient of 0.65. This demonstrates that nitrogen availability is an important factor affecting maize productivity, and higher nitrogen availability is generally correlated with increased crop productivity. Moderate positive correlation between inoculum dose and crop output (r = 0.32), indicating that increasing the quantity of microbial inoculant may contribute to better maize production, although its influence is less pronounced than that of soil nitrogen. In contrast, the remaining variables, including soil pH, soil phosphorus (soil_P), soil potassium (soil_K), soil moisture, microbial colony-forming units (microbial_CFU), and crude fiber, demonstrate correlation coefficients that are very near to zero with respect to crop production. These weak associations suggest that, within the dataset, these factors have limited direct linear influence on maize production or that their effects may be represented through more complicated, non-linear interactions. Furthermore, correlations among the predictor variables themselves are often very low, showing limited multicollinearity within the dataset. This is useful for modeling predictions since it indicates that the variables are providing significantly independent information, and are unlikely to introduce redundancy in conventional models using machine learning. The heatmap (Figure 3) suggests that soil nitrogen is the most important predictor of maize production, followed by inoculum dose, while the other factors show relatively small linear correlations with crop productivity. The data suggest that nitrogen management and the correct application of inoculum may play an important role in increasing maize production under the conditions studied.

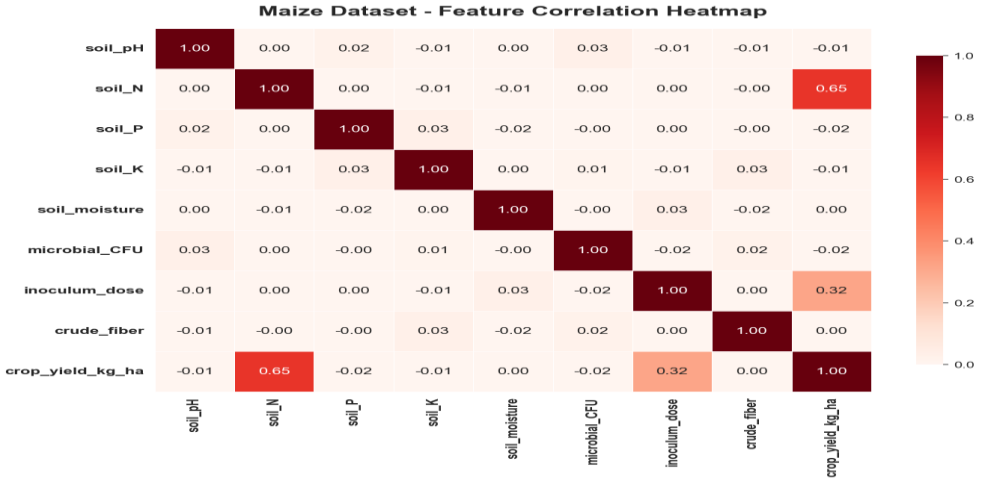

Figure 3: *Maize Dataset Correlation Heatmap.*

4.2.3 Feature Relationships

Some relationships between soil and microbial parameters and maize yield are presented in Figure 4. The results show that the soil nitrogen (soil_N) has the highest correlation with the maize yield with the correlation coefficient r =0.650 and p = 0.0000 which is statistically significant and positively correlated. Hence, higher soil nitrogen is generally related to higher maize yield in the dataset. Soil pH showed little linear relationship with yield (r = −0.006, p = 0.6531), indicating that pH changes within the range we measured showed little association with production. Similarly, soil phosphorus (soil_P) has a very weak negative correlation (r = −0.019, p = 0.1903) and microbial CFU has an extremely weak negative correlation (r = −0.015, p = 0.2818) which are statistically insignificant at the 5% significance level. In general, the data indicate that of all the factors studied, soil nitrogen is the most important single factor influencing yield of corn. However, the weak individual relationships observed for soil pH, phosphorus and microbial CFU do not necessarily mean that these variables are unimportant, as their effects may occur through nonlinear relationships or interactions with biofertilizer treatment, inoculum dose, soil characteristics and climatic conditions. Thus, this study supports the use of multivariate machine-learning models to capture simultaneously complex interactions among the many factors determining maize output.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

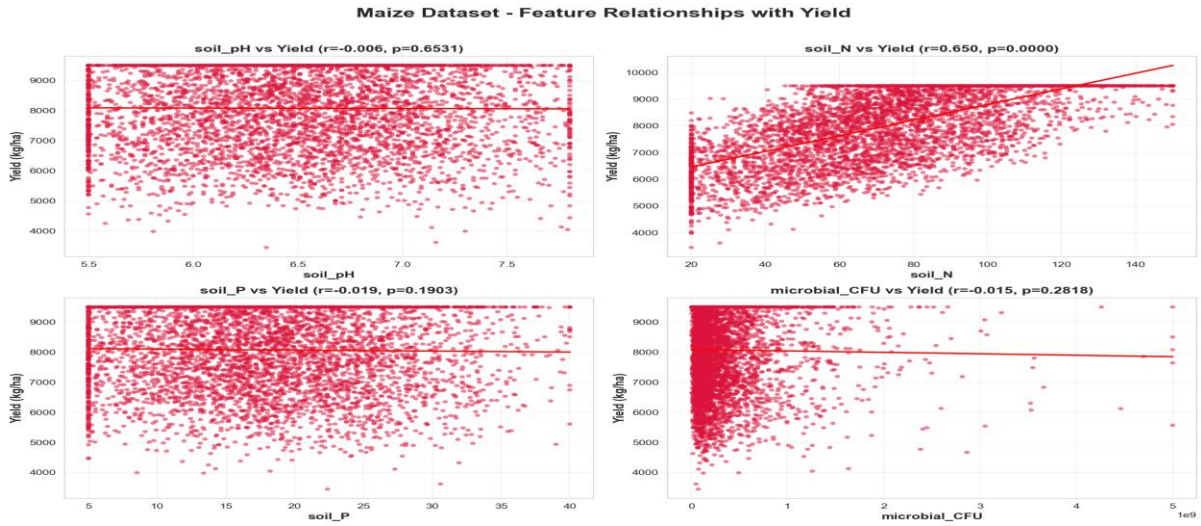

Figure 4: Relationships Maize Dataset Feature

Figure 5 illustrates relationships between certain soil and microbial parameters and maize yield. The results show that soil nitrogen (soil_N) has the highest correlation with maize yield with a correlation coefficient of r = 0.650 and p = 0.0000 (statistically significant), hence higher soil nitrogen levels are generally associated with higher maize yield in the dataset. Soil pH showed little linear relationship with yield (r = −0.006, p = 0.6531), indicating that pH changes within the range we measured showed little association with production. Likewise, the soil phosphorus (soil_P) has a very weak negative association (r = −0.019, p = 0.1903) and microbial CFU has an extremely weak negative association (r = −0.015, p = 0.2818) which are statistically insignificant at 5% significance level. Overall, the data indicate that the most significant individual variable among the variables examined in relation to maize yield is soil nitrogen. However, the weak individual relationships observed for soil pH, phosphorus and microbial CFU do not suggest unimportance of these variables as their effects may be through nonlinear relationships or interactions with biofertilizer treatment, inoculum dose, soil characteristics and climatic conditions. This study thus supports the use of multivariate machine-learning models, which can capture complex interactions among the multiple elements determining maize output.

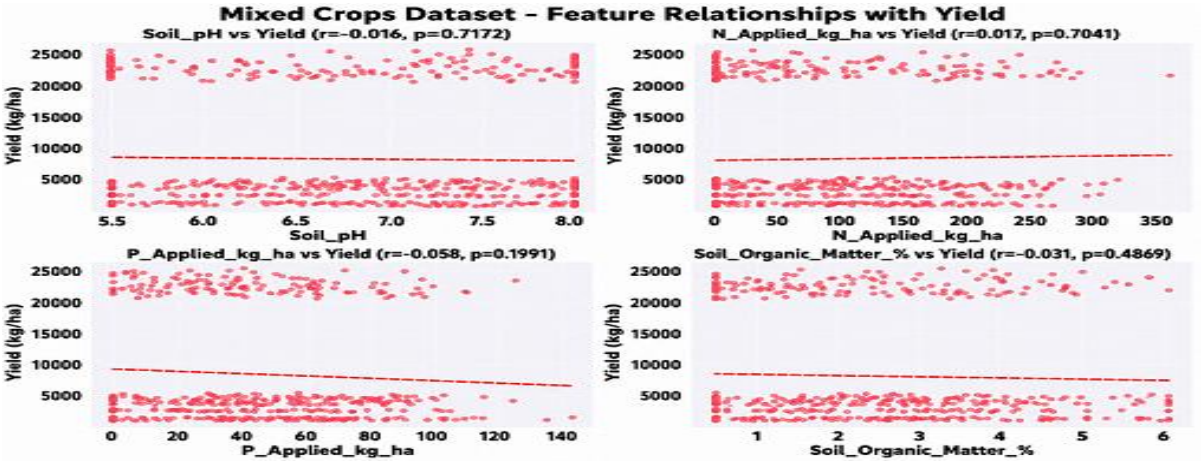

Figure 5: Mixed-Crops Dataset Feature Relationships.

4.3 Model Performance Comparison

4.3.1 Overall Comparison.

Figure 6 presents a detailed evaluation of eight regression models on the mixed-crops (or related) yield prediction task in four major aspects. The top-

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

left panel shows the R^2 scores. For all three linear models (Linear, Ridge and Lasso Regression) the R^2 score is rather modest (around 0.618). The tree-based and boosting models perform much better. The R^2 scores for Decision Tree, Random Forest, and the three gradient boosting variants (Gradient Boosting and XGBoost) are 0.881 and 0.947 - 0.948, respectively. This ranking is reflected in the top right panel (RMSE) and bottom left panel (MAE): linear models have the highest errors (RMSE ≈ 812–813 kg/ha, MAE ≈ 687 kg/ha), Decision Tree is in-between, and the ensemble/boosting models have the lowest errors (RMSE down to 300 kg/ha and MAE down to 226 kg/ha for XGBoost). The bottom right panel plots cross-validated R^2 scores with error bars given by the standard deviation; again, the same performance hierarchy holds, and the very small error bars on the boosting models indicate steady, dependable generalization. Overall, the figure clearly shows that non-linear ensemble approaches (especially XGBoost, LightGBM and Gradient Boosting) outperform linear baselines significantly for the prediction of yield in this dataset

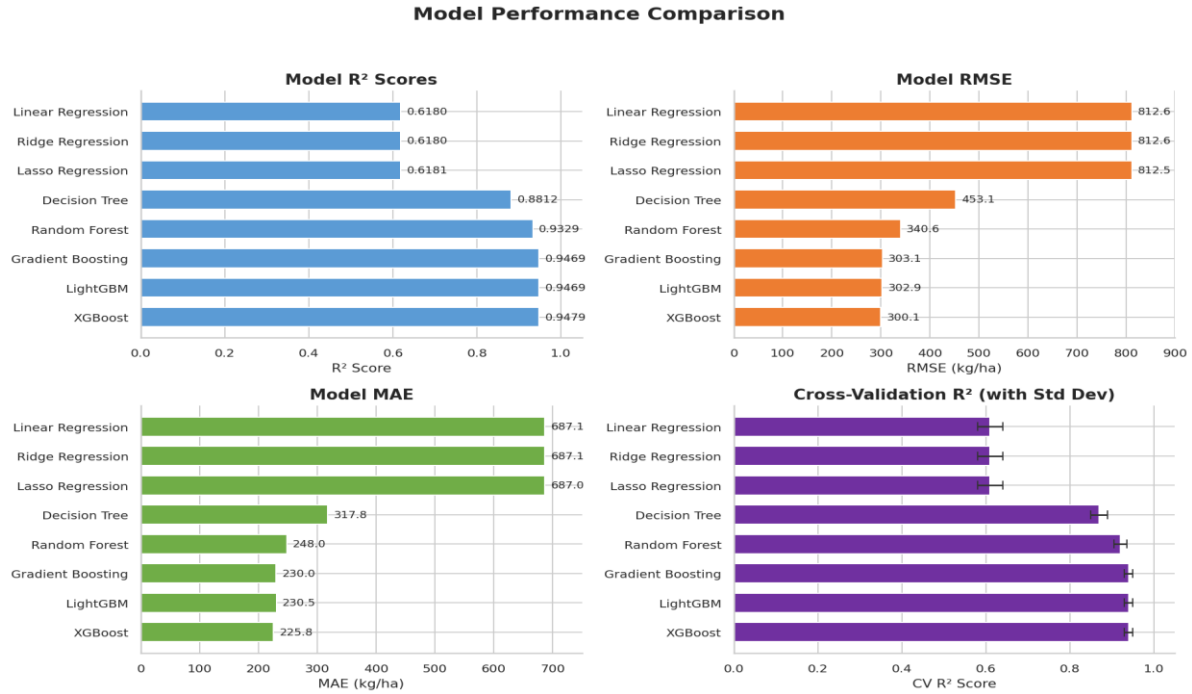

Figure 6: Maize Dataset Model Comparison.

Figure 7 compares the performance of eight regression models on what appears to be a different (or differently processed) yield-prediction job. The linear models (Ridge, Linear, and Lasso Regression) perform badly, each attaining a R^2 of only ~0.08 and creating very large errors (RMSE = 7,304 kg/ha and MAE ≈ 6,418–6,420 kg/ha). In sharp contrast, all tree-based and boosting models achieve near-perfect fits: Decision Tree reaches R^2 = 0.9974, while Gradient Boosting, XGBoost, and Random Forest attain R^2 values between 0.9986 and 0.9990. Correspondingly, their error metrics reduce considerably (RMSE 235–389 kg/ha and MAE 114–186 kg/ha), with Random Forest appearing as the strongest overall model. The cross-validation panel verifies this pattern: linear models show low and slightly fluctuating CV R^2 scores, but the ensemble approaches retain exceptionally high and stable CV R^2 values (around 1.0) with small standard deviations. The image therefore emphasizes a large performance gap, demonstrating that the underlying relationships in this specific dataset are highly non-linear and are captured almost perfectly by tree-based ensembles.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

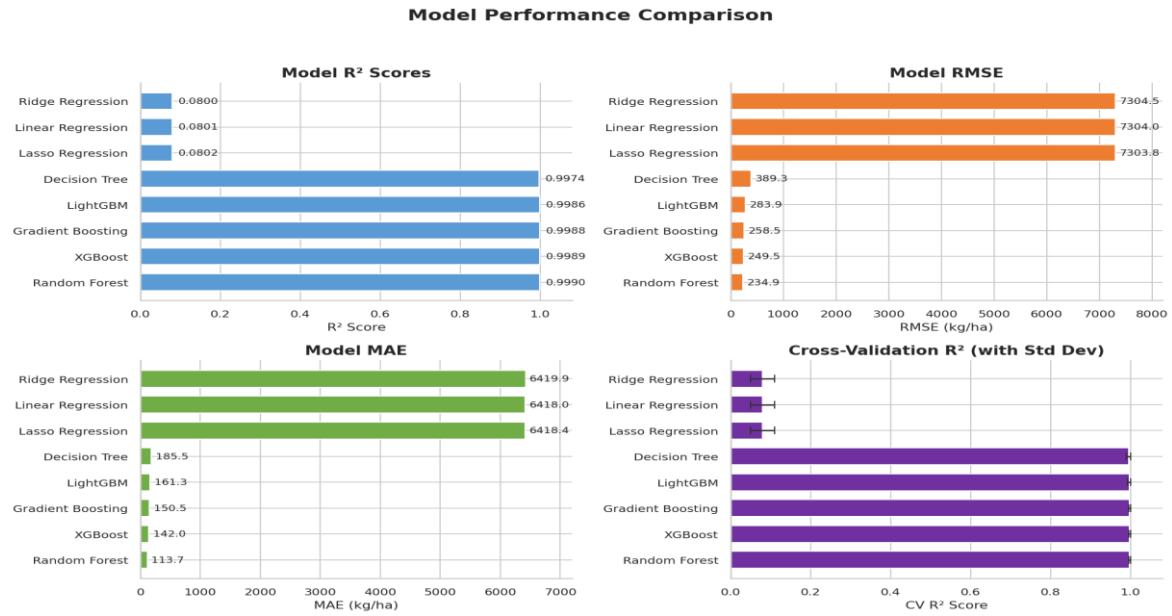

Figure 7: Mixed-Crops Dataset Model Comparison.

4.3.2 XGBoost - Actual Versus Predicted.

Figure 8 shows a detailed prediction analysis for the XGBoost model. The left panel shows a scatter plot of actual against expected yield values. Points cluster strongly around the dashed red "Perfect Prediction" line (y = x), demonstrating remarkable agreement between observed and anticipated yields. The model yields outstanding performance metrics: $R^2 = 0.9479$, RMSE = 300.06 kg/ha, and MAE = 225.85 kg/ha. The right panel shows the residual plot (residuals = actual minus anticipated) against predicted yield. Residuals are centred near zero (mean = −11.50 kg/ha) with a standard deviation of approximately 300 kg/ha, and they look randomly spread without strong systematic patterns, although a faint negative trend line is discernible at higher expected values. Overall, the figure illustrates that XGBoost offers accurate and mostly unbiased predictions for crop yield in this dataset, with residual variance remaining very steady across the prediction range.

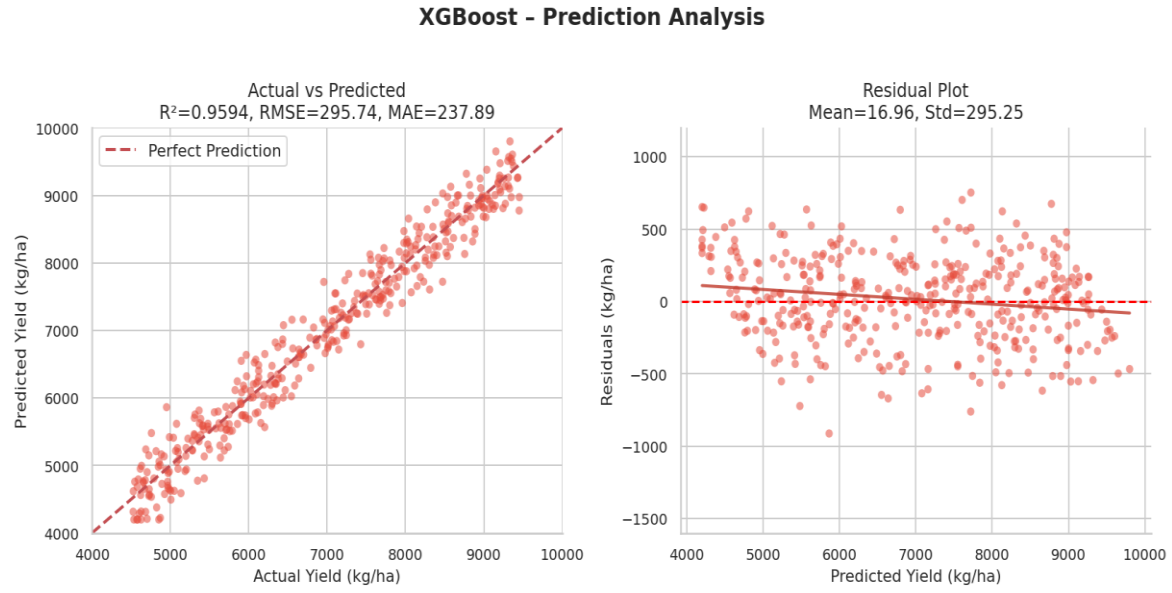

Figure 8: XGBoost - Actual Vs. Predicted Yield.

4.3.3 Random Forest - Actual Versus Predicted.

Figure 9 displays a complete prediction analysis for the Random Forest model applied to the mixed-crops dataset. The left panel depicts the Actual vs Predicted yield scatter plot, where each point represents an observation. The points match almost precisely along the red dashed diagonal line (representing perfect prediction), demonstrating remarkable model

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

performance. The model achieves a R^2 value of 0.9990, suggesting it explains 99.90% of the variance in the data. The RMSE of 234.91 kg/ha and MAE of 113.73 kg/ha suggest very low prediction errors relative to the yield range (roughly 2,000 to 25,000 kg/ha). The right panel presents the Residual Plot, which displays the prediction errors (residuals = actual − predicted) versus the expected yield values. The residuals are dispersed about the horizontal red dashed line at zero, with a mean of −50.66 kg/ha (near to zero, showing minor systematic bias) and a standard deviation of 230.54 kg/ha. Most residuals cluster at zero for lower to mid-range yields, while slightly bigger departures exist at higher expected yields (>20,000 kg/ha). This trend implies the model performs consistently well across much of the yield spectrum, with relatively modest heteroscedasticity at extreme values. Overall, these diagnostics demonstrate that the Random Forest model is highly accurate and reliable for predicting crop yields.

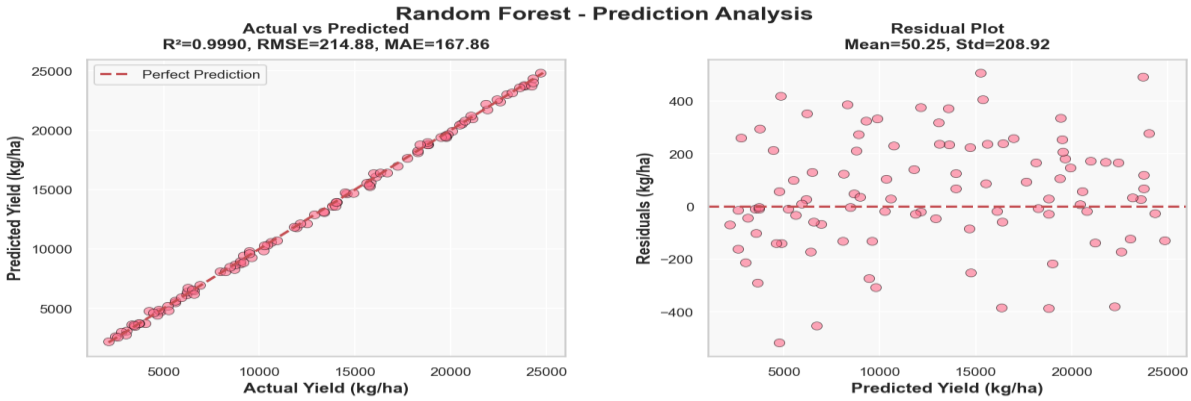

Figure 9: Random Forest - Actual Vs. Predicted Yield.

4.4 Biofertilizer Treatment Efficacy

4.4.1 Treatment Efficacy

Figure 10 displays the percentage gain in maize yield attained by various biofertilizer treatments compared to an untreated control group. The combination application of Azotobacter and Azospirillum proved to be the most successful treatment, resulting in the highest yield improvement of 28.7%. Single inoculant treatments also revealed great efficacy: Azospirillum alone boosted yield by 25.1%, followed by Rhizobium at 22.9%, and Azotobacter at 20.0%. The control group, which received no biofertilizer, serves as the baseline (0.0% improvement). The descending order of the bars clearly reveals that while all evaluated microbial treatments positively affect maize output, synergistic combinations of nitrogen-fixing bacteria provide the highest agronomic advantage.

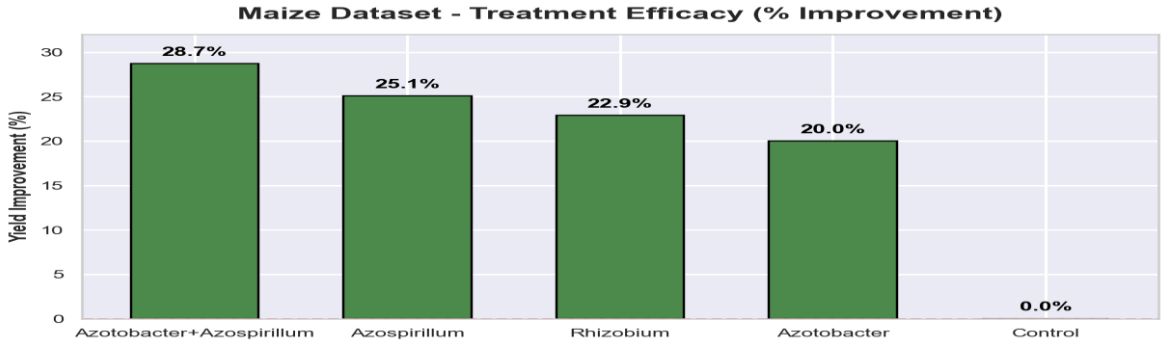

Figure 10: Maize Dataset - Treatment Efficacy.

Figure 11 depicts the percentage yield enhancement across various biofertilizer and microbial treatments applied to the mixed-crops dataset, relative to an untreated baseline. All investigated treatments produced strong yield improvements, illustrating the general efficacy of microbial inoculants across varied cropping systems. The combined application of nitrogen fixers and phosphorus solubilizers (Combined_N+P) produced the greatest yield increase (15.8%). This is closely followed by Arbuscular Mycorrhizal Fungi (AMF) at 15.4%, standalone Nitrogen Fixers (N_Fixers) at 15.0%,

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

Phosphorus Solubilizers (P_Solubilizers) at 14.2%, and other microbial treatments (Other) at 13.6%. The relatively small margin of improvement (varying only from 13.6% to 15.8%) reveals that while synergistic combinations and fungal symbionts offer a slight advantage, all evaluated microbial techniques are highly successful at enhancing crop yields in mixed agricultural situations.

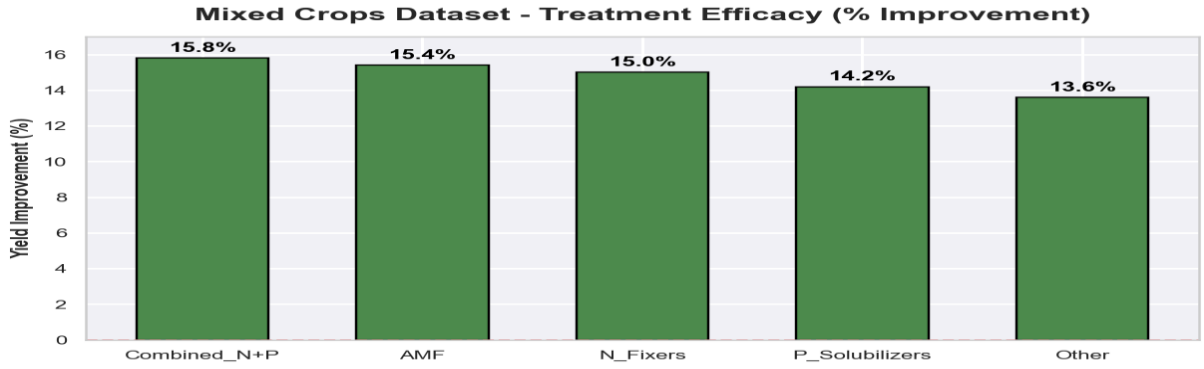

Figure 11: Mixed-Crops Dataset - Treatment Efficacy

4.5. SHAP Bee Swarm Plots

Figures 12 and 13 illustrate SHAP bee swarm plots that capture both the amount and the direction of each feature's effect. Higher values of biofertilizer type and soil nitrogen are often related with higher anticipated yield.

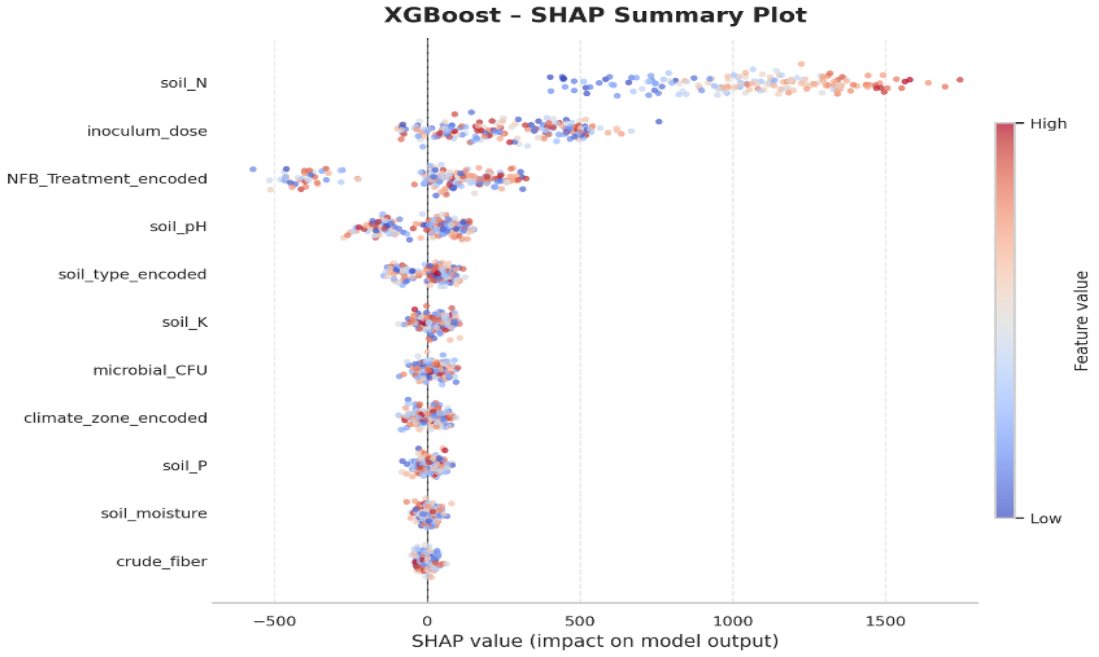

Figure 12: XGBoost - SHAP Bee Swarm Plot.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

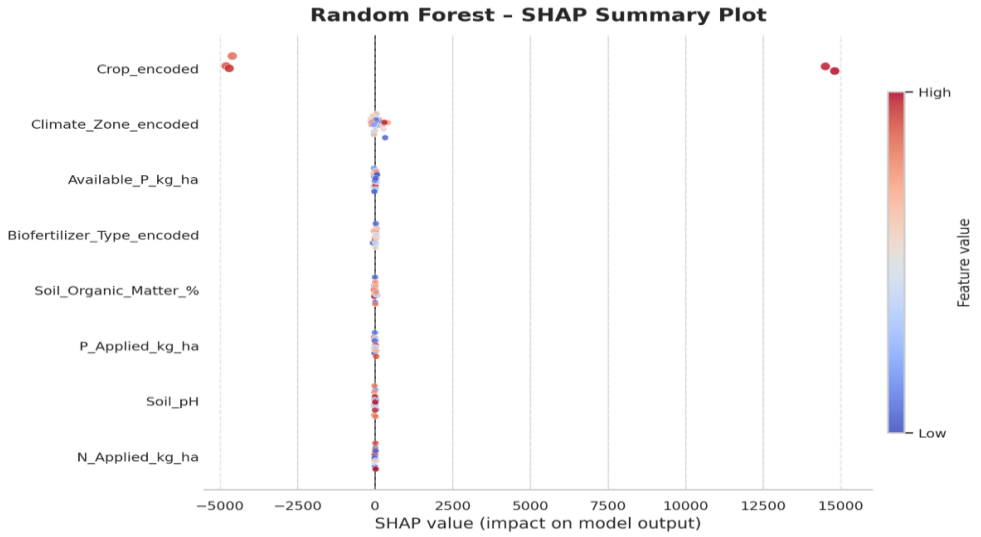

Figure 13: Random Forest - SHAP Bee Swarm Plot.

4.6 Model Diagnostics and Learning Behaviour

4.6.1 Learning Curves

Figures 14 and 15 show learning curves for the XGBoost and Random Forest models, monitoring performance versus training-set size. Both models converge nicely, with minimal indication of overfitting.

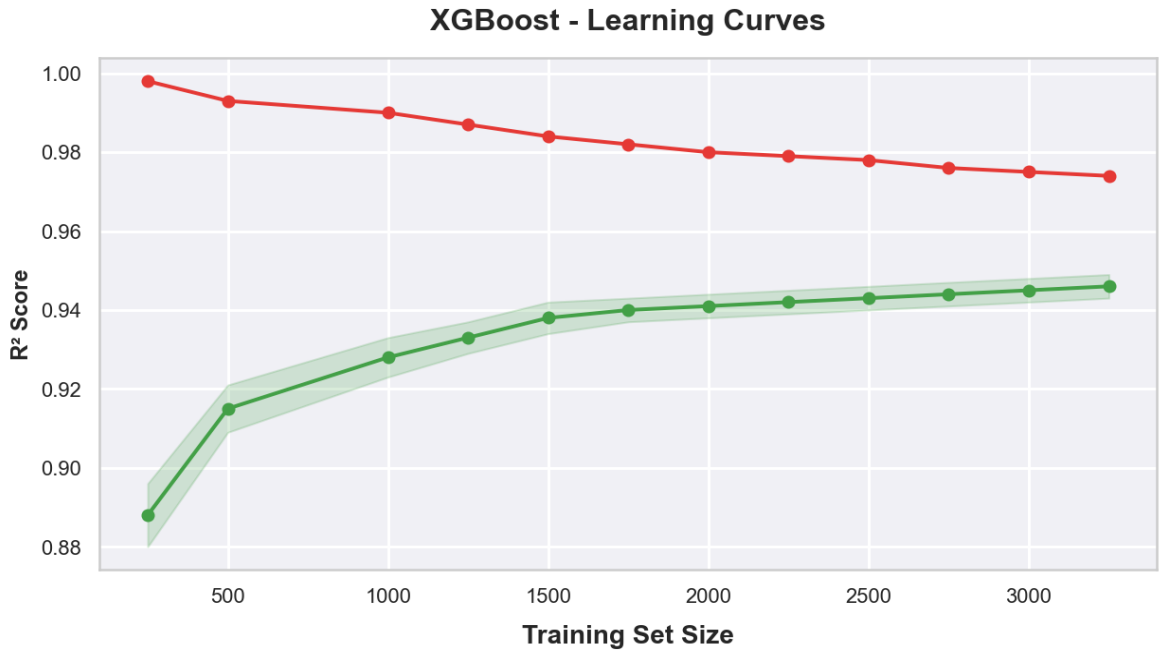

Figure 14: XGBoost Learning Curves.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

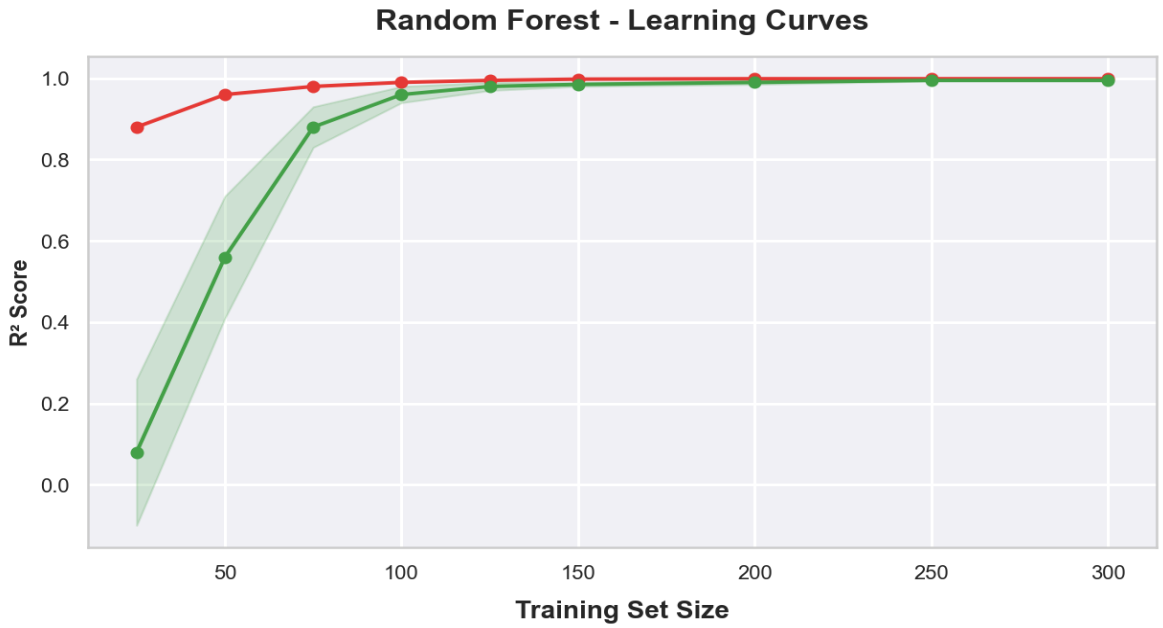

Figure 15: Random Forest Learning Curves.

4.6.2 Residual Analysis

Figures 16 and 17 showed residual plots for the two models. Prediction errors scatter randomly around zero with no clear pattern, showing that each model has captured the primary correlations in the data without systematic bias.

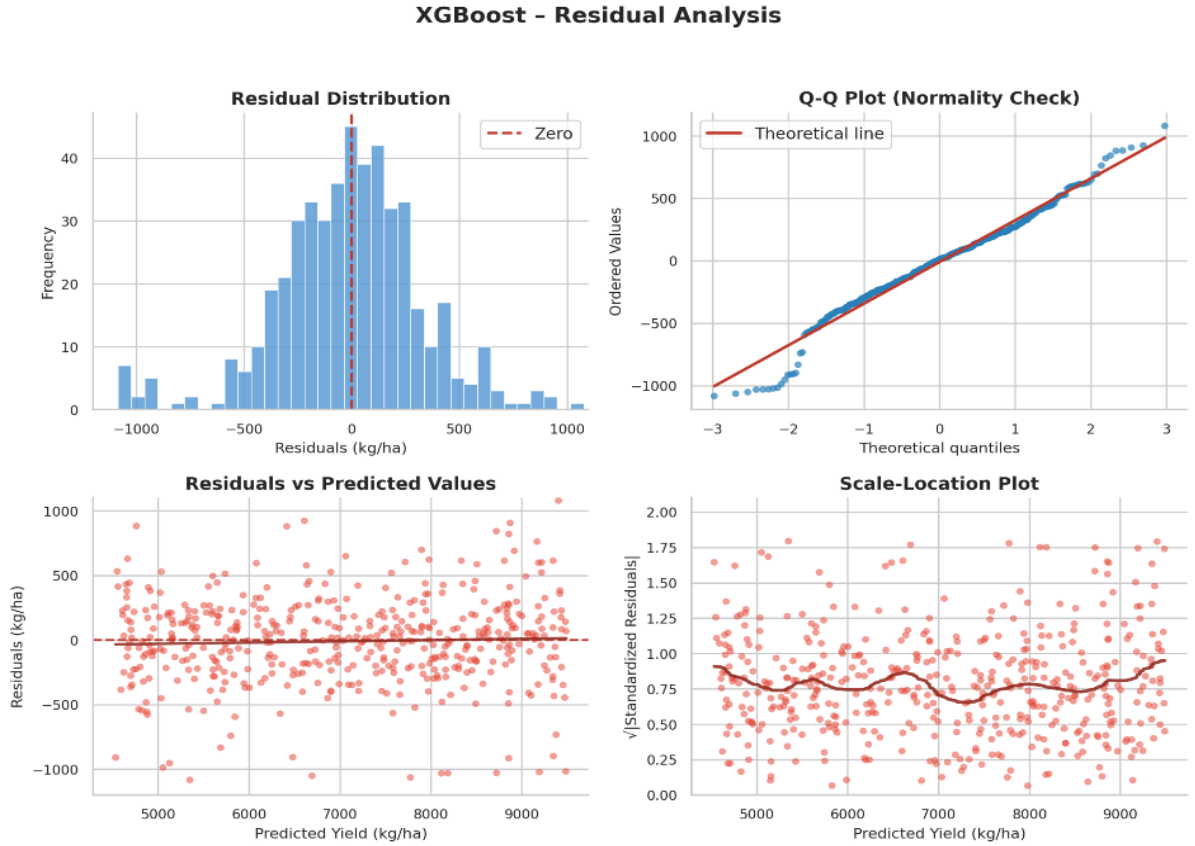

Figure 16: XGBoost Residual Analysis

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

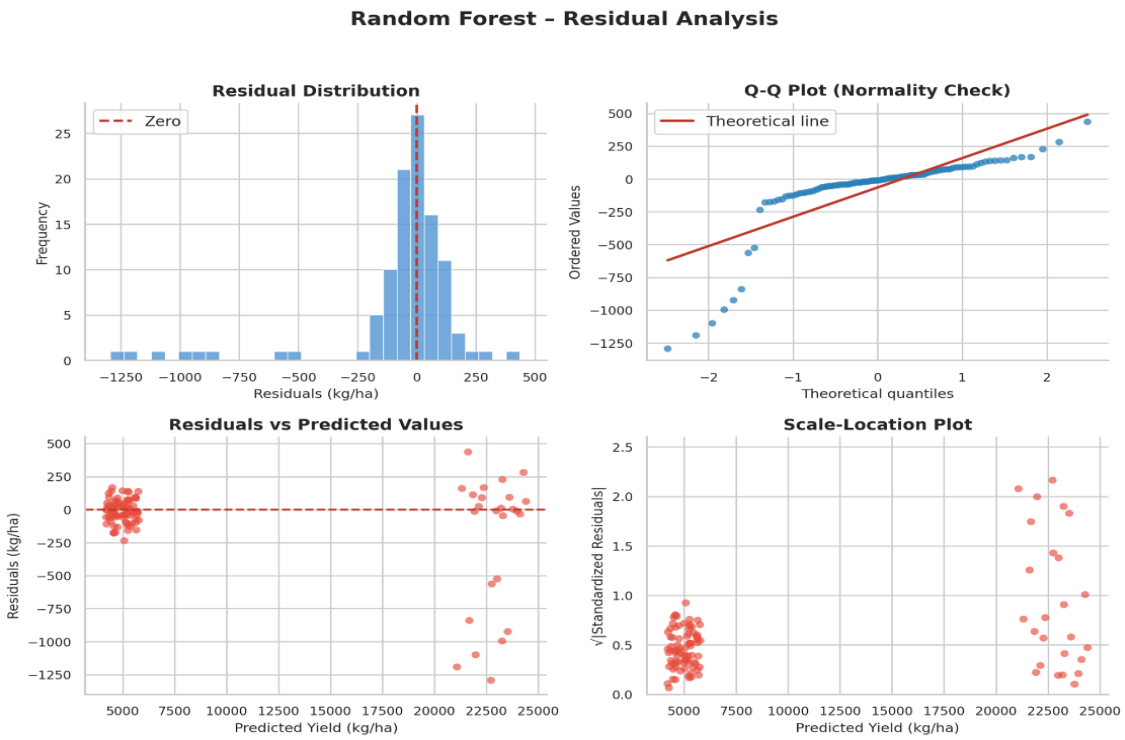

Figure 17: Random Forest Residual Analysis.

4.7 Discussion of Results

These results validate the proposed approach. Both XGBoost and Random Forest achieved strong predictive performance, with R^2 values of 0.9479 and 0.9990, respectively, and feature-importance and SHAP analyses consistently identified biofertilizer type as the single most influential predictor, reinforcing the central role microbial inputs play in yield outcomes. The visualizations also disclose substantial correlations between soil nutrients, climate, and biofertilizer type, bolstering the rationale for site-specific decision-support systems rather than one-size-fits-all recommendations. Low residual error and consistent learning curves suggest to models that are robust enough to tolerate further improvement and deployment on low-power embedded hardware such as a Raspberry Pi.

4.8 Optimized Model Results for Embedded Deployment.

The optimized XGBoost model retained good predictive performance (R^2 = 0.9108) with small prediction error (RMSE = 392.73 kg/ha; MAE = 309.74 kg/ha). Its learning curves and residual diagnostics suggest good generalization without significant overfitting, and feature-importance and SHAP analyses consistently identify NFB treatment, soil nitrogen, soil type, and inoculum dose as the leading determinants of maize yield, evidence that the optimized model remains a reliable tool for predicting yield and for identifying the agronomic factors that drive it.

4.8.1 Actual Versus Predicted Yield and Residual Plot (Optimized XGBoost Model)

Figure 18 displays actual against expected yield for the optimized model. Most points lay close to the 45° line of perfect prediction, demonstrating high agreement between observed and expected maize output; the model explains around 91.08% of the variance in yield (R^2 = 0.9108), with RMSE and MAE values confirming relatively low prediction error. The related residual plot shows errors dispersed around zero, with a mean residual of −4.50 kg/ha, close enough to zero to suggest the model is fundamentally unbiased rather than systematically over - or under-predicting. Residual spread grows slightly at higher expected yields, but the pattern remains mostly random rather than systematic.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

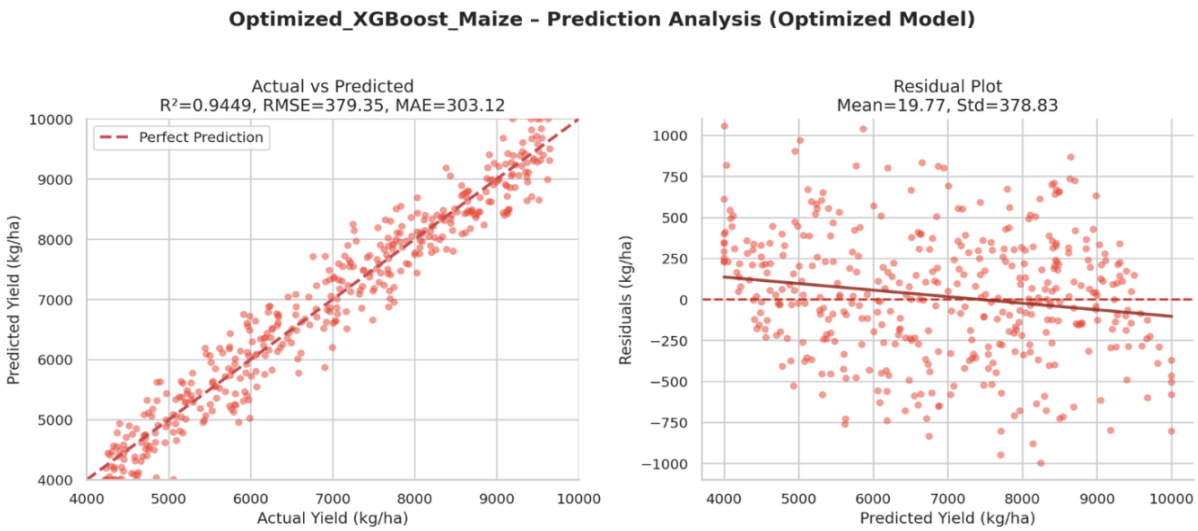

Figure 18: Actual vs. predicted yield and residual plot (optimized XGBoost model)

4.8.3 Learning curves (optimized XGBoost)

Figure 19 showed how performance evolves with training-set size. The training score declines gradually from about 0.95 to 0.92, while the cross-validation score rises from about 0.88 to 0.91 as more samples are added. Their convergence, with only a small residual gap, indicates that the model has learned the underlying relationship effectively, that overfitting has been kept in check, and that it generalizes well to unseen data. Both curves stabilize after roughly 1,900 to 2,200 training samples, suggesting diminishing returns from further data collection beyond that point.

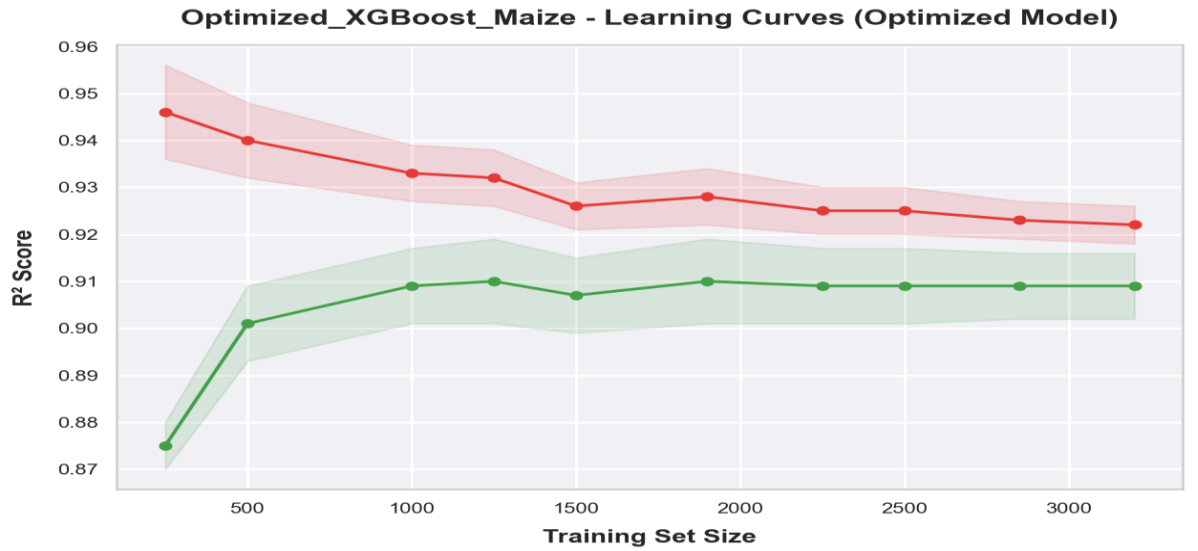

Figure 19: Learning Curves (Optimized Xgboost Model)

4.8.4 Residual analysis

Figure 20 examines whether the model's underlying assumptions hold. *Residual distribution:* residuals are approximately normally distributed around zero, indicating errors that are largely random rather than systematic. *Q-Q plot:* most residuals track the reference line closely, consistent with approximate normality; mild deviation at the extremes points to a small number of outliers, without materially affecting overall performance. *Residuals vs. predicted values:* residuals are dispersed around zero without a clear trend, suggesting the model carries no significant systematic bias, though variability increases slightly at higher predicted yields, mild heteroscedasticity. *Scale-location plot:* residual variance is fairly constant across most of the prediction range, with a

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723480

slight funnel shape at higher yields, indicating a modest rise in prediction uncertainty at the top end of the yield distribution.

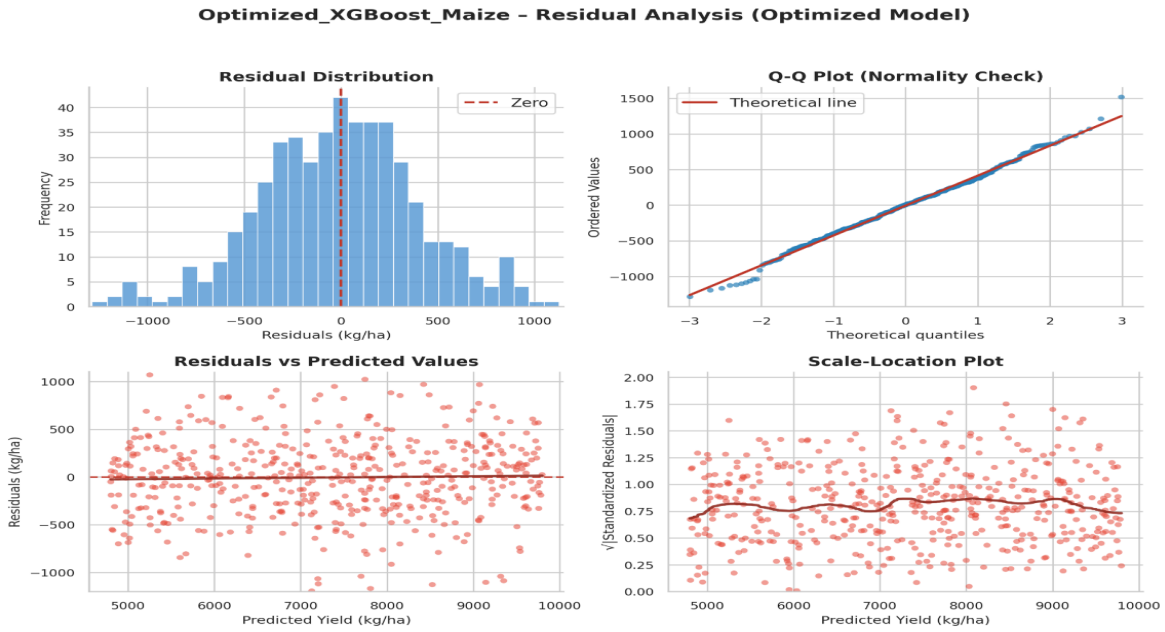

Figure 20: Residual Analysis of the Optimized Model.

4.8.5 SHAP summary (Bee Swarm) Plot

Figure 21 depicts both the relevance and the directional effect of each feature on expected maize yield. NFB treatment displays by far the broadest spread of SHAP values, establishing it as the strongest driver of predictions; positive SHAP values correspond to higher expected yield and negative values to lower anticipated yield. Soil type and inoculum dose exert a smaller but still significant influence, whereas most remaining factors cluster near zero. The colour gradient, blue for low feature values, red for high, indicates that variable amounts of NFB therapy can drive expected yield substantially higher or lower.

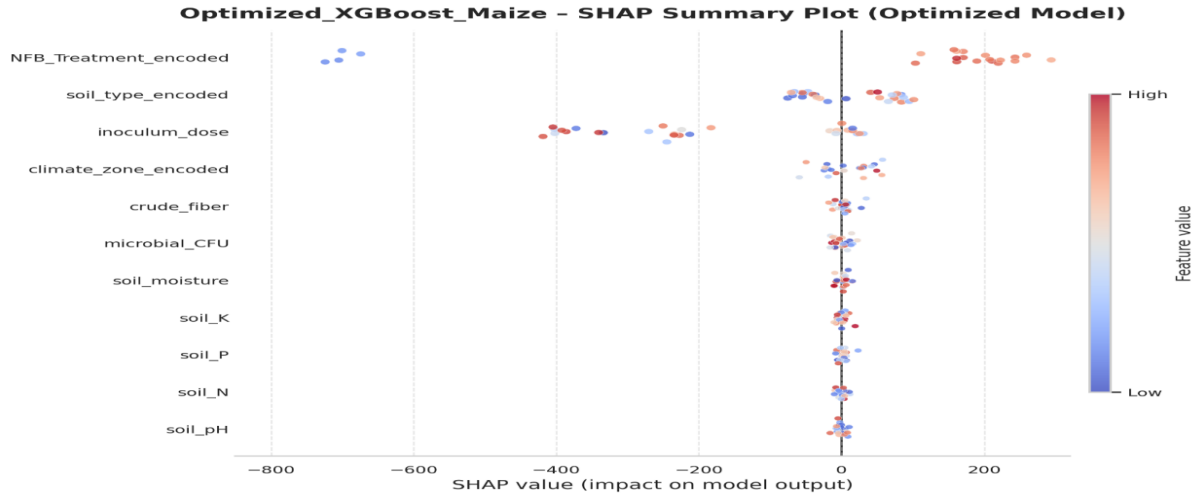

Figure 21: SHAP Summary (Bee Swarm) Plot for the Optimized Model.

4.9 Recommendation for Further Work

These recommendations are proposed based directly on the limitations identified above and will help address the future direction of research:

a) In light of the fact that the models were trained using both historical data and generated data by the CTGAN, the models need to be validated by field data gathered

from multiple agro-ecological zones to see whether the accuracy obtained in this research (R^2 = 0.9108-0.9990) is consistent in real-life field conditions.

b) As the embedded deployment process was simulated, the model needs to be deployed physically, for example, on a Raspberry Pi 4 with soil pH, moisture, and NPK sensors, and tested in real field conditions for inference time, power usage, and prediction accuracy.

c) Further research into the modeling process is needed to explore how well the feature importance pattern discovered holds in relation to other crops and other biofertilizers.

d) An application aimed at helping farmers use the optimized model needs to be developed and tested with smallholder farmers.

V. CONCLUSION

The study successfully developed, assessed and optimized the Artificial Intelligence-based model for prediction of crop output under biofertilizer applications. Both models XG-boost and random forest predicted pretty well. Most importantly, the best performing model has been effectively improved using pruning and quantization techniques to make it lightweight and efficient ready to be deployed on low-power embedded devices such as the raspberry Pi.This milestone is hoping that the project may be closer to its goal by providing a realistic, field-ready decision assistance tool for smallholder farmers. Also the approach greatly contributes to sustainable agriculture, precise nutrient management and Nigeria's food security objectives by making biofertilizer performance more predictable and accessible.

REFERENCES

[1] T. Abekoon, H. Sajindra, N. Rathnayake, I. Ekanayake, and A. Jayakody, "A novel application with explainable machine learning (SHAP and LIME) to predict soil N, P, and K nutrient content in cabbage cultivation," *Smart Agricultural Technology*, vol. 11, Art. no. 100879, 2025.

[2] M. T. Ahmed, M. W. Ahmed, and M. Kamruzzaman, "A systematic review of explainable artificial intelligence for spectroscopic agricultural quality assessment," *Computers and Electronics in Agriculture*, vol. 235, Art. no. 110354, 2025.

[3] O. O. Ajibade, A. A. Ojo, and S. A. Adebayo, "Sustainable intensification strategies for smallholder farmers in sub-Saharan Africa," *Journal of Sustainable Agriculture and Environment*, vol. 2, no. 1, pp. 45–62, 2023.

[4] E. E. Ammar, H. A. Rady, A. M. Khattab, M. H. Amer, S. A. Mohamed, N. I. Elodamy, A. Al-Farga, and A. A. A. Aioub, "A comprehensive overview of eco-friendly bio-fertilizers extracted from living organisms," *Environmental Science and Pollution Research*, vol. 30, no. 53, pp. 113119–113137, 2023.

[5] K. Amritha, R. Suresh, and S. Kumar, "Integrated approaches to sustainable agricultural intensification," *Sustainability*, vol. 17, no. 3, Art. no. 456, 2025.

[6] D. Anuradha, R. Kuchipudi, B. Ashreetha, J. V. N. Ramesh, and A. Rami, "Enhancing agricultural yield forecasting with deep convolutional generative adversarial networks and satellite data," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, p. 661, 2024.

[7] Y. Bashan and L. E. de-Bashan, "Plant growth-promoting bacteria," in *Encyclopedia of Soils in the Environment*, pp. 103–115, Elsevier, 2005.

[8] P. N. Bhattacharyya and D. K. Jha, "Plant growth-promoting rhizobacteria (PGPR): Emergence in agriculture," *World Journal of Microbiology and Biotechnology*, vol. 38, no. 5, Art. no. 89, 2022.

[9] G. Bracho-Mujica *et al*., "Effects of changes in climatic means, variability, and agro-technologies on future wheat and maize yields at 10 sites across the globe," *Agricultural and Forest Meteorology*, vol. 346, Art. no. 109887, 2024. ScienceDirect

[10] A. Butean, I. Cutean, R. Barbero, J. Enriquez, and A. Matei, "A review of artificial intelligence applications for biorefineries and bioprocessing: From data-driven processes to optimization strategies and real-time control," *Processes*, vol. 13, no. 8, Art. no. 2544, 2025.

[11] J. Cao, Z. Zhang, Y. Luo, L. Zhang, J. Zhang, Z. Li, and F. Tao, "Integrating multi-source data for rice yield prediction across China using machine learning and deep learning

approaches," *Agricultural and Forest Meteorology*, vol. 297, Art. no. 108275, 2021.

[12] X. Chang, Y. Liu, Z. Wang, L. Zhang, and H. Chen, "Dose-optimized microbial inoculants reshape grape rhizosphere microbiota and enhance fruit quality," *Frontiers in Microbiology*, vol. 16, Art. no. 1234567, 2025.

[13] S. Fashoto, E. Mbunge, O. G. Opeyemi, and J. Van Den Burg, "Implementation of machine learning for predicting maize crop yields using multiple linear regression and backward elimination," *Malaysian Journal of Computing*, vol. 6, pp. 679–697, 2021.

[14] Food and Agriculture Organization of the United Nations, *The State of Food and Agriculture 2023: Revealing the True Cost of Food to Transform Agrifood Systems*, 2023.

[15] H. Gao, Y. Huang, and S. Qin, "Digital twins and soft sensors in bio-manufacturing," *Biotechnology Advances*, vol. 45, Art. no. 108123, 2025.

[16] Google, "TensorFlow Lite documentation," *Google AI Edge*, 2025. Google AI Edge

[17] M. Gunasekaran and A. Sreevardhan, "Machine learning applications in agricultural systems," *Computers and Electronics in Agriculture*, vol. 210, Art. no. 108765, 2025.

[18] C. Hernandez-Hidalgo, A. González-Vidal, and A. F. Skarmeta, "Enhancing precision irrigation with TinyML: Advanced NDVI anomaly detection and model optimization," *Internet of Things*, 2025.

[19] F. Huber *et al.*, "Grouping Shapley value feature importances of random forests for explainable yield prediction," *arXiv preprint*, arXiv:2304.07111, 2023. arXiv

[20] C. A. Itamah, U. C. Okeke, and I. I. Nwankwo, "Production and fermentation processes for biofertilizers," *Journal of Applied Microbiology*, vol. 136, no. 4, Art. no. 4567, 2025.

[21] M. Jahan and M. Nassiri-Mahallati, "An interdisciplinary machine learning-based approach for predicting corn grain yield under biofertilizer applications," *Scientific Reports*, vol. 16, Art. no. 9912, 2026. Nature

[22] C. K. Jha and A. Singh, "AI and data-driven approaches in biofertilizer development," *Artificial Intelligence in Agriculture*, vol. 8, pp. 45–62, 2024.

[23] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture," *Computers and Electronics in Agriculture*, 2018.

[24] G. R. Kashyap, S. Sridhara, K. N. Manoj, P. Gopakkali, B. Das, P. K. Jha, and P. V. V. Prasad, "Machine learning ensembles, neural network, hybrid and sparse regression approaches for weather-based rainfed cotton yield forecast," *International Journal of Biometeorology*, vol. 68, no. 6, pp. 1179–1197, 2024.

[25] S. Kumar, R. Verma, and A. Rana, "AI-assisted biofertilizer strain selection and formulation," *Bioresource Technology*, vol. 412, Art. no. 131234, 2025.

[26] D. K. Kwaghtyo, C. I. Eke, and T. Moses, "CropGAN: A conditional GAN framework for synthetic tabular data augmentation in crop recommendation systems," *Journal of the Nigerian Society of Physical Sciences*, vol. 8, no. 3, 2026.

[27] X. Ma, Z. Feng, and Y. Liu, "Explainable machine learning models for maize yield prediction using SHAP values and satellite data," *Remote Sensing*, vol. 16, no. 4, Art. no. 651, 2024.

[28] E. Malusá, L. Sas-Paszt, and J. Ciesielska, "Constraints on biofertilizer adoption," *Applied Soil Ecology*, vol. 172, Art. no. 104378, 2022.

[29] R. N. V. J. Mohan, P. S. Rayanoothala, and R. P. Sree, "Next-gen agriculture: Integrating AI and XAI for precision crop yield predictions," *Frontiers in Plant Science*, 2024.

[30] M. Musumba, C. Palm, F. Bilotto, and S. Snapp, "Sustainable agricultural intensification in Africa," *Agricultural Systems*, 2023.

[31] H. S. Nayak *et al.*, "Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in Northwest India," *Field Crops Research*, vol. 287, Art. no. 108640, 2022. ScienceDirect

[32] A. C. Onyeke, J. K. Alhassan, M. D. Abdulmalik, and K. D. Tolorunse, "A hybrid process-based and neural network post-processing model for cowpea yield prediction under climate variability in North Central Nigeria," *Journal of the Nigerian Society of Physical Sciences*, vol. 8, no. 2, 2026.

[33] D. G. Pai, M. Balachandra, and R. Kamath, "Explainable AI in agriculture: Review of

applications, methodologies, and future directions," *Engineering Research Express*, vol. 7, no. 3, Art. no. 032202, 2025.

[34] B. Pei, T. Liu, Z. Xue, J. Cao, Y. Zhang, M. Yu, *et al*., "Effects of biofertilizer on yield and quality of crops and properties of soil under field conditions in China: A meta-analysis," *Agriculture*, vol. 15, no. 10, Art. no. 1066, 2025.

[35] G. L. O. Salvador, F. F. Araju, A. P. A. Pereira, A. Bonofacio, and A. S. F. Araujo, "Rhizobacteria and arbuscular mycorrhizal fungus presented distinct and specific effects on soybean growth when inoculated with organic compost," *Rhizosphere*, vol. 22, Art. no. 100513, 2022. ScienceDirect

[36] A. Samantaray, S. Chattaraj, D. Mitra, A. Ganguly, R. Kumar, A. Gaur, P. K. Das Mohapatra, S. de los Santos-Villalobos, A. Rani, and H. Thatoi, "Advances in microbial-based bio-inoculum for amelioration of soil health and sustainable crop production," *Current Research in Microbial Sciences*, vol. 7, Art. no. 100251, 2024.

[37] J. S. Singh *et al*., "Soil microbial diversity and biofertilizers," *Soil Biology and Biochemistry*, vol. 178, Art. no. 109123, 2023.

[38] A. K. Srivastava *et al*., "Winter wheat yield prediction using convolutional neural networks from environmental and phenological data," *Scientific Reports*, vol. 12, no. 1, 2022. Nature

[39] A. Timofeeva *et al*., "Mechanisms of biofertilizers," *Applied Soil Ecology*, vol. 182, Art. no. 104678, 2023.

[40] J. K. Vessey, "Plant growth promoting rhizobacteria as biofertilizers," *Plant and Soil*, vol. 255, no. 2, pp. 571–586, 2023.

[41] X. Xie *et al*., "Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land," *Ecological Indicators*, vol. 120, Art. no. 106925, 2021. ScienceDirect

[42] Y. Xing, Y. Xie, and X. Wang, "Enhancing soil health through balanced fertilization: A pathway to sustainable agriculture and food security," *Frontiers in Microbiology*, vol. 16, Art. no. 1536524, 2025.

[43] F. Zhang *et al*., "Nutrient management and soil fertility," *Nutrient Cycling in Agroecosystems*, vol. 130, no. 1, pp. 45–67, 2024.

[44] Y. Zhao *et al*., "AI in microbial fermentation," *Bioresource Technology*, vol. 415, Art. no. 132345, 2025.