Advanced and Sustainable Technologies
ASET
E-ISSN 2976-2294

Volume 4, No 2, December 2025 [385-402]

Analysis of NGBoost and XGBoost for Added Amount Fertilizer (NPK) Prediction: Uncertainty Estimation and Model Performance in Precision Agriculture

Erdy Sulino bin Mohd Muslim Tan[1]*, Marni Azira Binti Markom[2], Abu Hassan Abdullah[1], Norasmadi Abdul Rahim[1], Lee Yit Leng[3], Fathinul Syahir Ahmad Saad[1], Allan Melvin Andrew[1], Imaduddin Helmi Wan Nordin[3], Pubalan Nadaraja[1], and Mohd Amri Zainol Abidin[3]

[1]Faculty of Electrical Engineering and Technology, Universiti Malaysia Perlis, Malaysia.
[2]Faculty of Electronic Engineering and Technology, Universiti Malaysia Perlis, Malaysia.
[3]Faculty of Mechanical Engineering and Technology, Universiti Malaysia Perlis, Malaysia.

Received 31 July 2025, Revised 31 August 2025, Accepted 12 September 2025

ABSTRACT

Optimizing fertilizer application is essential for enhancing crop yield and minimizing the environmental impact of precision agriculture. This study presents a comparative analysis of XGBoost and NGBoost for predicting the amount of NPK fertilizer added, focusing on model performance and uncertainty estimation. The dataset collected from the Harumanis mango orchards included key soil parameters, such as nitrogen, phosphorus, potassium, pH, EC, soil moisture, temperature, and rainfall. The methodology involved data preprocessing, feature scaling, and model training using XGBoost and NGBoost. XGBoost, a gradient boosting model, provides highly accurate deterministic predictions, whereas NGBoost, a probabilistic model, quantifies the uncertainty in the predictions. Model performance was evaluated using R^2, MAE, RMSE, and Negative Log-Likelihood (NLL). The results indicate that XGBoost outperforms NGBoost in accuracy, achieving $R^2 = 0.9984$, MAE = 2.0388, and RMSE = 2.7618, whereas NGBoost provides uncertainty estimation but with slightly lower accuracy ($R^2 = 0.9909$, MAE = 4.9620, RMSE = 6.5654, and NLL = 2.6001). Further analysis included residual plots, prediction error plots, learning curves, and validation curves to assess the reliability and generalization of the models. These findings suggest that, while XGBoost is ideal for deterministic NPK prediction, NGBoost offers probabilistic insights that aid in the development of risk-aware fertilization strategies. This study contributes to data-driven precision agriculture by enhancing fertilizer management efficiency and sustainability.

Keywords: Fertilizer prediction, XGBoost, NGBoost, Precision agriculture, Soil nutrients.

1. INTRODUCTION

Precision agriculture plays a transformative role in enhancing fertilization management by integrating advanced technologies to optimize both nutrient use efficiency and environmental sustainability. At its core, precision agriculture employs technologies like proximal and remote sensor surveys to monitor variations within fields, allowing for the targeted application of fertilizers based on site-specific conditions. This approach is crucial in matching nutrient application to the precise needs of crops at specific locations, minimizing the risk of nutrient leaching and accumulation, which can lead to environmental issues [1].

The practice of precision agriculture involves using advanced technologies such as GPS, GIS, and variable rate technology (VRT) to ensure accurate fertilizer application. These technologies

*Corresponding author: erdysulino@unimap.edu.my enable precise mapping and management of field variability, resulting in improved nutrient use efficiency. By applying the correct amount of nutrients where they are needed most, this approach not only enhances crop production but also mitigates the excessive use of fertilizers that can harm the environment [2, 3].

The adoption of smart sensors and IoT in precision agriculture offers real-time monitoring and data-driven decision-making capabilities. This integration allows farmers to make informed decisions about nutrient application, thereby improving nutrient use efficiency (NUE) and reducing waste. Real-time nitrogen sensing, in particular, enables precise nitrogen management, which is pivotal in promoting sustainable agriculture and reducing the environmental footprint of farming practices [4, 5]. The application of precision agriculture is not without challenges. The effective implementation of these technologies requires investments in equipment and training, as well as overcoming data management issues. Despite these challenges, the strategic use of precision agriculture technologies holds significant promise for improving agricultural productivity while maintaining ecological balance [6].

Gradient boosting techniques, particularly XGBoost, have gained significant attention in various prediction tasks due to their high accuracy and robustness. However, as deterministic models, they lack the ability to provide uncertainty estimates, which can be a limitation in contexts like agronomy where decision-making under uncertainty is crucial. This is where NGBoost, a natural gradient boosting technique, shows promise by offering probabilistic predictions.

XGBoost has been employed successfully in several studies. For instance, it outperformed other models in predicting shear strengths of rockfill materials by achieving the highest prediction performance metrics, indicating its reliability in various engineering applications [7]. Similarly, it has been used effectively in estimating urban water levels and ice phenomena prediction due to its robustness and ability to utilize large datasets efficiently [8, 9]. Its integration with other techniques, like evolutionary algorithms, further enhances its predictive accuracy in complex tasks [9].

NGBoost offers a different approach by directly producing probabilistic predictions, which allow for estimating prediction uncertainties. This feature makes NGBoost particularly suitable for applications where understanding the uncertainty of predictions is as important as the predictions themselves. A study on predicting the California bearing ratio (CBR) values demonstrated NGBoost's ability to provide reliable confidence intervals, thereby offering a comprehensive view of prediction reliability [10]. Furthermore, its application in structural engineering showed it could achieve comparable mean prediction accuracy levels to other machine learning algorithms while providing robust uncertainty estimates [11].

XGBoost remains a powerful tool for deterministic predictions with high accuracy, while NGBoost significantly contributes to areas where capturing the uncertainty of predictions is critical. This makes NGBoost a valuable asset in fields like agronomy, where decision-making must often contend with variable environmental conditions and inherent uncertainties.

This study aimed to compare the performance of XGBoost and NGBoost in predicting the amount of NPK fertilizer added using soil parameters such as pH, EC, moisture, temperature, and rainfall. The evaluation includes traditional performance metrics (R^2, MAE, RMSE) and uncertainty quantification through Negative Log-Likelihood (NLL). By integrating explainable AI techniques and visualization tools, this study provides insights into model reliability, uncertainty estimation, and practical implications for precision agriculture.

Research Objectives:

1. To evaluate the predictive performance of XGBoost and NGBoost for added fertilizer prediction.

2. To analyze uncertainty estimation using NGBoost in the context of soil nutrient variability.

A comprehensive model assessment was performed using visualization tools, such as residual plots, prediction error plots, learning curves, and validation curves. The findings of this study will help farmers and agronomists make informed fertilization decisions, thereby reducing costs and minimizing the environmental impact. Furthermore, it contributes to data-driven precision agriculture by integrating machine learning and uncertainty-aware models for sustainable fertilizer management.

2. Methodology

2.1 Data Collection

Soil nutrient data were obtained from two primary sources: the Harumanis tree Orchard A data collected at Orchard A (Muzium Mempelam, Kuala Sala, Department of Agriculture Kedah (5°58’05.5 “N 100°24’05.1” E)) and data collected from an Orchard B (Individual Orchard at Guar Nangka in Perlis (6°28’34.4 “N 100°17’05.7” E)). This is followed by a comprehensive presentation of the laboratory test data, which provides analysis results that complement the sensor data and ensure the reliability and accuracy of the findings. Data collection at Muzium Mempelam, Kuala Sala, and Kedah for this study involved monitoring 24 Harumanis trees distributed across six sampling points within an area of 14,850 square feet. Sampling was conducted daily from 7:00 AM to 3:00 PM, resulting in 30,816 data samples. Each day, eight data samples were collected, with one data sample being the average of six subsamples taken at five-minute intervals. This intensive sampling regimen ensured comprehensive data coverage of the nutritional status and environmental conditions of the trees.

2.2 Data Preprocessing

Data preprocessing is a critical step in preparing datasets for machine learning models and involves several key techniques, including handling missing data, detecting and removing outliers, normalization and scaling, and feature engineering.

Missing data can significantly affect the performance of machine learning models. Various imputation techniques are employed to address this issue, such as random sample (RS) imputation and one-hot encoding methods. These techniques fill in the gaps in datasets to maintain the integrity of the model. For instance, in scenarios with high missing rates, one-hot encoding has proven to be effective, demonstrating robustness and accuracy [12]. Another approach is the cyclical hybrid imputation technique, which combines row-based and column-based imputation techniques to handle missing data effectively. This method has shown promise in increasing model accuracy when tested on multiple datasets [13].

Outliers can be identified using various methods, such as sliding window techniques, which detect anomalies by identifying values that fall outside a normal pattern distribution [14]. Another novel approach for outlier detection involves using Gaussian mixture models for cellwise detection, which allows for the identification and imputation of contaminated cells rather than discarding them, thereby preserving valuable information [15].

Normalization and scaling are vital for improving the convergence and stability of machine learning models. The choice between standardization and normalization can depend on the dataset size and the machine learning algorithm used. For instance, normalization has been shown to enhance the performance of linear models on smaller datasets, whereas standardization is more beneficial for larger datasets with linear models [16]. Moreover, robust scaling techniques, such as LSBZM normalization, specifically tailored for time series data, can address issues like skewness by applying a combination of transformations to ensure uniform scaling [17].

Feature engineering involves creating new variables that enhance the predictive power of a model. This can include deriving additional variables like NPK decay over time or cumulative rainfall impact to provide the model with more informative features. The systematic review of feature selection methods suggests using advanced techniques like metaheuristics and hyper-heuristics to improve the model's performance by efficiently reducing the dimensionality of the dataset and eliminating redundancy [18].

Each of these preprocessing steps enhances the model's ability to learn from the data effectively by addressing issues related to data integrity, scale, and feature relevance, ultimately leading to more accurate and reliable predictions.

2.3 Target Variable

The target variable, also known as the dependent or response variable, is the variable that the ML model aims to predict or explain based on input features. In the context of NPK prediction for Harumanis Mango cultivation, the target variable is typically the amount of fertilizer (NPK) nutrients added to the soil.

The target variable is essential in training the ML model, as it serves as the benchmark against which the model's predictions are evaluated. The model learns from the input features to predict the target variable. For example, in a regression problem, the target variable may be the actual measured levels of NPK nutrients in the soil, and the model aims to predict these levels based on the input features.

It is important to choose the target variable carefully to ensure that it accurately represents the phenomenon of interest and is measurable and relevant to the research objectives of this study. In soil NPK prediction, the target variable should ideally reflect the actual levels of NPK nutrients in the soil, which is crucial for effective crop management and agricultural decision-making.
Using this methodology, a systematic approach was presented to predict and manage the levels of NPK in the soil over a 14-day period [12]. This process involved measuring the initial NPK levels in the soil, denoted as (A), and then using decay equations to estimate the amount of each nutrient remaining after 14 days.

The decay of each nutrient was modelled using the following equation:

$$[N_{14} = N_{\text{initial}} \times e^{-k_N \times 14}] \tag{1}$$
$$[P_{14} = P_{\text{initial}} \times e^{-k_P \times 14}] \tag{2}$$
$$[K_{14} = K_{\text{initial}} \times e^{-k_K \times 14}] \tag{3}$$

where Equations (1), (2), and (3) represent the amounts of NPK remaining after 14 days, respectively. (k_N), (k_P), and (k_K)are the decay constants for each nutrient.

The total target amount of NPK after 14 days, denoted as (B) equation (4), is the sum of the remaining amounts of each nutrient:

$$[B = N_{14} + P_{14} + K_{14}] \tag{4}$$

To achieve the desired nutrient levels in the soil, an ML model was employed to predict the necessary amount of NPK to be added, referred to as (C). This prediction ensured that, after 14

days, the soil retained its target nutrient levels. The Critical Soil Test Value (CSTV) was set to 140 ppm (an example) and served as a benchmark for optimal soil fertility.
The target variable for this methodology is defined as:

$$[\text{Target Variable} = \text{CSTV} - (A - B)] \tag{5}$$

Equation (5) accounts for the difference between the CSTV and the actual change in NPK levels, considering both the initial value (A) and the target amount (B) after 14 days.

By following this method, precise adjustments can be made to soil nutrient levels to ensure that they remain within the optimal range over time. This approach combines decay equations and ML predictions to effectively manage soil fertility.

Table 1 presents the methodology for determining the amount of fertilizer required to replenish N, P, and K in the soil, ensuring optimal nutrient levels for the growth of Harumanis mango trees. The process followed a systematic approach, considering soil nutrient readings, nutrient decay over a period of 14 days, and CSTV using the MALCC to calculate the fertilizer requirement.

Table 1: Example calculations to get the total added amount of NPK.

	N	**P**	**K**
(A) Amount Available in Soil (mg/kg) Reading from TDR Sensor	80	90	100
(B) Decay (14 days)	19.7	22.2	24.7
Amount after 14 days (A)-(B)	60.3	67.8	75.3
Fertilizer YaraMila 13-13-21	N	P2O5	K20
MALCC CSTV	190	150	200
CSTV - (A-B)	90.3	37.8	75.3
mgkg	90.27	86.61	90.76
Total Amount (mgkg)	267.65		
Total added Amount Fertilizer (gram)	569.46		

2.4 Machine Learning Model

The machine learning model architecture, utilized in this study, is based on an ensemble learning technique known as Gradient Boosting Decision Trees. This model is particularly suitable for regression tasks, such as predicting the added amount of N, P, and K fertilizers based on soil and environmental parameters.

Figure 1 illustrates the proposed method for the overall workflow of the gradient boosting model, which begins with input data collected from sensors, including variables such as soil temperature, moisture, pH, electrical conductivity, and rainfall. These raw input features are passed through a preprocessing phase, which includes several critical steps: data cleaning, feature scaling, encoding categorical values, and splitting the dataset into training and testing sets.

Once the data is preprocessed, it is fed into the boosting model composed of multiple decision trees. The model is trained sequentially. Tree 1 learns from the original training data, and Tree 2 is trained to correct the errors made by Tree 1 by learning the residuals. Tree K continues this process iteratively, minimizing prediction errors at each step. Each decision tree in the ensemble contributes to the final prediction, and their outputs are combined to produce a more accurate and generalized result. The model learns from the prediction errors at each stage, adjusting and improving subsequent tree structures. Once training is complete, the model is used to make predictions on the test data.

The output from the model provides an estimated value for the added amount of NPK fertilizer needed, helping to support precision agriculture practices by optimizing nutrient management in crop cultivation. This workflow not only ensures improved prediction accuracy through iterative learning but also enables the model to handle complex, nonlinear relationships between soil parameters and nutrient requirements efficiently.

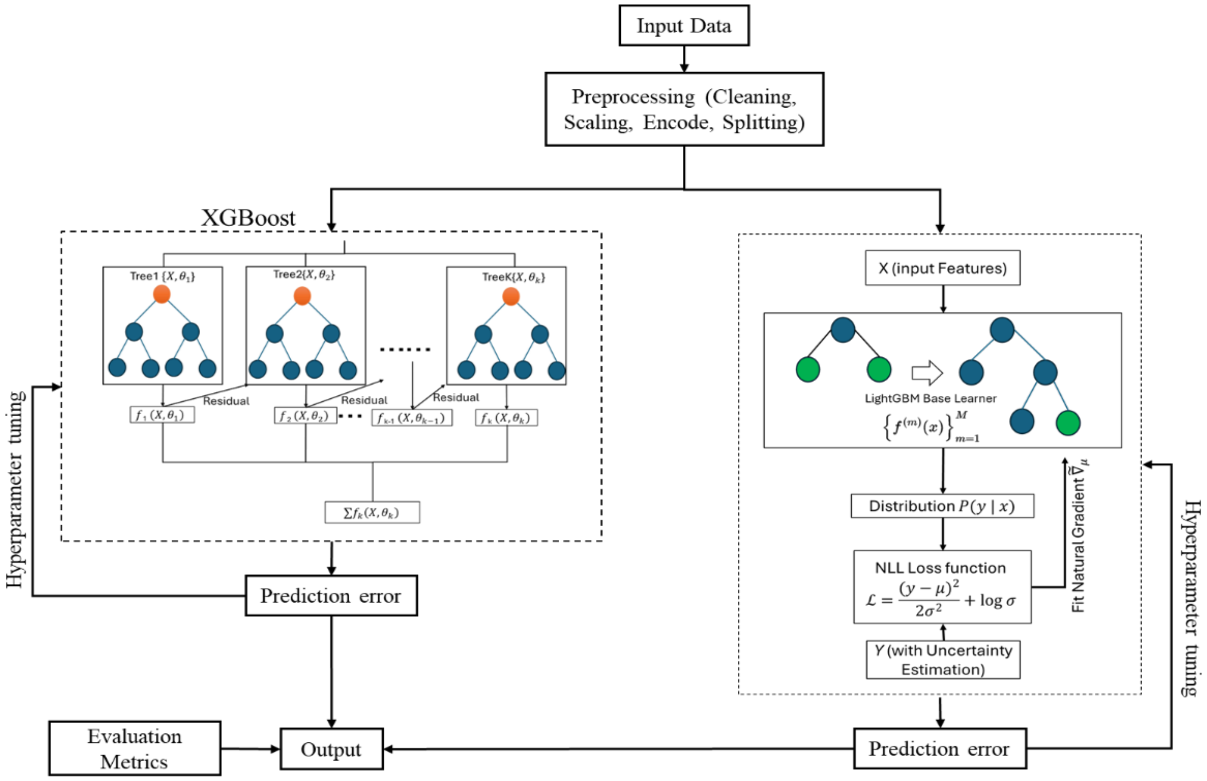

Figure 1: Proposed Architecture of a Gradient Boosting Decision Tree Model Workflow. The model uses an ensemble of decision trees trained sequentially to minimize prediction error, where each subsequent tree corrects the residuals from previous predictions. Input data undergoes preprocessing before being passed into the model, followed by output generation and evaluation.

2.4.1 XGBoost Model Algorithm for Soil NPK Prediction

XGBoost is an ensemble learning algorithm that builds multiple decision trees to optimize prediction accuracy. The model was selected because of its ability to handle nonlinear relationships in soil nutrient variations, reduce overfitting through regularization techniques and optimize computational efficiency using parallel processing.

XGBoost is a gradient boosting algorithm that constructs an ensemble of decision trees by minimizing the MSE while incorporating regularization to prevent overfitting.

a) Objective Function

XGBoost minimizes the following function:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \qquad (6)$$

where:

- y_i is the actual NPK value obtained from the soil sensors.
- $\hat{y}_i$ is the predicted NPK value obtained using XGBoost.
- $\Omega(f_t)$ is a regularization term that penalizes complex trees.

b) Gradient Boosting Process

Each new tree fits the residual errors from the previous trees as follows:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{7}$$

At each iteration, XGBoost updates its predictions by minimizing the second-order Taylor approximation as follows:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{8}$$

where:

- $g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}$ (Gradient).
- $h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$ (Hessian).

This allows for faster convergence compared to traditional gradient boosting.

c) Decision Tree Splitting

Each tree is split based on a gain function that determines the best feature split as follows:

$$\text{Gain } = \frac{1}{2} \left[\frac{(\sum g_i)^2}{\sum h_i + \lambda} + \frac{(\sum g_j)^2}{\sum h_j + \lambda} - \frac{(\sum g_p)^2}{\sum h_p + \lambda} \right] - \gamma \tag{9}$$

where:

- λ is the L2 regularization term used to control overfitting.
- where γ is the threshold for pruning unnecessary splits.
- $g_i, g_j, g_p$and h_i, h_j, h_p represent the gradients and Hessians of the parent and child nodes.

d) Final Prediction

The final NPK prediction is obtained by summing the contributions from all trees:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i) \tag{10}$$

Where T is the total number of trees.

XGBoost is effective for predicting soil NPK levels with high accuracy and robust performance. The hyperparameters of the XGBoost model were optimized using grid search cross-validation, focusing on parameters such as the learning rate, maximum depth, and number of estimators to achieve the best performance.

2.4.2 NGBoost Model

NGBoost extends traditional boosting methods by incorporating probabilistic predictions and allowing uncertainty quantification. This model was chosen to assess the confidence levels in soil nutrient predictions, which is crucial for decision-making in fertilization management in orchards. NGBoost, or Natural Gradient Boosting, enhances conventional boosting methods by focusing on probabilistic predictions and enabling uncertainty quantification. This approach is particularly advantageous in the context of soil nutrient prediction for fertilization management, as it allows for the assessment of confidence levels in the predictions made, thereby aiding decision-making processes in agricultural practices.

In agriculture, precise nutrient management is critical for optimizing crop yields and minimizing environmental impacts. By employing models like NGBoost, which can provide uncertainty estimates along with predictions, farmers and agricultural scientists can make more informed decisions about fertilization strategies [18, 19]. For instance, in orchard management, understanding variations and uncertainties in soil nutrient content can guide targeted fertilization, reducing the risks of over-fertilization and under-fertilization. This approach also supports sustainable agricultural practices by optimizing nutrient application rates based on accurate predictions and uncertainty assessments [20, 21].

NGBoost enhances traditional approaches by incorporating a probabilistic framework that is beneficial for capturing the complex interactions in nutrient dynamics across different soil and crop conditions. This allows for the development of precise fertilization plans tailored to specific environmental and soil variability, ultimately improving crop yields and reducing environmental degradation [22, 23]. Define the parametric distribution, instead of predicting a single value $y\hat{}\_i$, NGBoost models the distribution of y using a parametric distribution family. For example, using a Gaussian:

$$y \sim \mathcal{N}(\mu(x), \sigma^2(x)) \tag{11}$$

where:

- $\mu(x)$ (mean) is the predicted central value,
- $\sigma^2(x)$ (variance) represents the uncertainty.

The model's goal is to learn the functions $\mu(x)$ and $\sigma^2(x)$.

Compute the Natural Gradient of the Log-Likelihood. Instead of using the standard gradient ∇_θ, NGBoost computes the natural gradient, which respects the geometry of the probability distribution:

$$\widetilde{\nabla}_\theta = F^{-1}\nabla_\theta \mathcal{L} \tag{12}$$

where:

- $\mathcal{L}$ is the negative log-likelihood (NLL) loss function,
- $\nabla_\theta \mathcal{L}$ is the standard gradient of the loss,
- F is the Fisher Information Matrix, which adjusts for the curvature of the parameter space.

For a Gaussian distribution, the log-likelihood is:

$$\mathcal{L} = \frac{(y-\mu)^2}{2\sigma^2} + \log\ \sigma \tag{13}$$

The natural gradient updates are then derived using:

- For mean μ:
  $$\widetilde{\nabla}_\mu = \frac{y-\mu}{\sigma^2} \tag{14}$$
- For variance σ^2:
  $$\widetilde{\nabla}_{\sigma^2} = \frac{(y-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2} \tag{15}$$

Fit gradient boosting trees to update parameters. NGBoost uses decision trees as base learners to fit these natural gradients and update the parameters iteratively. For each boosting iteration t:

1. Compute the natural gradient residuals $\widetilde{\nabla}_\mu$ and $\widetilde{\nabla}_{\sigma^2}$.
2. Fit regression trees $f_t(x)$ to predict these residuals.
3. Update the parameters using a step size ρ:
4. Repeat until convergence.

$$\begin{aligned}\mu_{t+1} &= \mu_t + \rho f_\mu(x) \\ \sigma_{t+1}^2 &= \sigma_t^2 + \rho f_{\sigma^2}(x)\end{aligned} \tag{16}$$

Make predictions with uncertainty estimation. After training, NGBoost outputs the final predicted distribution $P(y \mid x)$.

- The predicted value $\hat{y}_i$, is usually taken as the mean $\mu(x)$.
- The model also provides prediction intervals using the variance $\sigma^2(x)$, allowing for uncertainty-aware predictions.

NGBoost differs from XGBoost in that it:

- Outputs full probability distributions instead of single-point estimates
- Uses natural gradient updates to optimize likelihood functions
- Provides uncertainty intervals, improving reliability in real-world applications

Both models were trained on the Orchard_A dataset and tested on the Orchard_B dataset to compare their effectiveness in predicting soil NPK levels.

2.5 Model Evaluation

To assess and compare the predictive performance of XGBoost and NGBoost, multiple evaluation metrics were used, R^2, MAE, RMSE, and Uncertainty Analysis.

3. RESULTS AND DISCUSSION

This study evaluates and compares XGBoost and NGBoost for predicting the added amount of NPK fertilizer in precision agriculture. The objective is to assess their predictive accuracy and uncertainty estimation capabilities, ensuring optimal fertilizer application based on soil nutrient conditions. The models were trained and tested using three datasets: training data (to fit the models and analyze their learning performance), testing data (to evaluate generalization ability), and new unseen data (to validate model performance in practical applications).

The results indicate that XGBoost outperforms NGBoost in predictive accuracy, achieving lower errors and higher R^2 scores across all datasets. XGBoost consistently demonstrated a lower MAE, RMSE, and MASE, confirming its reliability in estimating the required NPK fertilizer. In contrast, NGBoost, while slightly less accurate, provides uncertainty quantification through NLL, which is valuable for decision-making under uncertain conditions. This highlights NGBoost's ability to offer probabilistic insights, making it useful in scenarios where confidence intervals are necessary. The findings confirm that XGBoost is the optimal choice when high precision is required, whereas NGBoost offers a balanced approach by incorporating predictive uncertainty. This comparative analysis provides important insights for integrating machine learning into precision agriculture, enabling data-driven fertilization strategies that optimize soil nutrition while managing variability and uncertainty.

3.1 Pre-processing

Kernel Density Estimation (KDE) plots were employed to evaluate the distribution of soil properties across different datasets. Figure 2 illustrates the probability density functions for key soil parameters, comparing Orchard A and Orchard B. KDE plots provide a smoothed estimation of the probability distribution of each variable, facilitating a detailed comparison of feature distributions between datasets.

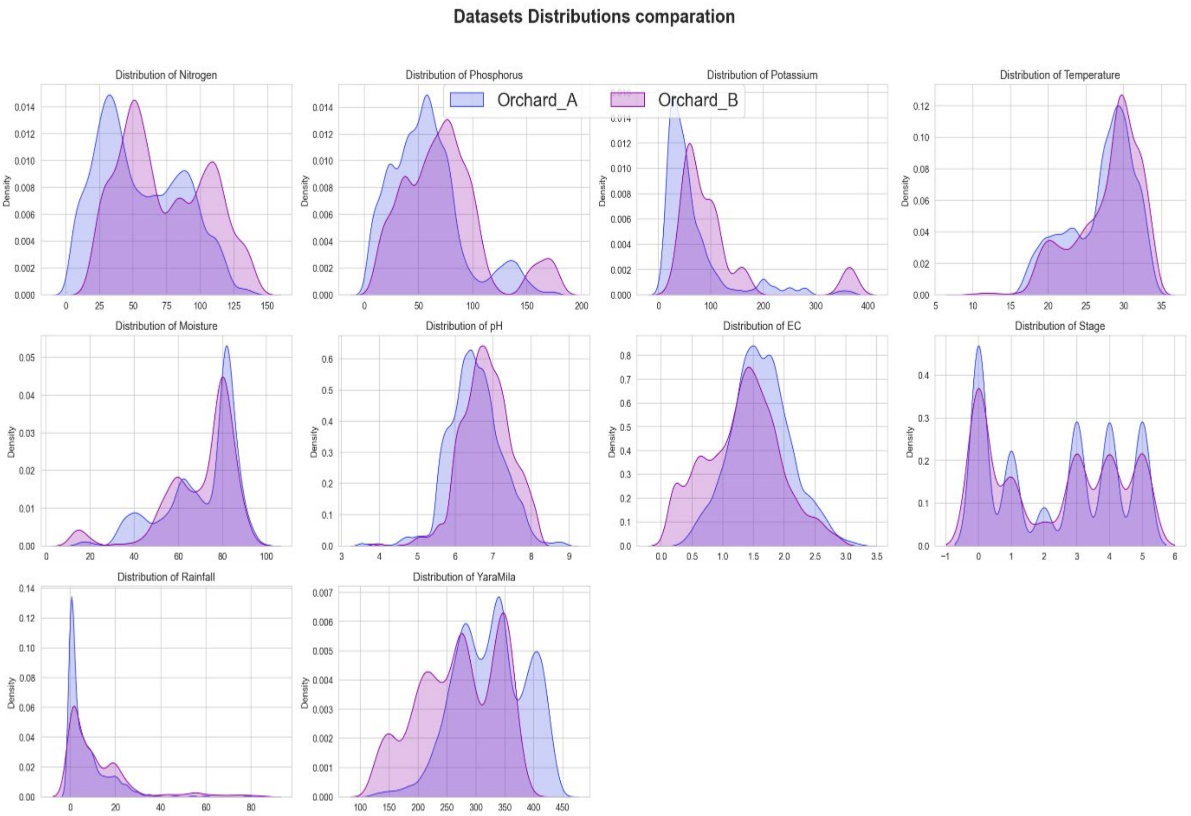

Figure 2: Comparative KDE Analysis of Soil Properties Across Different Orchard Datasets.

In the nitrogen distribution, Orchard_A exhibits a higher density in the lower nitrogen range, whereas Orchard_B shows a broader distribution, indicating potential variability in soil nutrient availability. Similar trends are observed for phosphorus and potassium, where differences in density highlight spatial variations in nutrient content between the two orchards. Additionally, the pH and EC distributions indicate minor shifts in soil chemical properties, which could impact model generalization. The rainfall and moisture distributions demonstrate distinct peaks, suggesting environmental variations that may influence soil nutrient dynamics.

These KDE-based visualizations play a crucial role in understanding the underlying data distribution before model training. Identifying discrepancies between training and testing datasets ensures appropriate pre-processing steps, such as normalization and transformation, to enhance model robustness and predictive performance.

EDA was conducted on two datasets in Figure 4: Orchard A (training set) and Orchard B (testing set), focusing on soil nutrients (N, P, K), pH, EC, moisture, temperature, rainfall, and phenological stages. Nitrogen in Orchard_A is left-skewed (30–60 mg/kg), while Orchard_B has a broader spread (peaking at 80–100 mg/kg), indicating different fertilization practices. P shows a multimodal distribution, with higher values in Orchard B (>125 mg/kg). K in Orchard A is right-skewed (<100 mg/kg), whereas Orchard_B has extreme values (>200 mg/kg), suggesting high variability.

Soil properties differ between orchards. pH follows a normal distribution (6.5–7.5), though Orchard_B has more extreme values (<5, >8). EC is right-skewed, with higher values in Orchard B, indicating possible salinity issues. Moisture distribution is bimodal (40% and 80%), likely due to irrigation differences.

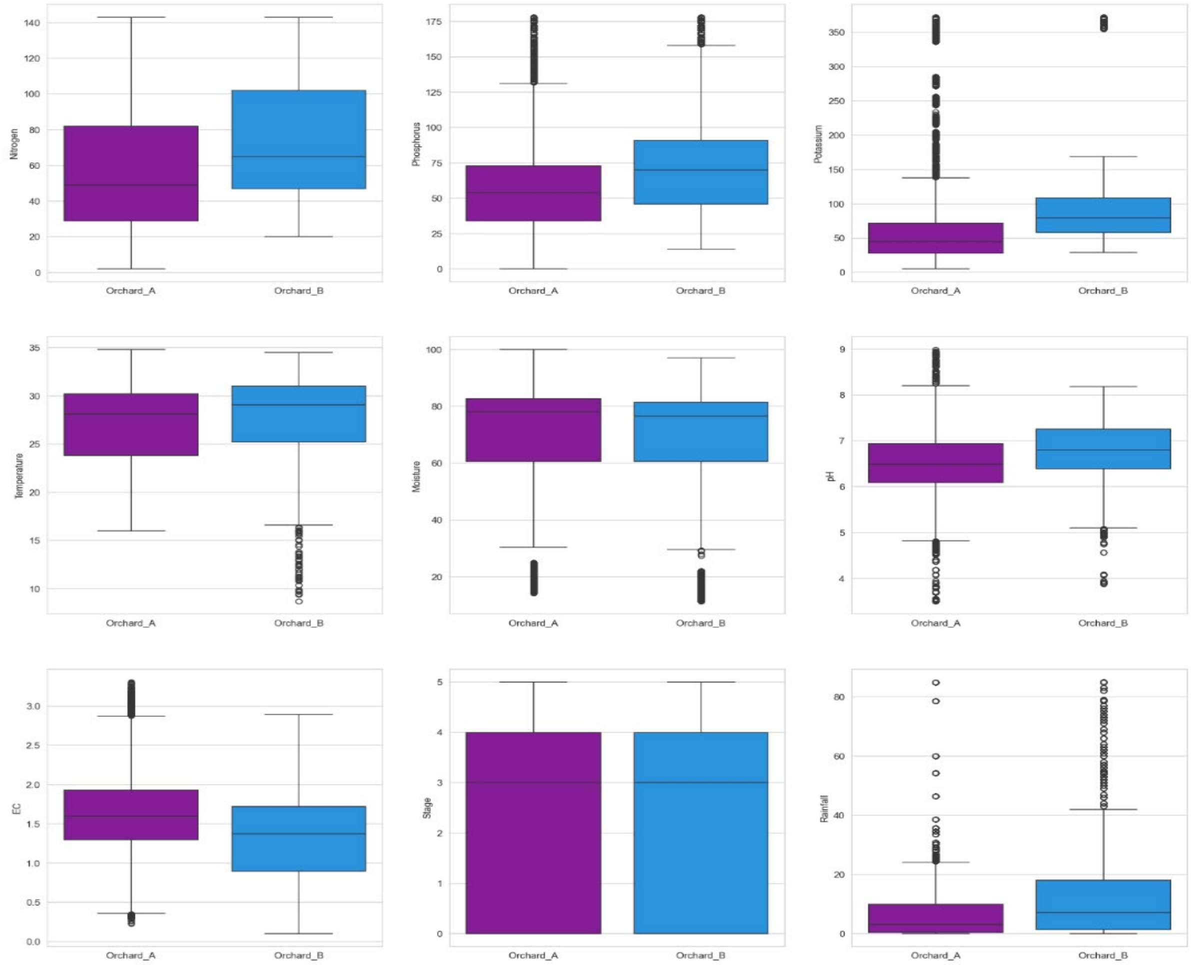

Figure 4: Exploratory Data Analysis (EDA): Distribution Comparison of Training and Testing Sets,

Environmental conditions also vary. Temperature distributions are similar (25–30°C), but Orchard_B experiences extreme rainfall events (>50 mm), impacting nutrient leaching. Phenological stages are evenly represented, ensuring model generalizability. These findings highlight the need for data normalization and feature engineering to address skewness and site-specific factors, improving predictive model accuracy and robustness.

3.2 Model Performance on Training and Testing Data

The performance of XGBoost and NGBoost was evaluated using MAE, RMSE, and R^2 scores for both training and testing datasets, shown in Table 2. XGBoost outperformed NGBoost in predictive accuracy, achieving an R^2 of 0.9991 on the training set and 0.9988 on the testing set, demonstrating strong predictive capability. The model exhibited low error rates, with an MAE of 1.3033 and RMSE of 1.8219 for training, and an MAE of 1.5037 with RMSE of 2.1276 for testing, indicating high reliability in predicting NPK values. The learning curve of the XGBoost model for training and testing data is presented in Figure 5, further confirming the model's robustness and stability.

NGBoost, designed for probabilistic prediction, performed slightly lower, with an R^2 of 0.9956 on training and 0.9949 on testing. It recorded a higher MAE (2.8867 for training, 3.0651 for testing) and RMSE (4.0689 for training, 4.3638 for testing), indicating greater prediction errors compared to XGBoost. However, NGBoost's probabilistic nature allows for uncertainty estimation, providing additional insights into prediction confidence.

These results highlight XGBoost's superior accuracy in predicting the added amount of fertilizer, while NGBoost offers valuable uncertainty quantification. The trade-off between precision and uncertainty estimation is crucial for optimizing fertilizer application in precision agriculture.

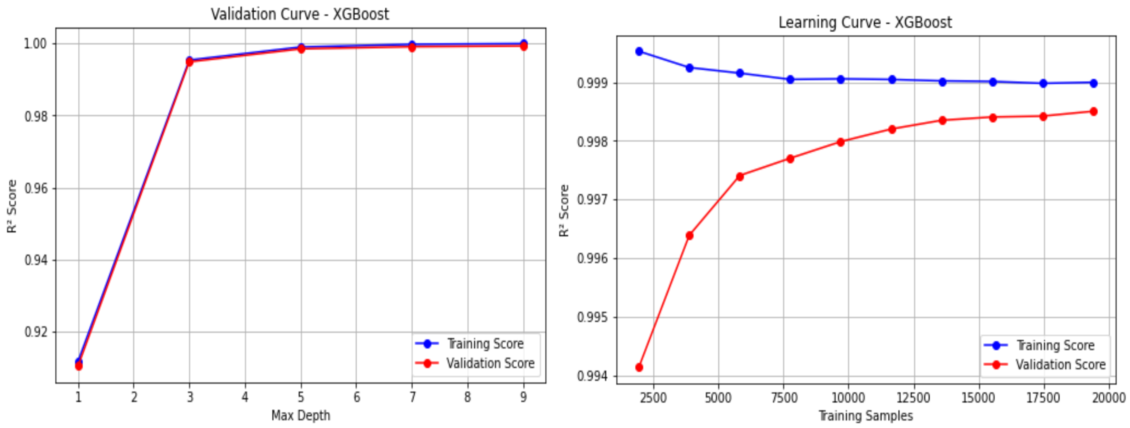

Figure 5: Learning curve for Training and Testing Data using XGBoost Model.

Table 2: Model Performance for XGBoost and NGBoost models.

Model	Dataset	MAE	RMSE	R^2
XGBoost	Training	1.3033	1.8219	0.9991
XGBoost	Testing	1.5037	2.1276	0.9988
NGBoost	Training	2.8867	4.0689	0.9956
NGBoost	Testing	3.0651	4.3638	0.9949

Evaluation and compare the predictive performance of the XGBoost and NGBoost models, both residual analysis and prediction error plots were utilized. These visual diagnostics provide insight into the models' accuracy, bias, and generalization capabilities. As shown in Figure 6, the residual plots reveal the distribution of prediction errors for both models across the range of actual YaraMila values.

The residual analysis for both models shows that errors are mostly concentrated around zero, indicating minimal prediction bias. For XGBoost, residuals remain tightly clustered with only a slight increase in variance at higher YaraMila values, reflecting strong generalization capacity and minimal outliers. NGBoost also centers residuals around zero but exhibits slightly greater dispersion, particularly in the mid-to-high YaraMila range, a behavior expected from its probabilistic design, which models predictive uncertainty. The prediction error plots (Figure 6) further illustrate these differences: XGBoost displays a very tight clustering of points along the diagonal reference line, achieving an R^2 of 0.9984, while NGBoost also aligns closely with the reference line but with a slightly lower R^2 of 0.9909. This comparison highlights the trade-off between the two approaches: XGBoost excels in deterministic accuracy with lower residual spread, whereas NGBoost, though less precise, provides valuable uncertainty quantification that supports risk-aware fertilizer management. Together, these insights suggest that model choice may depend on context: XGBoost for high-accuracy predictions in stable conditions, and NGBoost for decision-making under uncertainty where confidence intervals are critical to avoid economic loss or environmental harm.

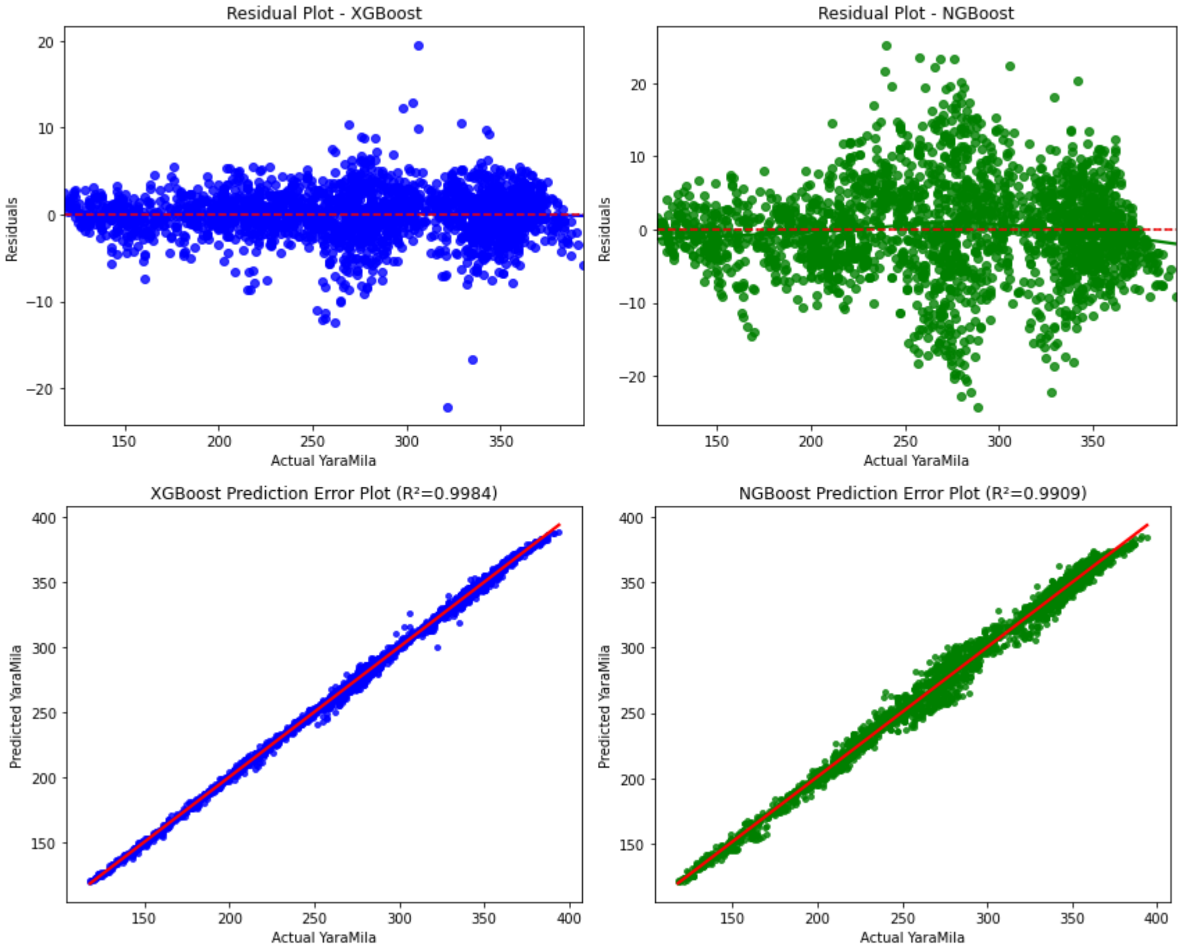

Figure 6: Prediction error Plot for XGBoost and NGBoost models.

3.3 Model Prediction on a Test Dataset (Orchard B)

The performance of XGBoost and NGBoost is shown in Table 4, and further evaluated on a new dataset to assess their generalization ability. XGBoost maintained high predictive accuracy with an R^2 score of 0.9984, indicating a strong fit to the data. It achieved a low MAE of 2.0388 and RMSE of 2.7618, demonstrating minimal prediction errors. The MASE of 0.1791 suggests stable performance, with minimal deviation from actual values.

In contrast, NGBoost exhibited a lower R^2 score of 0.9909, reflecting slightly reduced predictive accuracy. The model produced a higher MAE of 4.962 and RMSE of 6.5654, indicating larger errors in predictions. The MASE value of 0.4361 further confirms increased deviation from actual values compared to XGBoost. NGBoost provides uncertainty quantification through Negative Log-Likelihood (NLL), which was recorded at 2.6001. This suggests that NGBoost, while less precise, offers valuable probabilistic insights into prediction confidence.

XGBoost remains the more accurate model for predicting fertilizer requirements, while NGBoost presents an alternative for scenarios where uncertainty estimation is crucial. These results reinforce XGBoost's suitability for precision agriculture applications, ensuring reliable fertilizer recommendations with minimal prediction errors.

Table 4: Performance of XGBoost and NGBoost Model using test Orchard B.

Model	R^2 Score	MAE	MSE	RMSE	MASE	NLL
XGBoost	0.9984	2.0388	7.6273	2.7618	0.1791	-
NGBoost	0.9909	4.962	43.1048	6.5654	0.4361	2.6001

Figure 7 highlights NGBoost's ability to provide probabilistic estimates, with uncertainty increasing for lower actual values. The first subplot (top) compares XGBoost predictions against actual values, with a red dashed line representing a perfect fit. The second subplot (middle) shows NGBoost predictions with associated uncertainty bars. The third subplot (bottom) directly compares XGBoost and NGBoost predictions.

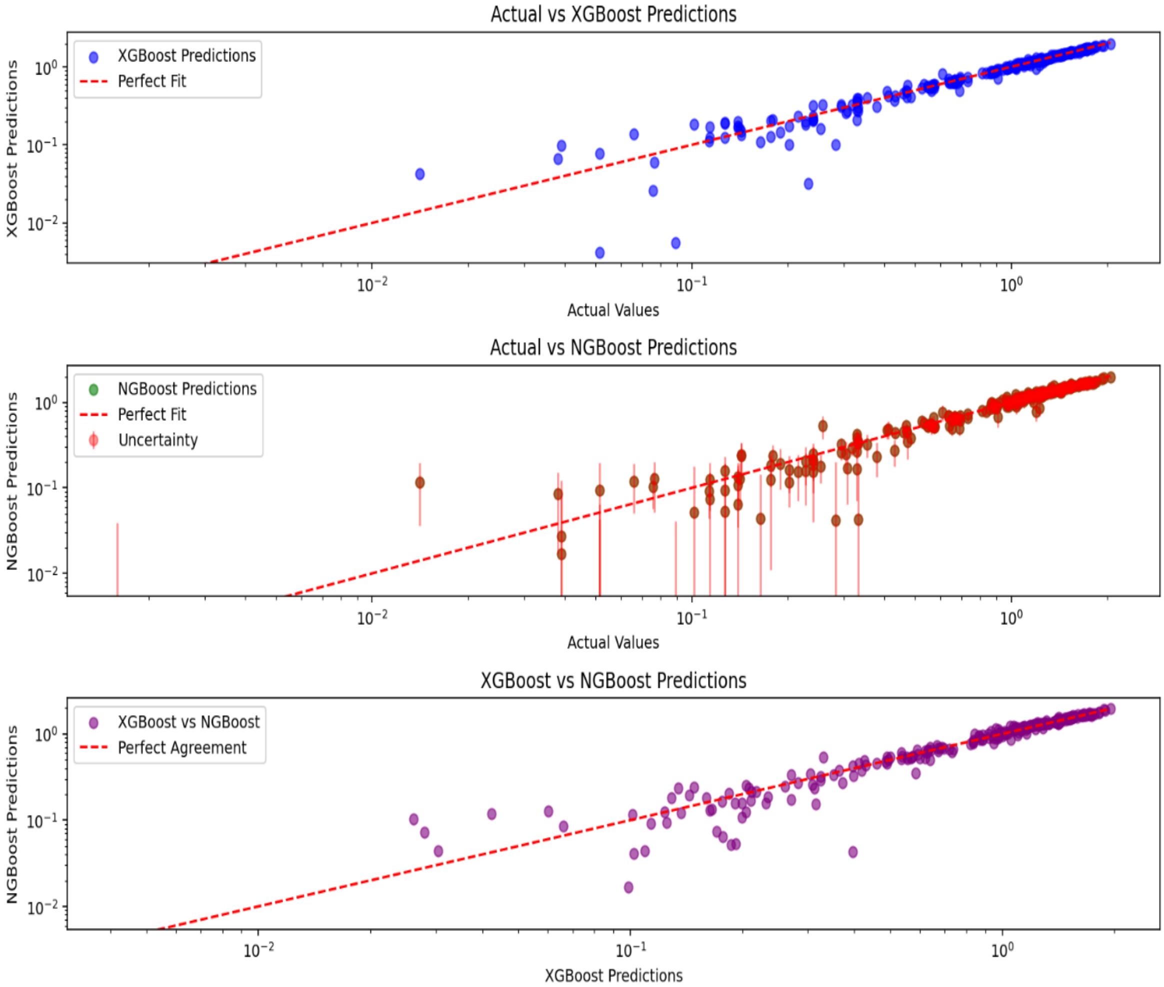

Figure 7: Comparison of XGBoost and NGBoost Model Prediction using test dataset (Orchard_B).

3.3.1 NGBoost Predictions with Uncertainty Intervals

Figure 8 illustrates the prediction results of the NGBoost model on a test dataset comprising 50 samples, highlighting the model's ability to provide both point predictions and uncertainty estimates. The red dots represent the actual target values (ground truth), while the blue dots denote the NGBoost predicted means along with their associated uncertainty intervals (typically ±1 standard deviation of the predictive distribution). The NGBoost model demonstrates a good agreement between the predicted and actual values across most test samples. The majority of actual values fall within the predicted uncertainty bounds, indicating the model's effectiveness in quantifying predictive uncertainty. This capability is particularly important in precision agriculture applications where incorrect fertilizer recommendations can have economic or environmental consequences.

The size of the uncertainty intervals varies across the test samples, reflecting the model's confidence in its predictions. Samples with higher uncertainty likely correspond to input conditions that are less represented in the training data or exhibit more variability. Conversely, narrow intervals indicate higher confidence and more consistent model behavior.

This result highlights the advantage of NGBoost over traditional deterministic models like XGBoost, as it not only makes accurate predictions but also provides a probabilistic framework for risk-aware decision-making in NPK fertilizer application. Such insights can support smarter resource allocation and reduce over- or under-fertilization in Harumanis mango cultivation.

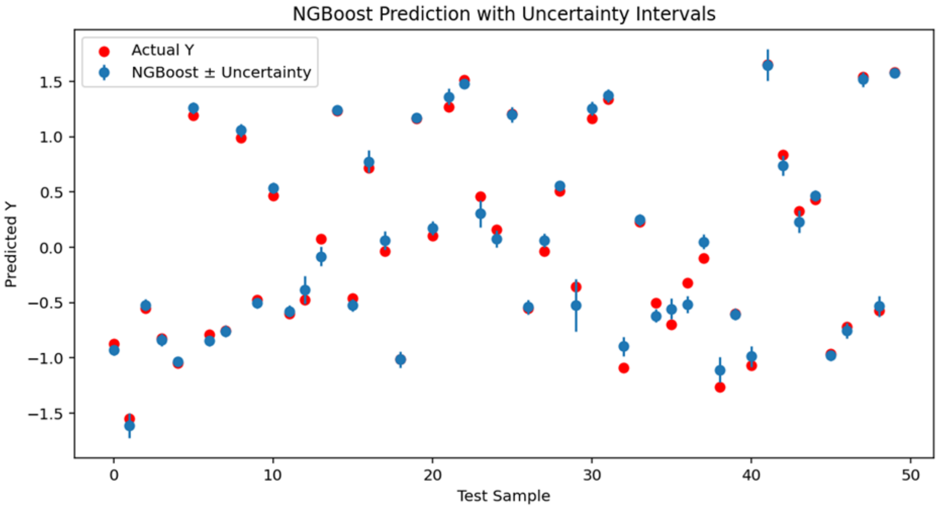

Figure 8: NGBoost Prediction with Uncertainty Intervals.

Figure 9 shows the NGBoost model's predictions against actual values. The blue points represent predictions, and the shaded region illustrates the 95% confidence interval. The black dashed line represents a perfect fit where predictions equal actual values. The spread of the uncertainty bands indicates how confident the model is in its predictions narrow bands suggest high certainty, while wider bands indicate more uncertainty.

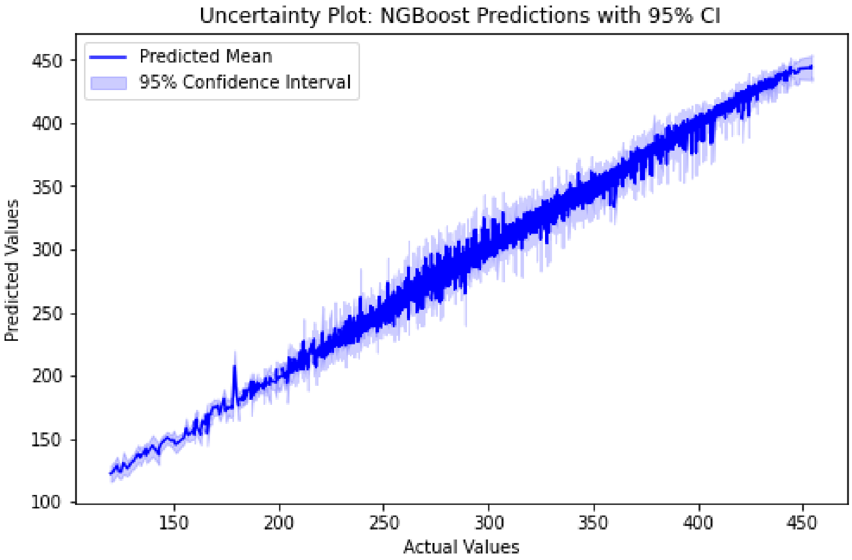

Figure 9: Uncertainty Plot for the NGBoost model's predictions with 95% Confidence Interval.

3.3.2 Distribution of Negative Log-Likelihood (NLL) for NGBoost.

Figure 10 represents the distribution of NLL values for NGBoost, with the red dashed line indicating the mean NLL (2.6001). Lower NLL values indicate better probabilistic calibration of predictions. The distribution's concentration around the mean suggests consistency, but a long tail may indicate occasional high-uncertainty predictions.

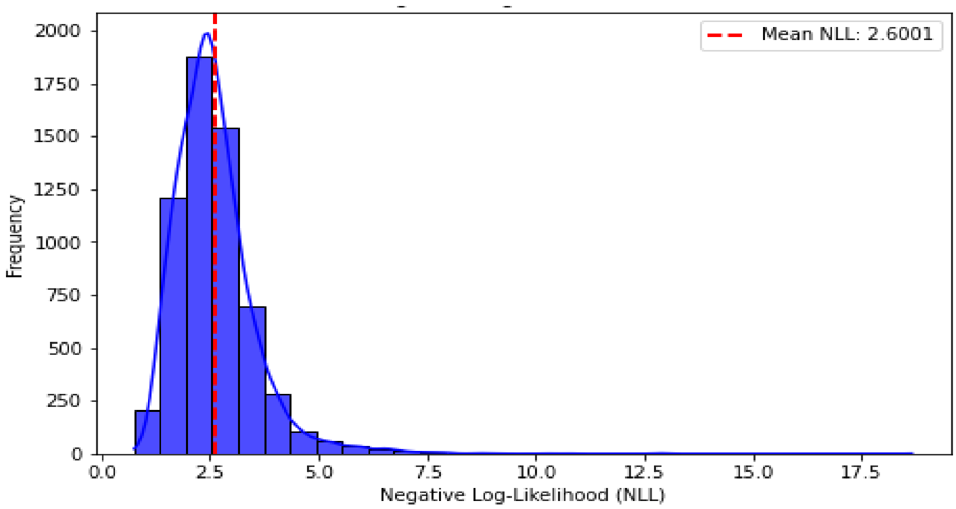

Figure 10: Distribution of negative Log-Likelihood (NLL) for NGBoost.

4. CONCLUSION

This study aimed to achieve three key research objectives, determine the optimal fertilizer rates based on CSTV at different phenological stages, develop machine learning models (XGBoost and NGBoost) to predict the added amount of NPK fertilizer, and validate model accuracy and performance, ensuring its reliability for real-world precision agriculture applications.

The results successfully demonstrated that XGBoost achieved superior predictive performance, with lower MAE, RMSE, and higher R^2 values, indicating its high accuracy in estimating added NPK. Meanwhile, NGBoost provided robust uncertainty quantification, as reflected in the NLL values, allowing for more probabilistic decision-making in fertilizer application. These findings confirm that the research objectives were met, as the study developed and validated an effective machine learning-based approach for soil nutrient prediction.

Furthermore, visualization techniques such as residual plots, prediction error plots, learning curves, and validation curves provided additional evidence of model robustness and generalization ability. The successful implementation of these models suggests that machine learning can significantly enhance precision agriculture by enabling data-driven fertilization strategies, reducing nutrient wastage, and minimizing environmental impact.

Future research should explore hybrid modeling approaches that combine deterministic and probabilistic frameworks for improved fertilizer recommendation accuracy. Additionally, incorporating spatiotemporal data, real-time sensor integration, and weather variability factors could enhance model performance and adaptability.

In conclusion, this study contributes to the advancement of AI-driven precision agriculture, demonstrating that machine learning models, particularly those with uncertainty estimation, can optimize soil nutrient management and promote sustainable farming practices.

ACKNOWLEDGEMENTS

This work was supported by Yayasan Inovasi Malaysia (YIM) and the Ministry of Science, Technology, and Innovation (MOSTI), Malaysia, under the MyGRiS grant scheme [Grant amount: RM50,000.00, 2023–2024]. The authors would like to express their sincere gratitude for the financial support that made this research possible.

REFERENCES

[1] Hedley, C. The role of precision agriculture for improved nutrient management on farms. Journal of the Science of Food and Agriculture, vol. 95, issue 1 (2014), pp. 12–19. https://doi.org/10.1002/jsfa.6734

[2] Karunathilake, E. M. B. M., Heo, S., Chung, Y. S., Mansoor, S., and Le, A. T. The path to smart farming: innovations and opportunities in precision agriculture. Agriculture, vol. 13, issue 8 (2023), p. 1593. https://doi.org/10.3390/agriculture13081593

[3] Xing, Y., and Wang, X. Precise application of water and fertilizer to crops: challenges and opportunities. Frontiers in Plant Science, vol. 15 (2024), p. 1444560. https://doi.org/10.3389/fpls.2024.1444560

[4] Shrestha, M. M., and Wei, L. Review—Perspectives on the roles of real-time nitrogen sensing and IoT integration in smart agriculture. Journal of The Electrochemical Society, vol. 171, issue 2 (2024), p. 027526. https://doi.org/10.1149/1945-7111/ad22d8

[5] Garg, S., Rumjit, N. P., and Roy, S. Smart agriculture and nanotechnology: technology, challenges, and new perspectives. Advanced Agrochem, vol. 3, issue 2 (2023), pp. 115–125. https://doi.org/10.1016/j.aac.2023.11.001

[6] Soussi, A., Trinchero, D., Fossa, M., Sacile, R., and Zero, E. Smart sensors and smart data for precision agriculture: a review. Sensors, vol. 24, issue 8 (2024), p. 2647. https://doi.org/10.3390/s24082647

[7] Graf, R., Zhu, S., and Kolerski, T. Predicting ice phenomena in a river using the artificial neural network and extreme gradient boosting. Resources, vol. 11, issue 2 (2022), p. 12. https://doi.org/10.3390/resources11020012

[8] Díaz, E., and Spagnoli, G. Natural gradient boosting for probabilistic prediction of soaked CBR values using an explainable artificial intelligence approach. Buildings, vol. 14, issue 2 (2024), p. 352. https://doi.org/10.3390/buildings14020352

[9] Chen, S.-Z., Feng, D.-C., Taciroglu, E., and Wang, W.-J. Probabilistic machine-learning methods for performance prediction of structures and infrastructures through natural gradient boosting. Journal of Structural Engineering, vol. 148, issue 8 (2022), p. 04022088. https://doi.org/10.1061/(ASCE)ST.1943-541X.0003401

[10] Nguyen, D. H., Heo, J.-Y., Hien Le, X., and Bae, D.-H. Development of an extreme gradient boosting model integrated with evolutionary algorithms for hourly water level prediction. IEEE Access, vol. 9 (2021), pp. 125853–125867. https://doi.org/10.1109/ACCESS.2021.3111287

[11] Ahmad, M., Sabri, M. M., Jamil, I., Kashyzadeh, K. R., Alguno, A. C., Keawsawasvong, S., and Al-Mansob, R. A. Extreme gradient boosting algorithm for predicting shear strengths of rockfill materials. Complexity, vol. 2022 (2022), pp. 1–11. https://doi.org/10.1155/2022/9415863

[12] Ahmad, I., Bibi, F., Ullah, H., and Munir, T. M. Mango fruit yield and critical quality parameters respond to foliar and soil applications of zinc and boron. Plants, vol. 7, issue 4 (2018), p. 97. https://doi.org/10.3390/plants7040097

[13] Yu, L., Zhou, R., Chen, R., and Lai, K. K. Missing data preprocessing in credit classification: one-hot encoding or imputation? Emerging Markets Finance and Trade, vol. 58, issue 2 (2020), pp. 472–482. https://doi.org/10.1080/1540496X.2020.1825935

[14] Sujon, K. M., Choi, K., Abdus Samad, M., Towshi, Z. T., Hassan, R. B., and Othman, M. A. When to use standardization and normalization: empirical evidence from machine learning

models and XAI. IEEE Access, vol. 12 (2024), pp. 135300–135314. https://doi.org/10.1109/ACCESS.2024.3462434

[15] Abiodun, E. O., Alabdulatif, A., Abiodun, O. I., Alawida, M., and Alkhawaldeh, R. S. A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. Neural Computing & Applications, vol. 33, issue 22 (2021), pp. 15091–15118. https://doi.org/10.1007/s00521-021-06406-8

[16] Eskandarinasab, M., Hamdi, S. M., and Filali Boubrahimi, S. Impacts of data preprocessing and sampling techniques on solar flare prediction from multivariate time series data of photospheric magnetic field parameters. The Astrophysical Journal Supplement Series, vol. 275, issue 1 (2024), p. 6. https://doi.org/10.3847/1538-4365/ad7c4a

[17] Kulanuwat, L., Chantrapornchai, C., Maleewong, M., Boonya-Aroonnet, S., Wimala, S., Sarinnapakorn, K., and Wongchaisuwat, P. Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series. Water, vol. 13, issue 13 (2021), p. 1862. https://doi.org/10.3390/w13131862

[18] Toselli, M., Baldi, E., Ferro, F., Rossi, S., and Cillis, D. Smart farming tool for monitoring nutrients in soil and plants for precise fertilization. Horticulturae, vol. 9, issue 9 (2023), p. 1011. https://doi.org/10.3390/horticulturae9091011

[19] Lu, W., Hao, Z., Lin, M., Fan, X., Gao, J., Li, J., Zhou, Y., Guo, J., and Ma, X. Effects of different proportions of organic fertilizer replacing chemical fertilizer on soil nutrients and fertilizer utilization in gray desert soil. Agronomy, vol. 14, issue 1 (2024), p. 228. https://doi.org/10.3390/agronomy14010228

[20] Reza, M. N., Lee, K.-H., Karim, M. R., Haque, M. A., Bicamumakuba, E., Dey, P. K., Jang, Y. Y., and Chung, S.-O. Trends of soil and solution nutrient sensing for open field and hydroponic cultivation in facilitated smart agriculture. Sensors (Basel, Switzerland), vol. 25, issue 2 (2025), p. 453. https://doi.org/10.3390/s25020453

[21] Agrahari, R. K., Kobayashi, Y., Tanaka, T. S. T., Panda, S. K., and Koyama, H. Smart fertilizer management: the progress of imaging technologies and possible implementation of plant biomarkers in agriculture. Soil Science and Plant Nutrition, vol. 67, issue 3 (2021), pp. 248–258. https://doi.org/10.1080/00380768.2021.1897479

[22] Arora, P., Singh, R., and Patel, S. Probabilistic modeling for nutrient variability analysis in precision agriculture using ensemble learning. Computers and Electronics in Agriculture, vol. 213 (2024), p. 108091.

[23] Ma, J., Zhang, Y., and Liu, T. Integration of uncertainty-aware deep learning models for soil nutrient prediction under varying environmental conditions. Sustainability, vol. 17, issue 4 (2025), p. 2330.

Conflict of interest statement: The authors declare no conflict of interest. The funding sponsor, Yayasan Inovasi Malaysia (YIM) under the Ministry of Science, Technology and Innovation (MOSTI), Malaysia, had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Author contributions statement: Conceptualization, E.S.M.M.T.; Methodology, E.S.M.M.T.; Data Collection, E.S.M.M.T. & M.A.Z.A.; Model Development, E.S.M.M.T.; Result Interpretation, E.S.M.M.T. & M.A.B.M.; Visualization, A.M.A. & P.N.; Literature Review, A.M.A.; Writing – Original Draft Preparation, E.S.M.M.T. & M.A.B.M.; Writing – Review & Editing, I.H.W.N, A.H.A., N.A.R., & F.S.A.S.; Supervision, A.H.A., N.A.R., & F.S.A.S.; Project Administration, A.H.A., N.A.R., & F.S.A.S.