Hybrid XGBoost and LSTM Model for Added Amount NPK Fertiliser Prediction in Mango Farm

Erdy Sulino Bin Mohd Muslim Tan[1]*, Marni Azira Binti Markom[2], Allan Melvin Andrew[1], Imaduddin Helmi Bin Wan Nordin[3], Shahriman Bin Abu bakar[3], Arni Munira Markom[5], Pubalan Nadaraja[1], Norazila Binti Shoib[1], Addzrull Hi-Fi Syam Bin Ahmad Jamil[4], Mohd Amri Bin Zainol Abidin[3], Farah Hanan Binti Mohd Faudzi[1]

[1]Faculty of Electrical Engineering and Technology, Universiti Malaysia Perlis, Perlis, Malaysia.
[2]Faculty of Intelligent Computing, Universiti Malaysia Perlis, Perlis, Malaysia.
[3]Faculty of Mechanical Engineering and Technology, Universiti Malaysia Perlis, Perlis, Malaysia.
[4]Department of Electrical Engineering, Polytechnic Seberang Perai, Pulau Pinang, Malaysia.
[5]Faculty of Electrical Engineering, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia.
*erdysulino@unimap.edu.my

Abstract. Efficient fertiliser management is critical for optimising crop yields and reducing environmental degradation. In Malaysia, however, optimal fertiliser use remains a challenge. Many farmers rely heavily on personal experience or weather patterns, often neglecting the actual nutrient status of the soil. This lack of data-driven fertiliser application can lead to both under and over-fertilisation, resulting in reduced productivity and increased environmental risks. This study introduces a hybrid predictive model that combines Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) neural networks to estimate the added amount of nitrogen (N), phosphorus (P), and potassium (K) fertilisers needed for Harumanis mango cultivation. The model integrates the robust feature extraction of XGBoost with LSTM's capability to model temporal nutrient decay using real-time sensor data, including soil pH, moisture, electrical conductivity (EC), temperature, and rainfall. After hyperparameter optimisation, the hybrid model achieved superior predictive accuracy compared to the individual models, with a Mean Squared Error (MSE) of 0.000581, Mean Absolute Error (MAE) of 0.018369, and R^2 score of 0.9826. These results confirmed that the hybrid approach effectively captured both the spatial and temporal dynamics of soil nutrients. This study contributes to the advancement of precision agricul-

A. Akhyar et al. (eds.), *Proceedings of the International Conference on Research, Innovation, Sustainability, and Educations (IRISECON 2026)*, Atlantis Advances in Applied Sciences 5,
https://doi.org/10.2991/978-94-6239-745-3_20

ture by offering a reliable data-driven method for nutrient recommendations throughout the phenological stages of mango farming.

Keywords: Hybrid Model, Predictive Modeling, Fertiliser Optimization, Agricultural Innovation

1. Introduction

Precision agriculture has emerged as a vital approach for enhancing agricultural productivity while minimising resource wastage. It leverages modern technologies and data-driven insights to make farming practices more efficient, sustainable, and responsive to real-time field conditions [1,2,3]. Among various agricultural operations, fertiliser management, particularly the prediction and application of N, P, and K, is one of the most critical components in mango cultivation. These nutrients play essential roles across different growth stages, directly influencing vegetative development, flowering, fruit set, and overall crop yield [4,5].

Despite the growing adoption of precision agriculture technologies, there remains a significant gap in the practical implementation of data-driven fertiliser management [1] in real farm environments, particularly for perennial crops such as Harumanis mango. Existing fertilisation practices are largely empirical, lacking real-time adaptation to soil nutrient variability and phenological requirements [2,3]. Furthermore, current machine learning approaches in fertiliser prediction often rely on static datasets and fail to adequately capture temporal nutrient dynamics, such as nutrient decay and environmental interactions. This limitation leads to suboptimal fertiliser recommendations, reducing both economic efficiency and environmental sustainability. Therefore, there is a critical need for an integrated predictive framework that can simultaneously model nonlinear relationships and temporal dependencies using real-time field data.

Compounding the issue, Harumanis mango trees produce fruit only once a year, and their full phenological cycle lasts approximately 10 months. Identifying the correct timing for nutrient application, particularly at the beginning of the season, is often difficult for farmers. Without real-time nutrient information, decision-making becomes largely based on guesswork. Traditional fertilisation methods fail to capture the dynamic interactions among soil nutrients, environmental factors, and plant uptake, often leading to over- or under-application of fertilisers. Overuse accelerates problems such as soil salinity and water pollution, while underuse results in poor plant growth and reduced yield potential [6,7].

Recent advancements in machine learning (ML) and deep learning (DL) have provided promising solutions to address these complexities in agriculture. These technologies have been widely applied in crop yield prediction, irrigation management, and fertiliser optimisation [8,9,10]. Hybrid artificial intelligence models have demonstrated superior capability in capturing nonlinear and multidimensional relationships within agricultural data [11,12].

However, the deployment of AI in real farm environments still faces several challenges, including the requirement for large datasets, limited interpretability, and dependency on digital infrastructure. Furthermore, the lack of real-time, field-deployable AI systems limits practical adoption [13,14].

In response to these challenges, this study proposes a hybrid model integrating XGBoost and LSTM to predict optimal NPK fertiliser requirements for Harumanis mango cultivation. The hybrid approach leverages XGBoost's nonlinear modelling strength and LSTM's temporal learning capability, enabling effective handling of both static and sequential data [15,16].

To address the identified gaps, this study aims to:

1. To develop a hybrid XGBoost–LSTM model for predicting the required amount of N, P, and K fertilisers in Harumanis mango cultivation.
2. To integrate real-time soil and environmental data, including pH, moisture, EC, temperature, and rainfall, into a predictive framework.
3. To evaluate and compare the performance of the hybrid model with individual XGBoost and LSTM models using standard evaluation metrics (MAE, MSE, RMSE, and R^2).
4. To analyse the effectiveness of the hybrid approach in capturing both spatial and temporal variations of soil nutrients across different phenological stages.

2. Related Previous Studies

Fertilizer prediction has evolved under Agriculture 4.0 and 5.0 paradigms, integrating AI, IoT, and precision agriculture for sustainable farming [1,9]. ML and DL models have been widely applied to predict fertiliser needs based on soil and environmental data [9,13,17].

ML models have demonstrated strong performance in crop yield prediction and soil analysis. Data-driven approaches enable efficient fertiliser recommendations and improved farm management [2,8,18]. These techniques effectively model complex environmental relationships, enhancing decision-making accuracy [11].

DL approaches, particularly LSTM, have shown strong capability in modelling temporal agricultural data. Applications include climate prediction, soil monitoring, and crop classification [10,19]. Hybrid DL models further improve performance by integrating multiple data sources [15,16].

XGBoost has emerged as a powerful ML algorithm for environmental modelling and fertiliser prediction. It has been widely used for soil analysis, forecasting, and feature selection due to its robustness and efficiency [20,21,22]. Hybrid XGBoost–LSTM models further enhance prediction accuracy by combining spatial and temporal learning [15,23].

Despite these advancements, limitations remain, including lack of real-time deployment, reliance on offline datasets, and limited adaptability across locations. Therefore, this study proposes a hybrid AI model that integrates XGBoost and LSTM to address these challenges and improve fertiliser prediction accuracy in real-world agricultural environments.

3. Methodology

The methodology integrates XGBoost and LSTM to improve prediction accuracy of NPK fertiliser requirements. XGBoost is used for feature importance analysis, identifying key variables such as soil pH, moisture, temperature, EC, and rainfall [20,22]. Meanwhile, LSTM captures temporal dependencies in nutrient dynamics and environmental variations [10,19].

The hybrid model combines the strengths of both approaches, enabling effective modelling of structured data and time-series patterns [15,23]. Sensor-based data collected from mango orchards are used to train the model, ensuring realistic and site-specific predictions. This approach provides a more accurate and data-driven solution compared to traditional fertiliser recommendation methods.

3.1. Study Area and Data Colletion

The study was conducted at Harumanis mango farms located in Muzium Mempelam, Kedah (Location: 5°58'06.0"N 100°24'03.1"E) (serving as the training site), and Guar Nangka, Perlis (Location: 6°28'34.8"N 100°17'06.2"E) (serving as the testing site). Real-time soil data were collected using TDR sensors (real-time soil data collection), which measured various soil parameters, including N, P, K, pH, temperature, EC, moisture, and rainfall.

Figure 1: Agri NPK sensor device photograph.

Data were gathered across different phenological stages of mango trees, including the vegetative, flowering, and fruiting stages, ensuring comprehensive data representation for model development. Figure 1 shows the developed sensor device and Table 1 below provides a detailed overview of the monitoring framework used in this study.

Table 1: Monitoring framework for NPK and environmental parameters in Harumanis mango orchards

Parameter	Description
Duration of Study	From 1 May 2021 to 15 January 2023
Data Collection Period	Season 1: May 2021 to May 2022, Season 2: May 2022 to January 2023
Sampling Points	6 designated sampling points within an area of 14,850 square feet
Sensor Setup	6 sensors installed (one at each sampling point)
Sensor Data Frequency	Daily, from 7:00 AM to 3:00 PM
Data Samples Collected	2,824 data samples per sensor per season
Total Data Samples	33,888 data samples collected across both seasons
Rainfall Data Source	Malaysia Meteorological Department (https://www.met.gov.my/projection/rain/48603/)
Rainfall Data Period	Daily records from 14 May 2021 to 15 January 2023, aligning with the soil nutrient monitoring timeline
Parameters Measured	Nitrogen (mg/kg), Phosphorus (mg/kg), Potassium (mg/kg), pH, Soil Temperature (˚C), Electrical Conductivity (EC), Soil Moisture (%), Rainfall (mm)
Temporal Scope	Phenological stages: Vegetative 3 Stage, Flowering, fruiting, and Pruning
Zoning Approach	Orchards divided into distinct sampling zones based on tree growth, soil texture, and management practices for accurate soil representation

3.2. Data Pre-processing

In this study, missing data were handled using a combination of interpolation and imputation techniques to ensure data completeness and integrity prior to model development. These approaches help minimise data loss and maintain the reliability of the dataset used for training and validation. To address skewness and improve data normality, several transformation techniques were applied, including logarithmic (log), Box–Cox, and Yeo–Johnson transformations. These transformations enhance the distribution of the data, making it more suitable for machine learning algorithms by improving model stability and predictive performance.

During the feature engineering phase, nutrient decay over time was estimated using established nutrient half-life formulations. This allows for a more dynamic representation of nutrient availability in the soil, reflecting real-world temporal variations that influence plant uptake and fertiliser efficiency.

Furthermore, the Critical Soil Test Value (CSTV) was determined based on Mitscherlich's law and a modified arsine logarithmic calibration curve [12]. This method enables accurate estimation of the optimal soil nutrient levels required at different phenological stages of crop development, thereby supporting precise and stage-specific fertiliser recommendations.

3.3. XGBoost Model Development

The XGBoost algorithm was employed to model and evaluate the influence of various environmental and soil features in predicting the required amount of NPK fertiliser for mango cultivation. This regression model leverages structured environmental data by integrating multiple variables relevant to soil and climatic conditions. To enhance the model performance, key hyperparameters, including the learning rate, maximum tree depth, number of estimators, and subsample ratio, were fine-tuned using a randomised search strategy. The final model was trained to predict the target variable, which was the optimal quantity of NPK fertiliser to be applied.

XGBoost constructs an ensemble of K regression trees f_k by minimising the regularised objective function. The predicted output $\hat{y}_i$ and the objective function, $\mathcal{L}^{\{(t)\}}$ are given by (1) and (2), respectively.

$$\hat{y}_i = \Sigma_{\{k=1\}}^{K} f_{k(x_i)}, \text{ where } f_k \in \mathbb{F} \quad (1)$$

$$\mathcal{L}^{\{(t)\}} = \Sigma_{\{i=1\}}^{n} l\left(y_i, \hat{y}_i^{\{(t-1)\}} + f_{t(x_i)}\right) + \Omega(f_t) \quad (2)$$

Where:

$l(y, \hat{y})$: typically Mean Squared Error (MSE)

$\Omega(f_t) = \gamma T + \left(\frac{1}{2}\right) \lambda \, \Sigma_{\{j=1\}}^{T} w_j^2$: regularization term

Table 2. The table below summarizes the optimized hyperparameters and their descriptions

Parameter	**Value**	**Description**
n_estimators	300	The number of boosting rounds (trees) in the model. More estimators generally improve the model's performance, but it can lead to overfitting if too high.
max_depth	10	The maximum depth of each tree. It controls the complexity of the model and helps prevent overfitting by limiting the depth of the decision trees.
learning_rate	0.1394	The rate at which the model updates the weights after each iteration. A lower learning rate helps the model converge more slowly, potentially improving accuracy, but requires more estimators.
subsample	0.7498	The fraction of samples used to build each tree. This parameter helps prevent overfitting by randomly sampling a subset of the data for each tree.
colsample_bytree	0.8585	The fraction of features (columns) used for each tree. By randomly selecting features, it can help the model generalize better and prevent overfitting.
gamma	0.0022	The minimum loss reduction required to make a further partition on a leaf node. A higher value prevents the model from overfitting by controlling the complexity of the tree.
min_child_weight	1	The minimum sum of instance weight (hessian) needed in a child. It controls overfitting by making sure that leaf nodes have enough data to avoid overfitting.

3.4 LSTM Model Development

An LSTM model was developed to capture the temporal dependencies within soil and environmental data, accounting for nutrient decay patterns and dynamic weather influences over time. LSTM is well suited for sequential data modelling and was employed in this study to process time-series inputs related to soil and climate variables, enabling the prediction of future NPK fertiliser requirements.

The model architecture consisted of stacked LSTM layers, followed by fully connected layers for the regression output. It was trained using sequences of historical data, including features such as soil temperature, moisture, and rainfall, to estimate the required fertiliser at the upcoming time steps. Hyperparameters parameter setting show at table 3, including the number of LSTM units, learning rate, batch size, and number of epochs, were optimised using a randomised search strategy to enhance predictive performance and generalisation. During training, the LSTM unit maintains a hidden state h_t and a cell state c_t updated through several gate mechanisms: the forget gate $f_t = \sigma\left(W_f x_t + U_f h_{\{t-1\}} + b_f\right)$, input gate $i_t = \sigma\left(W_i x_t + U_i h_{\{t-1\}} + b_i\right)$, and output gate $o_t = \sigma\left(W_o x_t + U_o h_{\{t-1\}} + b_o\right)$. The cell state is updated with $\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{\{t-1\}} + b_c\right)$, followed by $c_t = f_t \odot c_{\{t-1\}} + i_t \odot \tilde{c}_t$, and hidden state $h_t = o_t \odot \tanh(c_t)$. The final predicted output is computed as $\hat{y}_t = W_y h_t + b_y$.

Table 3: Table shows final architecture and components of the LSTM network are summarized

Layer	**Output Shape**	**Parameter**	**Description**
LSTM	(None, 10, 64)	19,200	The first LSTM layer with 64 units. It processes the sequential data and outputs sequences of 64-dimensional vectors for each time step in the input sequence.
Batch Normalization	(None, 10, 64)	256	Applies batch normalization to the output of the LSTM layer, helping to stabilize the learning process and improve the model's generalization by normalizing the activations.
Dropout	(None, 10, 64)	0	A dropout layer applied to the LSTM output to reduce overfitting by randomly setting a fraction of input units to zero during training.
LSTM_1	(None, 32)	12,416	A second LSTM layer with 32 units that processes the sequence output from the first LSTM layer and reduces the sequence length. This layer outputs a 32-dimensional vector.

Layer	Output Shape	Parameter	Description
BatchNormalization_1	(None, 32)	128	Applies batch normalization to the output of the second LSTM layer, helping to stabilize the learning and reduce internal covariate shift.
Dropout_1	(None, 32)	0	A dropout layer applied to the output of the second LSTM layer to reduce overfitting, similar to the previous dropout layer but on the 32-dimensional vector.
Dense	(None, 32)	1,056	A dense (fully connected) layer with 32 units, used to map the 32-dimensional output to another 32-dimensional representation, adding non-linearity to the model.
Dense_1	(None, 1)	33	The final output layer with a single neuron, providing the model's prediction. This layer is typically used for regression tasks or binary classification with a linear activation function.

3.5 Hybrid XGBoost-LSTM Model Development

A hybrid XGBoost–LSTM model was developed to combine the complementary strengths of both tree-based and sequence-based models. Proposed method hybrid XGBoost-LSTM shows in Figure 2. This ensemble approach combines the effectiveness of XGBoost in handling structured tabular data and extracting feature importance with LSTM's capability of LSTM in capturing temporal dependencies and time-series trends in environmental and soil variables.

In this framework, the XGBoost model is first used to analyse the feature space and provide predictions based on static and structured input variables. Simultaneously, the LSTM model processed time-series sequences and modelled temporal variations such as nutrient decay, soil moisture changes, and rainfall effects.

The final predicted output was calculated by performing a weighted average of the outputs from the two models, assigning more influence to the LSTM predictions owing to their ability to model dynamic time-dependent patterns. The ensemble prediction is defined as

$$\hat{y}_{Hybrid} = \alpha \cdot \hat{y}_{XGB} + (1 - \alpha) \cdot \hat{y}_{LSTM}$$

With $\alpha = 0.4$ (XGBoost weight),

so:

$$\hat{y}_{Hybrid} = 0.4 \cdot \hat{y}_{XGB} + 0.6 \cdot \hat{y}_{LSTM}$$

$\hat{y}_{XGB}$: predicted NPK from XGBoost

$\hat{y}_{LSTM}$: predicted NPK from LSTM

To achieve higher accuracy in predicting fertiliser requirements, a hybrid model combining the strengths of XGBoost and LSTM was developed. This approach integrates XGBoost's superior capability in handling structured/tabular data with LSTM's strength in modelling sequential dependencies, particularly useful for time-series environmental and soil nutrient data. As outlined in Algorithm 1, the process begins with Step 1 to step 8.

Algorithm 1: NPK Prediction using Hybrid XGBoost-LSTM

1: Data Preprocessing

2: Training XGBoost Model

3: Hyperparameter Tuning for XGBoost

Perform hyperparameter tuning for XGBoost using Random Search

4: Training LSTM Model

5: LSTM Hyperparameter Tuning

Perform hyperparameter tuning for LSTM using Random Search.

6: Hybrid Model Tuning (40% XGBoost and 60% LSTM)

*Final Prediction = (0.4 * $xgb_{predictions}$) + (0.6 * $lstm_{predictions}$)*

Evaluate MAE, MSE, RMSE, and R^2

7: Testing the Hybrid Model

8: End

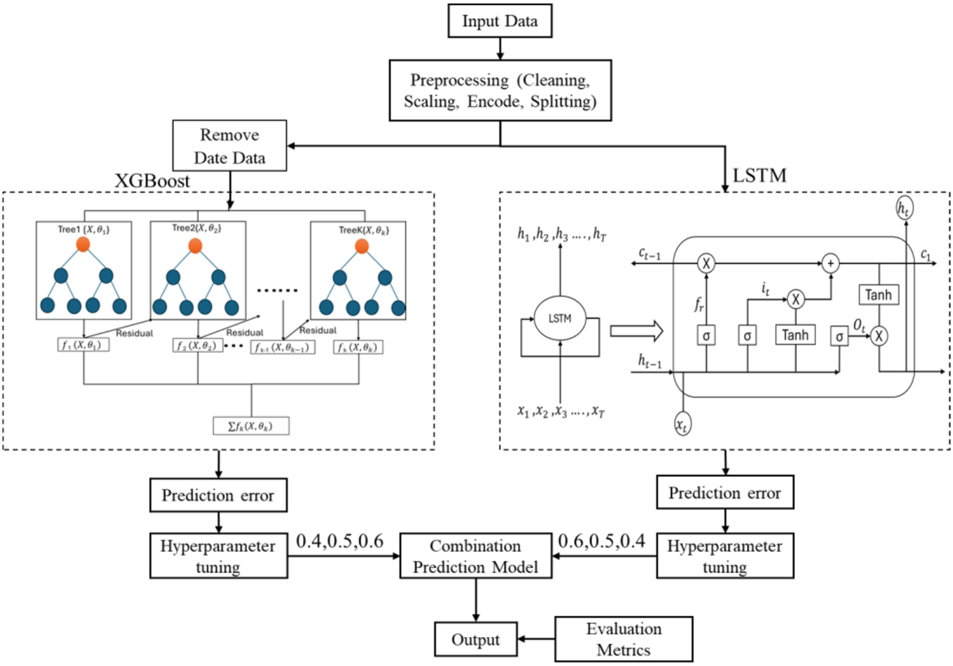

Figure 2: Hybrid XGBoost–LSTM workflow.

3.6 Evaluation Metrics

Several evaluation metrics were employed to assess the performance and accuracy of the predictive models developed in this study. These metrics provide quantitative insights into how closely the model's predictions align with actual observations, particularly for soil nutrient levels (N, P, and K). The following four metrics were used: MAE, MSE, RMSE, and R².

MAE calculates the average absolute difference between the actual and predicted values. It is a linear score, which means all individual differences are weighted equally in the average.

MAE: $\left(\frac{1}{n}\right) \Sigma \; |y_i - \hat{y}_i|$ (3)

Where:

- n = total number of observations
- y_i= actual value for the i^{th} observation
- $\hat{y}_i$= predicted value for the i^{th} observation

MSE measures the average of the squared differences between the actual and predicted values. By squaring the errors, MSE penalizes larger deviations more than smaller ones.

$$\text{MSE: } \left(\frac{1}{n}\right) \Sigma\ (y_i - \hat{y}_i)^2 \tag{4}$$

RMSE is the square root of the MSE and provides an error metric in the same unit as the original data. It gives a more interpretable scale of the average prediction error.

$$\text{RMSE: } \sqrt{MSE} \tag{5}$$

R^2 represents the proportion of variance in the actual values that is explained by the predictions. An R^2 close to 1 indicates that the model explains most of the variability in the response data.

$$\text{R}^2\text{: } 1 - \left[\frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})\wedge 2}\right] \tag{6}$$

Where:

- $\bar{y}$ = mean of actual values

The numerator represents the Residual Sum of Squares (RSS)
The denominator represents the Total Sum of Squares (TSS)
Together, these metrics provide a comprehensive evaluation framework for assessing the effectiveness of the machine learning models in predicting soil nutrient levels across different phenological stages.

4. Results and Discussion

4.1. Data Acquisition

The analysis of environmental and soil parameters from Orchards A and B revealed key variations in nutrient distribution that are essential for optimising fertiliser application. Figure 3 illustrates the N distribution, where Orchard A displays a wider spread, indicating inconsistent N availability. In contrast, Orchard B shows a narrower distribution, suggesting more uniform N content. The N levels in Orchard A showed a wider distribution, indicating inconsistent availability, whereas Orchard B exhibited a narrower range, suggesting a more uniform N content. This variation is critical for determining stage-specific nitrogen requirements.

P and K levels also differed between orchards. Orchard A had higher and more variable P and K contents, whereas Orchard B had lower and more consistent concentrations. These differences imply that fertiliser application rates must be customised for each orchard based on nutrient variability.

The soil pH in Orchard A generally higher, indicating a more alkaline environment that may influence nutrient absorption. Orchard B exhibited greater variability in EC, reflecting fluctuating salinity levels that could affect nutrient availability. Differences in temperature and moisture patterns were also observed, which affected soil water retention and fertiliser efficiency.

Rainfall varied significantly, with Orchard A receiving more precipitation, which influenced the timing and dosage of water-soluble fertiliser application. Furthermore, the phenological stages of mango trees differed; Orchard A showed a balanced growth stage distribution, whereas Orchard B had a concentration at specific stages, affecting nutrient demands at different times.

These site-specific variations highlight the need for tailored fertilisation strategies in the region. Understanding spatial and temporal nutrient dynamics supports the implementation of precision agriculture techniques. The hybrid XGBoost and LSTM model proposed in this study is well suited for capturing these complex interactions, enabling accurate and adaptive NPK fertiliser recommendations for improved mango yield and sustainability.

These insights underscore the importance of tailored fertilisation strategies for each orchard. By capturing these complex spatial-temporal dynamics, the proposed hybrid XGBoost-LSTM model (refer to Algorithm 1) offers a robust framework for precise and adaptive NPK fertiliser recommendations, thereby supporting sustainable mango cultivation in varying orchard conditions.

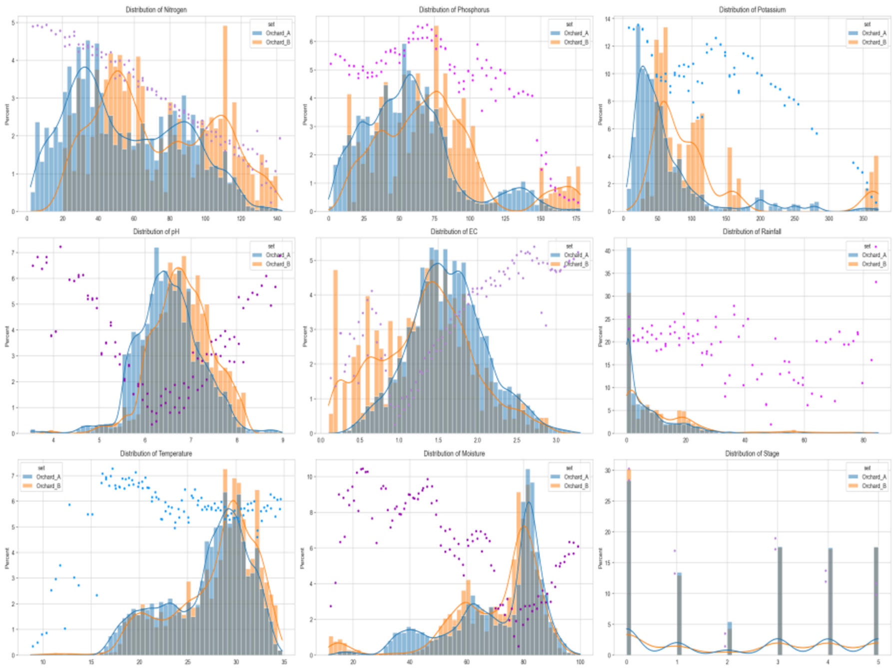

Figure 3. Comparative analysis of Orchard A and Orchard B features.

Figure 4 shows mean concentrations of N, P, and K across 50 samples for VF1, VF2,VF3, Flower, Fruit Development and Pruning stage. VF1 displays the most irregular nutrient trends, especially for nitrogen, with sharp fluctuations. K shows moderate variation and increases during the flowering and fruit development stages, reflecting the crop's high K demand. VF2 exhibits improved consistency, particularly in K, which follows a clearer pattern and peaks during fruiting, aligning with physiological needs. N and P are moderately stable, suggesting better nutrient scheduling than VF1. VF3 demonstrates the most structured nutrient profile. N and K are applied in a synchronized, timely manner, with high K levels during flowering and fruiting, and elevated N and P during the pruning stage to support new branch growth. Phosphorus remains consistently low. In summary, VF3 reflects optimized, phenology-based fertilization, VF2 shows moderate control, while VF1 lacks consistency in nutrient application.

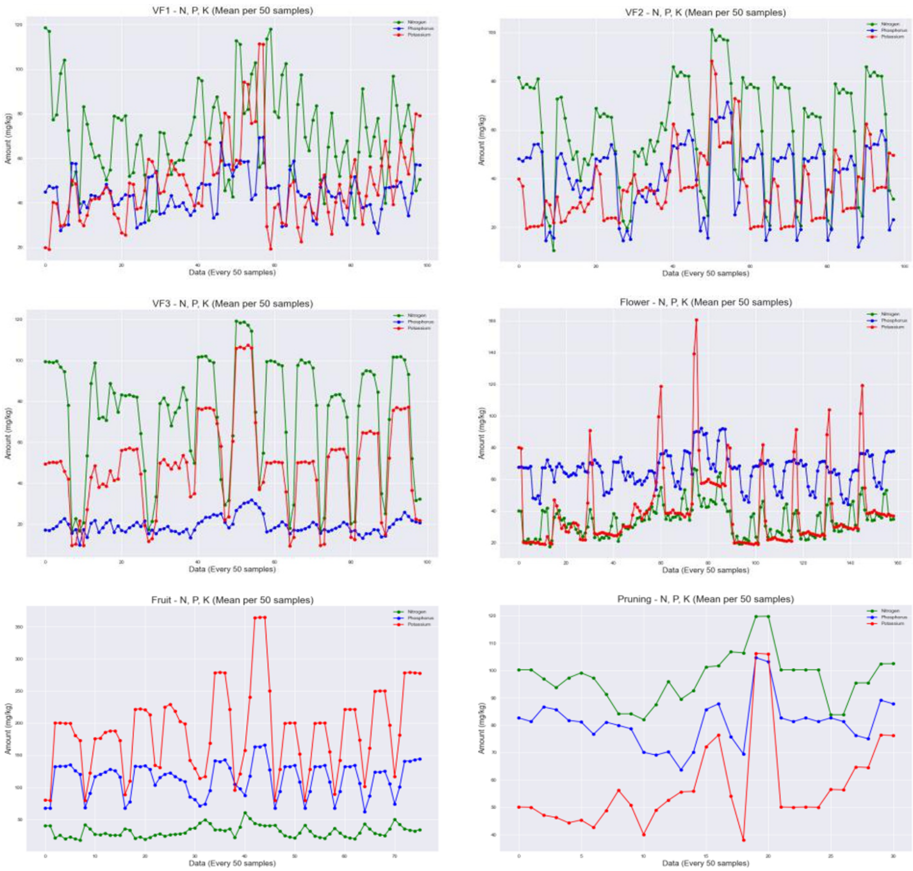

Figure 4: NPK levels across different phenological stages.

4.2. Pearson Correlation analysis

In this study, ten critical variables were incorporated as input features in the prediction model. These include N, P, K, pH, EC, Soil Temperature, Soil Moisture, Rainfall, Phenological Stage, and Yara fertiliser application. Pearson's correlation analysis was conducted to determine using an equation (5), linear relationships among these variables. The Pearson correlation coefficient (denoted as r) quantifies the degree of linear association between two variables and ranges from −1 to 1, where values close to 1 indicate a strong positive relationship, values near −1 suggest a strong negative relationship, and values around 0 imply no linear correlation between the two variables. The Pearson correlation coefficient is defined as (7)

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\Sigma(x_i - \bar{x})^2 * \Sigma(y_i - \bar{y})^2]}} \tag{7}$$

Where x and y are the variables being compared, and $\bar{x}$ and $\bar{y}$ represent their respective means.

The resulting correlation matrix is shown in Figure 5. Phosphorus and Potassium exhibited a strong positive correlation (r = 0.69), indicating that these two nutrients tended to increase together in the soil environment. In contrast, nitrogen showed a strong negative correlation with Yara fertiliser application (r = −0.70), suggesting an inverse relationship, possibly due to higher nitrogen uptake after Yara application.

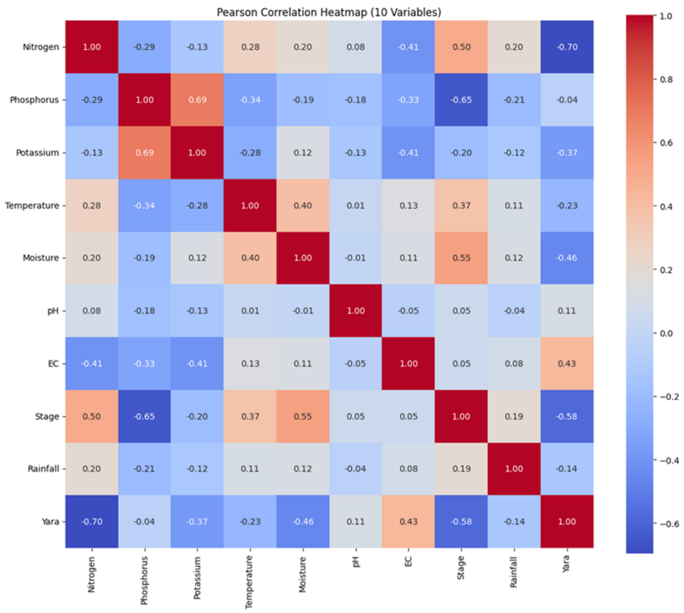

Figure 5: Pearson correlation heatmap.

The phenological Stage was positively correlated with nitrogen (r = 0.50) and moisture (r = 0.55), implying that nutrient demand and soil water content increased during active growth periods. Additionally, Stage was negatively correlated with phosphorus (r = −0.65) and Yara (r = −0.58), indicating a shift in nutrient dynamics across different growth stages. Moderate correlations were also observed between Moisture and Temperature ($r = 0.40$) and Moisture and Stage, suggesting seasonal or phenological dependencies. Notably, EC exhibited weak to moderate negative correlations with Nitrogen, Phosphorus, and Potassium, possibly indicating ion competition or salinity effects on nutrient transport.

Although some features showed weak correlations (pH vs. other variables), they were retained in the model to capture potential nonlinear effects that were not captured by the linear correlation alone. Correlation analysis provided essential insights into the interdependence of soil parameters and environmental factors affecting nutrient availability. This analysis serves as the foundation for feature selection and model interpretation in subsequent predictive modelling tasks.

4.3. Model Performance Results

The performance of the hybrid XGBoost and LSTM model was compared with the individual models, LSTM and XGBoost, based on key evaluation metrics, including the Mean MSE, MAE, and R^2 score. The results in Table 4 show of this comparison demonstrate the effectiveness of the hybrid model in improving prediction accuracy.

Table 4: Performance Evaluation of XGBoost, LSTM, and Hybrid Models (Before and After Hyperparameter Tuning) for NPK Fertiliser Prediction

Model	MAE	MSE	RMSE	R^2	MAPE
XGBoost	12.76	374.20	19.34	0.9367	3.30
XGBoost (Tuned)	9.77	203.83	14.27	0.9655	2.51
LSTM	12.00	248.60	15.76	0.9579	3.05
LSTM (Tuned)	8.77	133.12	11.53	0.9774	2.22
Hybrid XGBoost-LSTM	10.84	207.83	14.41	0.9648	2.77
Hybrid XGBoost-LSTM (Tuned)	7.59	104.23	10.21	0.9823	1.94

The hybrid XGBoost-LSTM model exhibited superior performance compared to the individual models across all nutrients. The model captured both the current nutrient status and nutrient decline over time, providing more accurate predictions of the required fertiliser additions. The following figures illustrate the comparison between the actual and predicted Yara fertiliser values for each model. The tuned hybrid model exhibited the closest fit to the ground truth.

Figure 6 shows the performance of XGBoost, XGBoost Tuned, LSTM, LSTM Tuned, Hybrid Model, and Hybrid model tuned. The plotin Figure 6 compares the actual values of the added NPK fertiliser (Yara Value) with the predicted values generated by the hybrid model. The hybrid architecture combines 40% of the predictions from the XGBoost model and 60% from the LSTM model to effectively handle both tabular feature relationships and temporal dependencies in the dataset. The superior performance of the hybrid XGBoost–LSTM model is consistent with findings from previous studies that emphasise the advantages of combining machine learning and deep learning techniques for agricultural prediction tasks. For instance, hybrid models have been shown to outperform standalone models by effectively capturing both nonlinear feature interactions and temporal dependencies in environmental datasets [15,23]. The results of this study further confirm that integrating tree-based algorithms with sequence learning models can significantly enhance prediction accuracy in complex agricultural systems

As shown in the hybrid-tuned prediction Figure 6, the predicted values closely followed the actual trend across all sample indices, indicating that the hybrid model was capable of capturing both high and low variation patterns in the NPK values. The narrow error bands and overlapping trajectories demonstrate the robustness and generalisation ability of the model.

This strong predictive performance confirms the advantage of integrating the strengths of both models: XGBoost for capturing nonlinear feature interactions and LSTM for learning time-dependent sequences, resulting in a more accurate and reliable prediction of fertiliser requirements for Harumanis mango cultivation. Furthermore, the findings align with recent research in precision agriculture that highlights the importance of site-specific nutrient management and data-driven decision-making [6,7]. By accurately predicting fertiliser requirements, the proposed hybrid model contributes to reducing excessive fertiliser application, mitigating environmental risks such as soil degradation and nutrient leaching, while improving crop productivity.

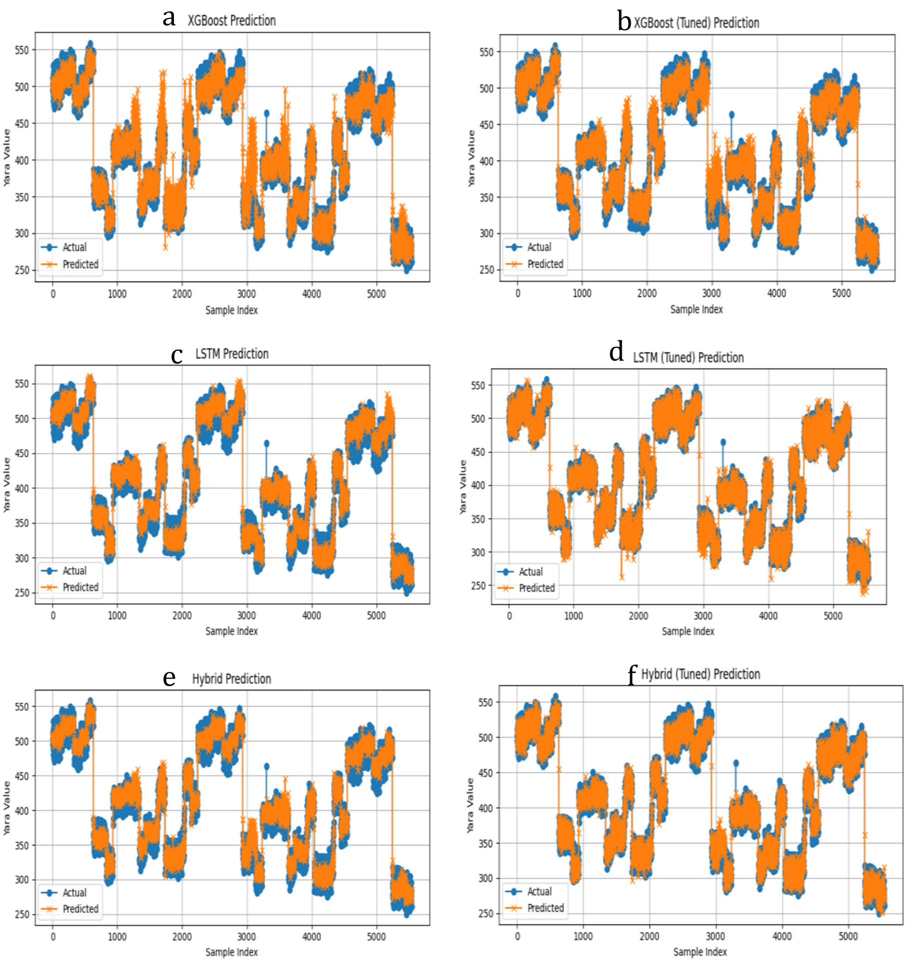

Figure 6: Comparison of prediction performance for all six models. (a)XGBoost Model. (b) XGBoost-tuned model. (c)LSTM Model. (d)LSTM Tuned model (e)Hybrid XGBoost-LSTM Model. (f) Hybrid XGBoost-LSTM Tuned model (40%-60%). Six line charts comparing actual and predicted fertilizer values generated by XGBoost, tuned XGBoost, LSTM, tuned LSTM, hybrid XGBoost–LSTM, and tuned hybrid XGBoost–LSTM models. The tuned hybrid model shows the closest agreement with the actual values, demonstrating the highest prediction accuracy among all evaluated models

5. Conclusion

This study proposed a hybrid XGBoost–LSTM model for predicting the required amount of NPK (Nitrogen, Phosphorus, and Potassium) fertiliser in mango cultivation, leveraging both structured environmental data and time-series nutrient patterns. By integrating the feature ranking and predictive power of XGBoost with the temporal modelling capabilities of LSTM, the hybrid approach achieved superior performance compared with standalone models.

The model demonstrated strong predictive accuracy, as evidenced by the improved evaluation metrics, such as MAE, RMSE, R^2, and MAPE, after tuning. These results highlight the capacity of the hybrid model to capture complex nonlinear relationships and sequential dependencies in agricultural data, offering a more precise and data-driven alternative to traditional fertiliser management methods.

This study contributes to the growing field of precision agriculture, aiming to minimise fertiliser waste, reduce environmental impact, and enhance crop yield through smart prediction systems. It also addresses a significant gap in the existing research by applying a hybrid machine learning–deep learning framework specifically tailored to perennial crops, such as mangoes, where nutrient requirements vary across phenological stages.

Future work will focus on real-time deployment using edge computing and IoT integration, enabling on-field applications that support sustainable and adaptive fertiliser practices. As agriculture continues to digitalise, these models represent a critical step forward in ensuring both productivity and environmental sustainability.

References

[1] V. Gawande, M. Chowdhury, B. V. Singh, G. N. Swamy, B. S. Sumithra, D. R. K. Saikanth, S. A. Aravind, Potential of precision farming technologies for eco-friendly agriculture, Int. J. Plant Soil Sci., 35 (19) (2023), pp. 101–112, 10.9734/ijpss/2023/v35i193528.

[2] C. Musanase, A. Vodacek, A. Uwitonze, I. Kabandana, D. Hanyurwimfura, Data-driven analysis and machine learning-based crop and fertilizer recommendation system for revolutionizing farming practices, Agriculture, 13 (11) (2023), 2141, 10.3390/agriculture13112141.

[3] R. P. Sishodia, R. L. Ray, S. K. Singh, Applications of remote sensing in precision agriculture: A review, Remote Sens., 12 (19) (2020), 3136, 10.3390/rs12193136.

[4] W. Widodo, DC resistivity and electromagnetic induction techniques for soil characterization in agriculture land, J. Environ. Eng. Geophys., 28 (4) (2023), pp. 197–208, 10.32389/jeeg23-004.

[5] R. K. Murthy, B. Nagaraju, K. Govinda, S. N. Uday Kumar, P. K. Basavaraja, H. M. Saqeebulla, G. V. Gangamrutha, S. Srivastava, P. Dey, Soil test crop response nutrient prescription equations for improving soil health and yield sustainability, Front. Plant Sci., 15 (2024), 10.3389/fpls.2024.1439523.

[6] H. H. Sarma, N. Talukdar, R. Chintey, S. K. Borah, H. Nath, Site specific nutrient management (SSNM): Principles and role in climate-resilient farming, J. Adv. Biol. Biotechnol., 27 (8) (2024), pp. 211–222, 10.9734/jabb/2024/v27i81133.

[7] Y. Xing, X. Wang, Precise application of water and fertilizer to crops: Challenges and opportunities, Front. Plant Sci., 15 (2024), 10.3389/fpls.2024.1444560.

[8] S. Maseko, H. Otterman, M. Delport, E. H. Tesfamariam, M. Van Der Laan, Evaluating machine learning models for spatial maize yield prediction, Eur. J. Agron., 157 (2024), 127193, 10.1016/j.eja.2024.127193.

[9] P. K. Kollu, M. L. Bangare, P. V. Hari Prasad, P. M. Bangare, K. P. Rane, J. L. Arias-Gonzáles, S. Lalar, M. Shabaz, IoT-driven multilinear regression technique for fertilizer recommendation, SN Appl. Sci., 5 (10) (2023), 10.1007/s42452-023-05484-8.

[10] I. Malashin, V. Tynchenko, A. Gantimurov, V. Nelyub, A. Borodulin, Applications of long short-term memory networks in polymeric sciences: A review, Polymers, 16 (18) (2024), 2607, 10.3390/polym16182607.

[11] T. S. T. Tanaka, G. B. M. Heuvelink, T. Mieno, D. S. Bullock, Can machine learning models provide accurate fertilizer recommendations?, Precis. Agric., 25 (4) (2024), pp. 1839–1856, 10.1007/s11119-024-10136-x.

[12] Y. Wu, X. Sun, B. Wang, L. Sun, A hybrid XGBoost-ISSA-LSTM model for dissolved oxygen prediction, Environ. Sci. Pollut. Res., 29 (12) (2021), pp. 18142–18159, 10.1007/s11356-021-17020-5.

[13] J. Xu, Z. Li, X. Li, Z. Li, Z. Wang, Prediction of daily climate using long short-term memory model, Int. J. Innov. Sci. Res. Technol., (2024), pp. 83–90, 10.38124/ijisrt/ijisrt24jul073.

[14] L. Xiao, Z. Luo, P. Zhang, Y. Wei, G. Wang, H. Zhou, S. Liu, Y. Zhu, Z. Shi, H. Zhang, J. Chang, E. Wang, Spatiotemporal co-optimization of agricultural

practices towards climate-smart production, Nat. Food, 5 (1) (2024), pp. 59–71, 10.1038/s43016-023-00891-x.

[15] A. Frifra, M. Maanan, M. Maanan, H. Rhinane, Harnessing LSTM and XGBoost algorithms for storm prediction, Sci. Rep., 14 (1) (2024), 10.1038/s41598-024-62182-0.

[16] F. Xia, D. Jiang, L. Kong, Y. Zhou, J. Wei, D. Ding, Y. Chen, G. Wang, S. Deng, Prediction of groundwater contaminants using XGBoost and LSTM, Int. J. Environ. Res. Public Health, 19 (15) (2022), 9374, 10.3390/ijerph19159374.

[17] F. Reuß, I. Greimeister-Pfeil, M. Vreugdenhil, W. Wagner, Comparison of LSTM and random forest for crop classification, Remote Sens., 13 (24) (2021), 5000, 10.3390/rs13245000.

[18] D. R. I. M. Setiadi, K. Nugroho, A. A. Ojugo, H.-S. Gan, A. Susanto, A. R. Muslikh, Rice yield forecasting using hybrid quantum deep learning model, Computers, 13 (8) (2024), 191, 10.3390/computers13080191.

[19] C. Sun, H. Si, M. Pei, S. Chang, B. Cao, Agricultural commodity price prediction using LSTM model, Agriculture, 14 (1) (2023), 60, 10.3390/agriculture14010060.

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

[21] Dikmen, F., Demir, A., Özkaya, B., Raza, M. O., Rasheed, J., Asuroglu, T., & Alsubai, S. (2025). AI-driven wastewater management through comparative analysis of feature selection techniques and predictive models. *Scientific Reports*, *15*(1), 25347. https://doi.org/10.1038/s41598-025-07124-0

[22] Duan, X. (2024). Assessment of resilient modulus of soil using hybrid extreme gradient boosting models. *Scientific Reports*, *14*(1), 31706. https://doi.org/10.1038/s41598-024-81311-3.

[23] Venkateswara, S. M., & Padmanaban, J. (2025). Interpretable deep learning models for independent fertilizer and crop recommendation. *Scientific Reports*, *15*(1), 41721. https://doi.org/10.1038/s41598-025-26910-4.