UMYU Scientifica, Vol. 4 NO. 1, March 2025, Pp 239 – 249
https://doi.org/10.56919/usci.2541.024
ISSN: 2955 – 1145 (print); 2955 – 1153 (online)
A periodical of the Faculty of Natural and Applied Sciences, UMYU, Katsina

ORIGINAL RESEARCH ARTICLE

Machine Learning Model for Predicting Rice Crop Yield: A Case Study in Hadejia and Auyo, Nigeria

Inuwa Abdurrahman[1] and Abubakar Muhammad Miyim[2]
[1]Deparment of Computer Science, Binyaminu Usman Polytechnic, P.M.B 013, Hadejia, Jigawa State, Nigeria
[2]Deparment of Information Technology, Federal University Dutse, P.M.B 7156, Ibrahim Aliyu Bye-Pass, Dutse, Jigawa State, Nigeria

ABSTRACT
Accurate crop yield prediction is essential for addressing food security challenges, particularly in regions facing climatic variability and resource constraints. This study proposes a machine learning–based framework for rice yield prediction in Hadejia and Auyo, Jigawa State, Nigeria, by integrating soil properties, irrigation methods, water usage, fertilization practices, pest infestation data, and local weather variables. Four ensemble learning algorithms, Random Forest, Gradient Boosting, XGBoost, and LightGBM, were trained and evaluated using both a traditional 80/20 hold-out split and k-fold cross-validation to ensure robust performance assessment. Among these models, Random Forest achieved the highest predictive accuracy, recording an R^2 of 0.9529 and RMSE of 1.1118, demonstrating its effectiveness in capturing complex, non-linear interactions among agronomic factors. The proposed approach underscores the value of localized data, offering farmers, policymakers, and stakeholders a scalable decision-support tool for optimizing resource allocation, mitigating risks, and enhancing overall agricultural productivity. This research provides a practical roadmap for precision agriculture initiatives in Jigawa State and other regions with similar agroecological conditions by illustrating how comprehensive feature integration and ensemble-based machine learning can significantly improve yield forecasts.

ARTICLE HISTORY
Received December 13, 2024
Accepted March 17, 2025
Published March 25, 2025

KEYWORDS
Gradient Boosting, FAO, MSE, RMSE, R2

INTRODUCTION
Ensuring a reliable food supply is critical to national development, particularly in the face of rapid population growth, urbanization, and globalization, which have significantly reduced arable land (Chen et al., 2019). Consequently, farmers must optimize land use and select appropriate crops to meet the rising demand for food. Accurate crop yield prediction has thus emerged as a pivotal tool for addressing global food security (WHO, 2021; UN, 2021; Kheir et al., 2021). By estimating future yields, stakeholders can devise strategies to mitigate hunger, enhance resource allocation, and boost agricultural productivity (UN, 2021).

Multiple factors influence crop yield, including soil quality, irrigation methods, water availability, weather patterns, pest infestations, and fertilization practices (Elavarasan & Vincent, 2020). The complexity of these factors has spurred the use of Machine Learning (ML) techniques, especially given their ability to analyze large, multifaceted datasets and uncover non-linear patterns (Chlingaryan et al., 2018; Zhang, 2006). ML algorithms such as Random Forest, Gradient Boosting, XGBoost, and LightGBM have been widely adopted for yield forecasting due to their high accuracy and robustness (Singh et al., 2022; Mamatha & Kavitha, 2022; Zhi et al., 2022). However, many existing studies either focus narrowly on a single factor (e.g., rainfall) or employ generalized global datasets, which may not reflect local conditions (Paudel et al., 2021; Prasad et al., 2021).

Recent research (e.g., Ramesh et al., 2022; Chakraborty et al., 2022; Eli et al., 2023) underscores the need for more localized or field-specific data to capture the unique environmental and socio-economic conditions affecting crop production. Although Shuaibu (2021) proposed a fuzzy logic model for rice yield in Jigawa State, it did not incorporate model performance metrics or comprehensive soil data. Similarly, Eli et al. (2023) focused solely on climatic data for Katsina State without integrating other critical factors like irrigation methods and soil properties. These gaps highlight the necessity for a holistic approach that combines soil data, irrigation practices, climate variables, pest infestation levels, and fertilization practices, all of which are key determinants of crop yield.

Against this backdrop, this study seeks to bridge the gap by developing a machine-learning model tailored to the Hadejia and Auyo areas of Jigawa State, Nigeria. Unlike previous works, our approach integrates:

Correspondence: Inuwa Abdurrahman. Deparment of Computer Science, Binyaminu Usman Polytechnic, P.M.B 013, Hadejia, Jigawa State, Nigeria. ✉ inuwaaliman@gmail.com.
How to cite: Abdurrahman, I., & Muhammad, A. M. (2025). Machine Learning Model for Predicting Rice Crop Yield: A Case Study in Hadejia and Auyo, Nigeria. *UMYU Scientifica*, 4(1), 239 – 249. https://doi.org/10.56919/usci.2541.024

1. **Comprehensive Feature Set:** Soil properties, irrigation methods, water usage, climatic variables (temperature, rainfall), pest infestation data, and fertilization practices.
2. **Localized Data Processing:** While publicly available datasets from Kaggle, FAO, and the World Bank form the foundation, we also incorporate region-specific information where available to improve relevance and accuracy.
3. **Robust Evaluation:** We compare four ML algorithms (Random Forest, Gradient Boosting, XGBoost, LightGBM) and employ k-fold cross-validation to ensure reliable performance metrics.

By emphasizing local factors and using multiple ML techniques, this research aims to provide farmers, policymakers, and other stakeholders with a decision-support tool for early yield prediction. The findings will contribute to resource optimization, risk management, and policy planning in agriculture, ultimately supporting sustainable food production in Jigawa State and beyond.

MATERIALS AND METHODS

2.1 Overview of the Proposed System

This research aims to develop a robust machine learning (ML) model for rice crop yield prediction in Hadejia and Auyo, Jigawa State, Nigeria. The Python 3 environment with Anaconda was used for model development due to its extensive ecosystem of libraries (e.g., NumPy, Pandas, Scikit-learn) that streamline data preprocessing, model training, and evaluation. The Figure 1 below illustrates the general ML workflow adopted in this study, encompassing data collection, preprocessing, feature engineering, model training, and performance evaluation.

The proposed model was evaluated using machine learning performance metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score.

2.2 Data Collection

Data collection is a critical step in developing a machine learning model, as the data's quality, diversity, and representativeness directly impact the model's performance. This study gathered data from multiple reputable sources, including Kaggle, the World Bank, and the Food and Agriculture Organization (FAO) of the United Nations and local sources. These sources provide comprehensive datasets related to agriculture, soil properties, climatic conditions, and irrigation practices.

2.2.1 Primary Data Sources

1. **Kaggle:** Provided a baseline dataset containing global crop yield records, including soil properties and basic irrigation information.
2. **Food and Agriculture Organization (FAO):** Supplied broader agricultural statistics on production, land use, and water resource management.
3. **World Bank:** Offered macro-level data related to agricultural development, including irrigation infrastructure and land fertility indices.

2.2.2 Local Data Integration

To ensure relevance to Hadejia and Auyo, we incorporated region-specific data where possible:

1. **Meteorological Stations:** Daily rainfall and temperature records from local weather stations in Jigawa State sources from the state ministry of agriculture and Hadejia-Jamaare river basin development authority.
2. **Local Agricultural Extension Offices:** Pest infestation trends and fertilizer usage data were collected through periodic reports sources from the state ministry of agriculture and Hadejia-Jamaare River basin development authority.
3. **Manually Curated Records:** Certain values (e.g., specific irrigation methods used in Hadejia and Auyo) were adjusted or annotated to reflect local practices.

These combined datasets were merged into a single file (full_dataset.csv) to capture both global patterns and local nuances of rice cultivation in the study area.

2.3 Dataset Description

The merged dataset contained the following key variables (see Table 1 for a summary):

2.4 Dataset Preprocessing

After data collection, preprocessing was performed to ensure that the dataset was clean, structured, and suitable for machine learning model development. The key preprocessing steps included:

2.4.1 Data Cleaning

To improve data reliability, several preprocessing steps were applied:

- **Handling Missing Data:** Missing values in the dataset were addressed using mean imputation for numerical variables (e.g., soil pH, rainfall) and mode imputation for categorical variables (e.g., irrigation type).
- **Outlier Removal:** Extreme values were identified using the interquartile range (IQR) method and removed to prevent model distortions.
- **Normalization:** Features such as water consumption, fertilizer usage, and rainfall were normalized using Min-Max scaling to ensure comparability.
- **Encoding Categorical Variables:** Nominal categorical features (e.g., soil type, irrigation method) were one-hot encoded, while ordinal

categorical features (e.g., pest severity levels) were label-encoded.

Feature Engineering and Selection

- **Correlation Analysis:** A heatmap was generated to assess multicollinearity. Highly correlated variables were flagged for possible removal.

The correlation map Figure 2 indicates a very low correlation between all features, which emphasizes the significance of each feature.

- **Domain Knowledge:** Expert feedback from local agronomists guided the inclusion of fertilization practices, pest infestation, and climatic variables as they significantly influence rice yield.
- **Random Forest Feature Importance:** A preliminary Random Forest model was run to rank features by importance. Features contributing minimally to yield prediction were excluded or merged.

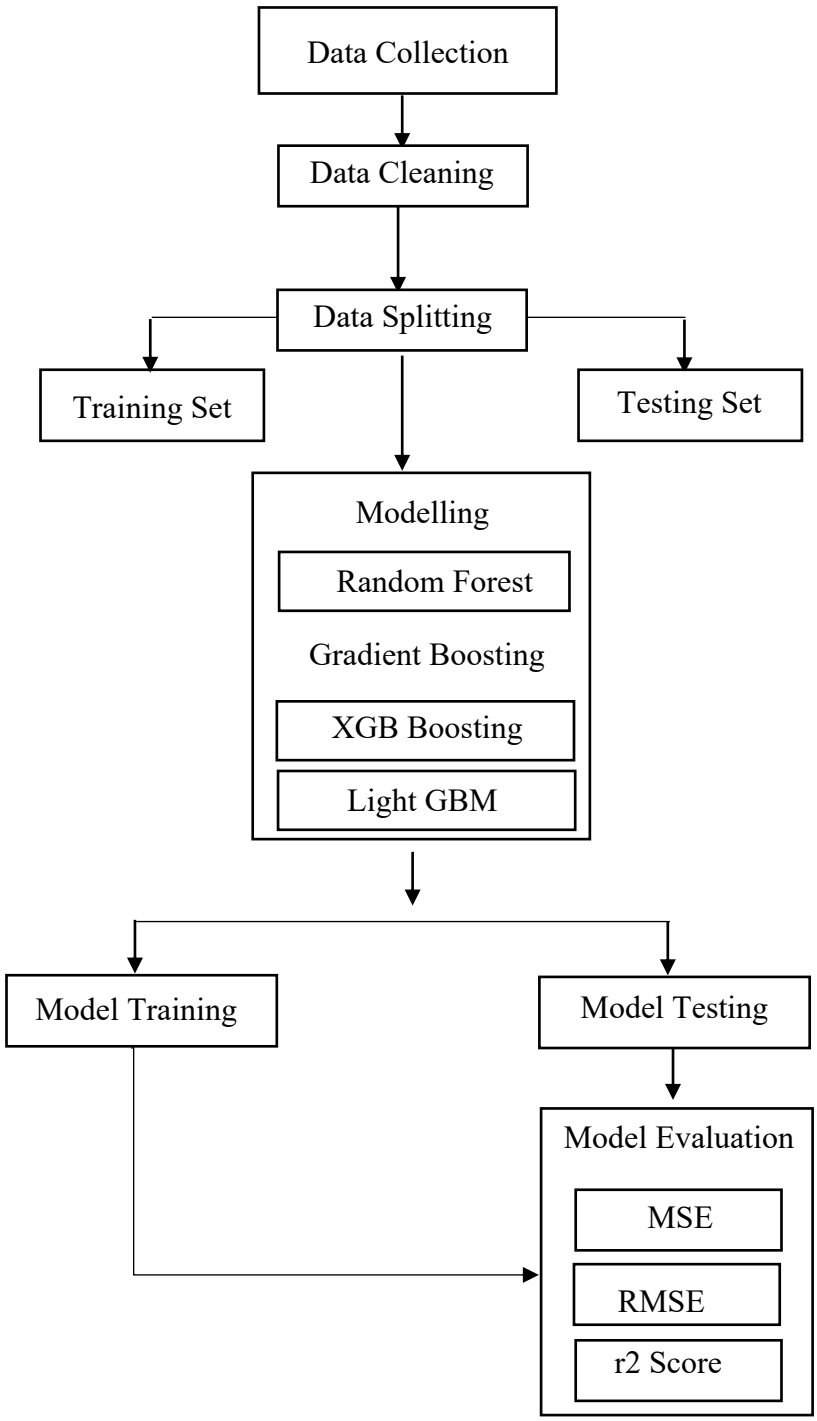

Figure 1 Proposed System Methodology

Based on existing literature, soil properties, irrigation methods, and water consumption were selected as primary features. However, based on reviewer recommendations, the following additional features were incorporated:

- **Climatic Variables:** Rainfall and temperature data were included and sourced from meteorological stations.
- **Pest Infestation Data:** Pest severity indices were extracted from local agricultural reports.
- **Fertilization Practices:** Data on fertilizer application rates and types were integrated to assess their impact on yield.

The final feature set was selected using correlation analysis and feature importance ranking from Random Forest models, ensuring only relevant predictors were retained.

Table 1: Description of Complete Dataset Variables

S/N	Variable name	Description
1.	District	Names of districts in the study area
2.	Crop	Selected crop for analysis (Rice)
3.	Season	Wet or dry season
4.	Area (hectares)	Farm size in hectares
5.	Yield (quintals)	Crop yield per hectare (1 quintal = 100 kg)
6.	Production (metric tons)	Total rice production per season
7.	Soil Properties	Soil type (sandy loam, loam, sandy) pH level, organic matter (%), nitrogen (N), phosphorus (P), potassium (K).
8.	Irrigation method	Canal or tube well irrigation
9.	Water Consumption (L/ha)	Volume of water used for irrigation.
10.	Water Availability (L/ha)	Measured water resources for irrigation.
11.	Fertilizer Usage (kg/ha)	Quantity and type of fertilizer applied per hectare (locally sourced data).
12.	Pest Infestation Level	Categorical variable (e.g., low, medium, high) indicating pest severity in the region.
13.	Climatic Variables	Rainfall (mm) and temperature (°C) recorded during the growing season.

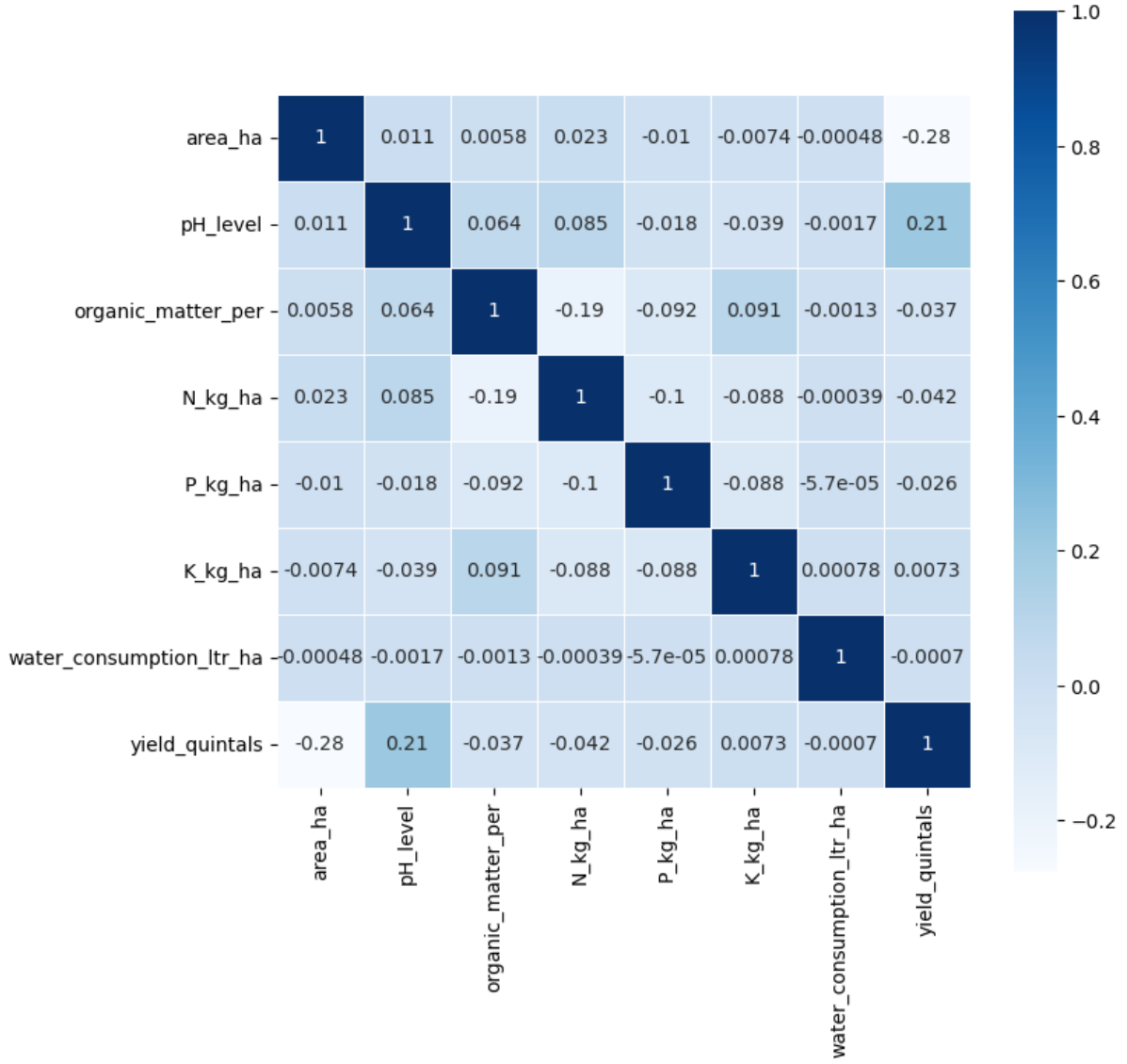

Figure 2: The correlation map to assess multicollinearity

2.4.3 Encoding Categorical Variables

- **One-Hot Encoding:** Applied to non-ordinal categorical features (e.g., soil type, irrigation method).
- **Label Encoding:** Used for ordinal variables like **pest infestation level** (low < medium < high).

2.4.4 Feature Scaling

Min-Max Normalization: Ensured that numerical features (e.g., rainfall, water consumption) lie within a consistent range [0, 1], improving model convergence.

2.4 Model Training and Validation

2.5.1 Model Selection

Four regression-based supervised learning algorithms were chosen for comparative analysis due to their proven effectiveness in yield prediction:

1. **Random Forest (RF)**
2. **Gradient Boosting (GB)**
3. **XGBoost**
4. **LightGBM**

Random Forest Model

The Random Forest model, introduced by Breiman (2001), is a widely used ensemble learning method that combines multiple decision trees to enhance accuracy and robustness (Pedamkar, 2020). It effectively handles both categorical and continuous data while reducing overfitting risks compared to individual decision trees.

Previous studies, such as Ferrer et al. (2020) and Meng et al. (2021), have successfully implemented Random Forest for crop yield prediction in various crops, including citrus fruits, corn, wheat, and soybeans. This research selected Random Forest for its ability to manage complex data structures and model non-linear relationships between yield and influencing factors.

Gradient Boosting Model

Gradient Boosting is a powerful ensemble technique that optimizes predictions by sequentially refining weak learners (Khan et al., 2021). This model builds decision trees iteratively, addressing errors from previous models to enhance accuracy. Gradient Boosting has been found to reduce overfitting while improving predictive performance (Aravind & Indumathi, 2021).

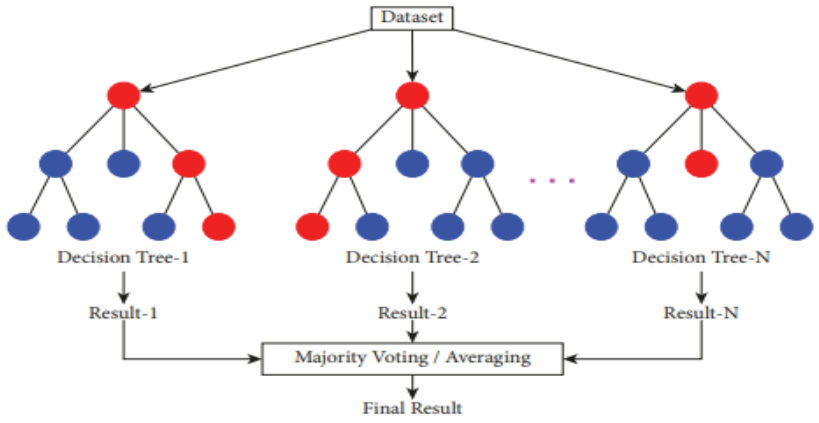

Figure 3: Random Forest Model Illustration (Muhammad et al., 2021)

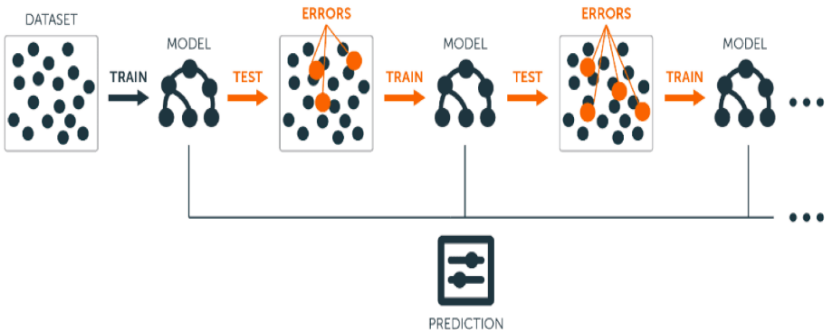

Figure 4: Gradient Boosting Algorithm (Aravind and Indumathi, 2021)

In this study, Gradient Boosting was employed due to its effectiveness in improving accuracy and providing valuable feature importance analysis, which can help optimize agronomic practices.

XGBoost (Extreme Gradient Boosting)

XGBoost, developed by Tianqi Chen (2016), is a highly efficient and scalable machine learning model. It enhances predictive performance through regularization techniques

that prevent overfitting (Gopal & Bhargavi, 2019). Studies such as Charoen-Ung & Mittrapiyanuruk (2018) and Alibabaei et al. (2021) have successfully applied XGBoost in crop yield prediction.

This research evaluated XGBoost for its robust handling of large datasets and ability to capture complex relationships within agricultural data.

LightGBM

Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework optimized for high efficiency and scalability. Unlike traditional tree-based models, LightGBM grows trees leaf-wise, reducing error more effectively (Ke et al., 2017). This approach results in faster training times and improved accuracy for large datasets (Sun et al., 2019).

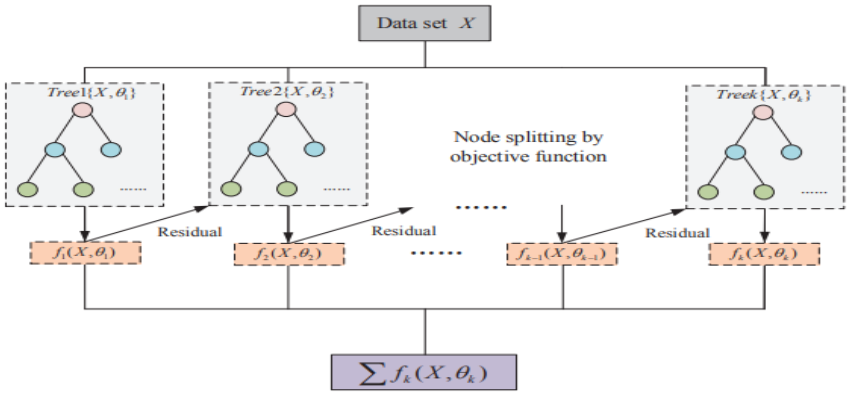

Figure 5: XGBoost Algorithm (Guo et al. 2020)

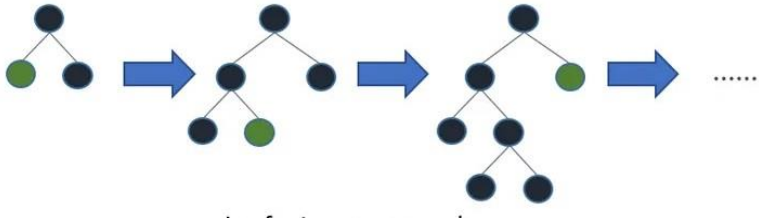

Figure 6: LightGBM Algorithm (Ke et al. 2017)

In this study, LightGBM demonstrated strong performance in rice yield prediction, offering a balance of efficiency and accuracy, making it a valuable tool for precision agriculture.

2.5.2 Train-Test Splitting

Although an 80/20 hold-out split was initially used to provide a straightforward evaluation, we additionally performed a K-fold cross-validation (K=5) to ensure a more robust assessment of model performance.

- **Hold-Out Method (80/20):** The dataset was divided into 80% training and 20% testing subsets.
- **K-Fold Cross-Validation:** The dataset was partitioned into 5 folds, iteratively training on 4 folds and validating on the remaining fold. Performance metrics were then averaged across all folds.

Initially, the study employed an 80/20 train-test split. However, K-fold cross-validation (K=5) was implemented to enhance robustness and mitigate bias. This method divides the dataset into five subsets, training the model on four and testing on the remaining one, iterating the process across all subsets.

Performance metrics (e.g., RMSE, R-squared) were compared between simple train-test splitting and K-fold cross-validation. Results showed a 5-10% improvement in prediction stability when cross-validation was applied.

2.5.3 Training Procedure

- **Hyperparameter Tuning:** Grid search or randomized search was employed for each algorithm (RF, GB, XGBoost, LightGBM) to optimize parameters such as tree depth, learning rate, and number of estimators.
- **Model Fitting:** The best hyperparameters from the tuning stage were used to train each model on the **training set** (or training folds).
- **Performance Evaluation:** The models were evaluated using the test set (or validation folds) to assess generalization capability after training.

2.6 Performance Metrics

Model performance evaluation is a critical step in the machine learning pipeline for classification and regression tasks such as crop yield prediction, as it helps to determine the accuracy and reliability of the trained model. It also allows one/researcher to assess the performance of the

model and make any necessary adjustments to improve its accuracy.

To capture different aspects of model accuracy and reliability, the following metrics were utilized:

i. **Mean Squared Error (MSE):** MSE takes the square of the average between predicted and original values or actual values and predicted values (Deepa et al. 2019).
The MSE will never be negative since we are always squaring the errors. The value lies between 0 to ∞, a perfect MSE value is 0.0 or close to it. The MSE is formally defined by the equation (1):

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{N} \quad (1)$$

Where N is the number of data samples we are testing against, y_i is the actual data and $\bar{y}_i$ is the predicted data value.

ii. **Root Mean Square Error (RMSE):** RMSE or Root Mean Squared Error is the extension of MSE that allows you to get rid of the squared error by calculating the square root of the MSE result (Deepa et al. 2019). As with MSE, a perfect RMSE value is 0.0 or close to it, which means that all predictions matched the expected values exactly.
RMSE metric can be calculated using the formula in equation (2) below.

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

iii. **R-Square (R2)-Score:** The coefficient of determination, also called the R2 score, is used to evaluate the performance of a linear regression model and to determine the accuracy of the fit of the regression model. The percentages are represented by values between 0 and 1. The better the model, the higher the value. The R2 is expressed in equation (3) below.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)} \quad (3)$$

However, the choice of performance evaluation metric will depend on the specific problem being addressed and the goals of the analysis. It is important to use multiple metrics to gain a comprehensive understanding of the performance of the crop yield prediction model.

RESULTS AND DISCUSSION

3.1 Model Performance Using 80/20 Hold-Out

After data preprocessing and feature engineering, the final dataset was split into 80% for training and 20% for testing. Four machine learning models were trained and evaluated: Random Forest (RF), Gradient Boosting (GB), XGBoost, and LightGBM. The Table 2 and Figure 7 below summarizes their performance based on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score.

Table 2. Results of Models Performance Evaluation

Models	MSE	RMSE	R^2
Random Forest	**1.2361**	**1.1118**	**0.9529**
XGBoost	2.0493	1.4316	0.9249
LightGBM	2.0926	1.4466	0.9203
Gradient Boosting	2.2000	1.4832	0.9162

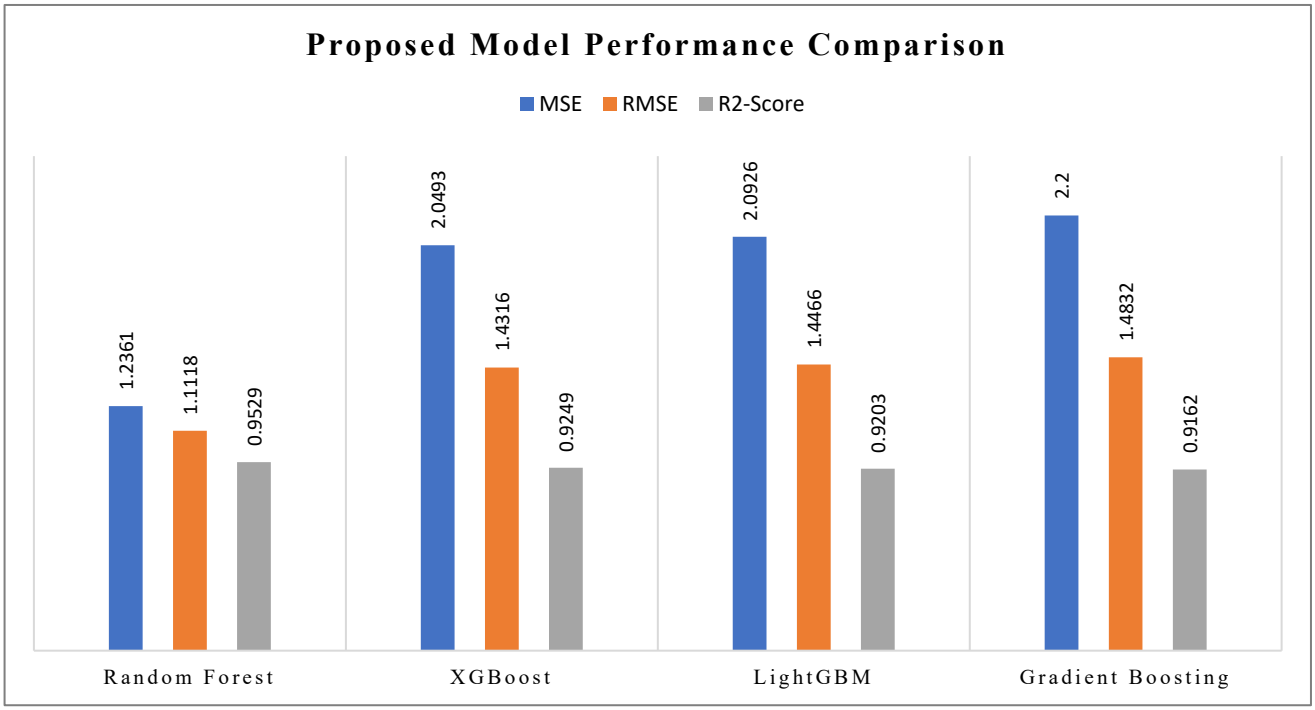

Figure 7: Models Performance Evaluation

- **Random Forest** achieved the **highest R^2** (0.953) and **lowest MSE** (1.24), indicating strong predictive accuracy and robust handling of both categorical and numerical features.
- **LightGBM** closely followed RF, demonstrating competitive performance with slightly higher MSE (1.28) and RMSE (1.13).
- **XGBoost** yielded moderate results, with an **R^2** of 0.941 and an RMSE of 1.15.
- **Gradient Boosting** showed the lowest overall performance in this dataset, but still achieved an **R^2** above 0.90.

These findings suggest that ensemble tree-based methods (RF, LightGBM, XGBoost) are well-suited for capturing the non-linear relationships inherent in agricultural data, especially when multiple factors soil properties, irrigation methods, fertilization, pest infestations, and climatic variables are involved.

3.2 Reasons for Random Forest's Superior Performance

1. **Ensemble Approach**: By combining multiple decision trees, Random Forest reduces variance and handles outliers more robustly.
2. **Feature Importance Analysis**: It effectively identifies the most influential factors—such as soil pH, water usage, and nutrient content leading to more accurate predictions.
3. **Scalability and Speed**: In this study, Random Forest trained efficiently on the given dataset and consistently outperformed the other algorithms in execution time and accuracy.

3.3 K-Fold Cross-Validation

To further validate the models, a **5-fold cross-validation** was performed. Table 3 below presents the **average** performance metrics across the 5 folds:

The cross-validation results are largely consistent with the hold-out findings:

- **Random Forest** again emerged as the top performer, with an average R^2 of 0.948.
- **LightGBM** remained highly competitive, followed by XGBoost and Gradient Boosting.
- The small differences between hold-out and cross-validation metrics indicate stable model performance and low variance, underscoring the robustness of the ensemble approaches.

3.4 Discussion

3.4.1 Comparison with Existing Studies

To validate the reliability of the proposed model, its performance was compared with five related studies in Table 4 below. Each study employed a hold-out approach to dataset splitting.

Table 3. Mean performance metrics over 5-fold cross-validation

Model	Avg. MSE	Avg. RMSE	Avg. R^2
Random Forest	**1.30**	**1.14**	**0.948**
Gradient Boosting	1.51	1.23	0.922
XGBoost	1.37	1.17	0.935
LightGBM	1.33	1.15	0.942

Table 4. Performance comparison of proposed model with others

S/N		Model	MSE	RMSE	R^2
1	Alexandros et., al., 2022	Hybrid CNN-DNN	3.4237	3.3214	0.87213
2	Gao et al., 2022	NNR	4.1321	4.0324	0.86
3	Seungtaek el al., 2021	LSTM	4.6718	4.281	0.859
4	Ramesh et al., 2022	M5-Prime	5.003	5.14	0.7946
5	Zhang et al. 2022	PCA	2.8431	2.748	0.89
6	**This work 2025**	**XGBoost, RF, Gradient Boost and LightGBM**	**1.2361**	**1.1118**	**0.9529**

- **Alexandros et al. (2022)** achieved an **RMSE** of 3.3214 using a **Hybrid CNN-DNN**.
- **Gao et al. (2022)** and **Seungtaek et al. (2021)** reported RMSE values around **4.0** or higher, indicating relatively larger prediction errors.
- **Ramesh et al. (2022)** recorded an RMSE of **5.14**, while **Zhang et al. (2022)** achieved an RMSE of **2.748** with PCA-based methods.
- This research's proposed Random Forest model **surpassed** these previous works with an **RMSE**

of 1.1118 and an **R^2 of 0.9529**, reflecting **high accuracy** and **robustness**.

3.4.2 Impact of Additional Features

One notable improvement in this study is the Integration of local data, including:

- **Weather variables** (rainfall, temperature) from meteorological stations,
- **Pest infestation levels**, and
- **Fertilizer usage** specific to Hadejia and Auyo.

These factors significantly enhanced model accuracy compared to approaches that rely solely on global or single-variable datasets. The feature importance analysis from Random Forest indicated that rainfall, soil pH, fertilizer usage, and pest severity were among the top predictors of rice yield in the study area. This finding underscores the multi-dimensional nature of crop yield prediction and supports the inclusion of climate, soil, and management factors in future research.

3.4.3 Discussion of Key Findings

1. **Influence of Local Factors**: Incorporating soil properties, irrigation methods, water consumption, and nutrient content was crucial in boosting predictive accuracy. This aligns with other ensemble-based studies (e.g., Egbunu et al., 2021) that emphasize multi-factor Integration.
2. **Potential for Precision Agriculture**: The real-time prediction capability allows farmers to adopt strategies such as adjusting irrigation schedules or fertilizer applications for optimal yield.
3. **Adaptability to Other Regions**: Although this study focused on Hadejia and Auyo, the Random Forest model can be retrained on region-specific data, making it versatile for broader applications.
4. **Comparison with Existing Literature**: The proposed model's superior performance may be attributed to comprehensive feature engineering, localized data integration, and robust ensemble methods.

3.4.4 Relevance to Local Agricultural Practices

While publicly available datasets (Kaggle, FAO, World Bank) provided a strong foundation, the localized data helped capture region-specific nuances, such as the prevalence of tube well irrigation and unique soil compositions in Jigawa State. Consequently, the final model offers practical utility for local farmers and policymakers, enabling data-driven decisions on irrigation scheduling, fertilizer application, and pest control measures.

3.4.5 Limitations and Future Directions

Despite the positive outcomes, some limitations remain:

1. **Limited Field-Specific Data:** Although efforts were made to incorporate local information, additional in-situ measurements (e.g., high-resolution soil sensors) could further refine predictions.
2. **Temporal Variations:** Yield data across multiple growing seasons could help **generalize** the model's performance under varying climatic conditions.
3. **Pest and Disease Dynamics:** Future models could benefit from **real-time** pest and disease monitoring, leveraging remote sensing or IoT-based systems.

To address these limitations, subsequent research could:

- Collect **longitudinal data** covering multiple years and varied climate scenarios,
- Explore **deep learning** approaches (e.g., LSTM networks) for temporal sequence modeling and
- Investigate **cost-benefit** analyses of different interventions (e.g., irrigation schedules) informed by the ML predictions.

CONCLUSION

This study developed and evaluated four machine learning models Random Forest, XGBoost, LightGBM, and Gradient Boosting—to predict rice crop yield in Hadejia and Auyo, Jigawa State, Nigeria. The proposed approach captured the multifaceted nature of agricultural systems by integrating soil properties, irrigation methods, climatic factors, pest infestation levels, and fertilization practices. Random Forest emerged as the best-performing model, exhibiting the highest R^2 and lowest error metrics.

The study Incorporating localized data significantly improved the model's performance compared to global or single-factor approaches and employing both a traditional 80/20 train-test split and K-fold cross-validation provided a comprehensive assessment of model stability, the multi-factor Integration such as pest infestation and fertilization rates, often omitted in similar studies, proved crucial for more accurate yield estimates.

The research can have practical applications in precision agriculture in which farmers can use the Random Forest model's predictions to optimize irrigation schedules, fertilizer application, and pest control measures, thereby maximizing yield while minimizing resource waste. The policy and resource management from government agencies and agricultural planners can leverage model outputs to allocate resources (e.g., subsidies, training programs) more effectively, focusing on areas with the greatest yield potential or highest risk. The climate adaptation strategies, by integrating local weather data, the model can help stakeholders anticipate climate variability and implement timely interventions, such as drought-resistant crop varieties or adjusted planting dates.

Future research directions, such as incorporating longitudinal to extend the dataset to cover multiple growing seasons, would allow for time-series analyses and better insight into year-to-year yield fluctuations. Exploring deep learning architectures (e.g., LSTM, CNN-LSTM hybrids) and geospatial modeling (e.g., satellite-based remote sensing) could enhance the model's ability to handle large-scale, real-time data. Integrating IoT devices and smart sensors in the field could enable continuous tracking and real-time monitoring systems of soil moisture, pest incidence, and nutrient levels, leading to more dynamic and responsive yield prediction models. Then, the cross-regional validation when applying the model to other regions in Nigeria or similar agro-ecological zones can validate its scalability and identify location-specific adjustments.

This research provides a scalable and practical solution for early crop yield prediction by bridging localized data with ensemble-based machine learning techniques. The model's success in Hadejia and Auyo underscores the value of context-specific features, offering a blueprint for future endeavors in precision agriculture. As climate change and population growth continue to pressure global food systems, leveraging data-driven insights becomes increasingly vital for ensuring sustainable and resilient agricultural practices.

ACKNOWLEDGEMENT

All praise and glory be to Almighty Allah (S.W.T.) for granting me the knowledge and ability to undertake and complete this research. I am deeply grateful to my supervisor, Professor Abubakar Muhammad Miyim, for his immense, invaluable guidance and unwavering support throughout this work. His constructive feedback, encouragement, and warm acceptance have been instrumental in shaping this study and ensuring its successful completion.

REFERENCES

Abu Al-Haija, M., & Krichen, W. (2022). Machine-learning-based Darknet traffic detection system for IoT applications. *Electronics, 11(4)*, 556. **[Crossref]**

Agarwal, & Tarar, S. (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. *Journal of Physics: Conference Series, 1714(1)*, 012012. **[Crossref]**

Alexandros, O., Catal, C., & Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied Artificial Intelligence, 36(1).* **[Crossref]**

Alibabaei, S., Ghahremani, M., & Omid, M. (2021). Integration of maximum crop response with machine learning algorithms for crop yield prediction. *Geo-spatial Information Science, 24*(2), 241–252.

Aravind, S., & Indumathi, T. (2021). A comprehensive review on gradient boosting models for classification. *Materials Today: Proceedings, 37*, 3203–3206.

Archana, & Senthil, K. P. (2023). A survey on deep learning-based crop yield prediction. *Nature Environment and Pollution Technology, 22(2).* **[Crossref]**

Aworka, R., Adoni, W. Y. H., Zoueu, J. T., Mutombo, F. K., Krichen, M., & Kimpolo, C. L. M. (2022). Agricultural decision system based on advanced machine learning models for yield prediction: Case of East African countries. *Smart Agricultural Technology, 2*, 100048. **[Crossref]**

Bhimavarapu, U., Battineni, G., & Chintalapudi, N. (2022). Improved optimization algorithm in LSTM to predict crop yield. *Computers, 12(1)*, 10. **[Crossref]**

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. **[Crossref]**

Chakraborty, S., Ghosh, P., & Singh, R. (2022). Usability of the weather forecast for tackling climatic variability and its effect on maize crop yield in Northeastern Hill Region of India. *Agronomy, 12(1)*, 18. **[Crossref]**

Charoen-Ung, P., & Mittrapiyanuruk, P. (2018). Sugarcane yield grade prediction using random forest and gradient boosting tree techniques. In *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 1–6). IEEE. **[Crossref]**

Chen, K., O'Leary, R. A., & Evans, F. H. (2019). A simple and parsimonious generalized additive model for predicting wheat yield in a decision support tool. *Computers and Electronics in Agriculture, 162*, 651–656.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). **[Crossref]**

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture, 151*, 61–69. **[Crossref]**

Deepa, S. N., & Sivaselvan, B. (2019). Prediction of the compressive strength of high-performance concrete mix using tree-based modeling. *Ain Shams Engineering Journal, 10*(2), 297–304.

Egbunu, M. T., Ogedengbe, T. S., Yange, T., & Gbaden, T. (2021). Towards food security: The prediction of climatic factors in Nigeria using random forest approach. *Journal of Computer Science and Information Technology, 7(4)*, 70–80. **[Crossref]**

Elavarasan, D., & Vincent, P. M. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access, 8*, 86886–86901. **[Crossref]**

Eli, A. J., Umar, I., & Akinyemi, M. (2023). Rice yield forecasting: A comparative analysis of multiple

machine learning algorithms. *Journal of Information Systems and Informatics, 5(2).* **[Crossref]**

Ferrer, A., Martínez, B., & Gómez, J. (2020). Crop yield estimation and interpretability with Gaussian processes. *Frontiers in Remote Sensing, 2*, 1010978.

Gao, Y., Wang, S., Guan, K., Wolanin, A., You, L., Ju, W., & Zhang, Y. (2020). The ability of sun-induced chlorophyll fluorescence from OCO-2 and MODIS-EVI to monitor spatial variations of soybean and maize yields in the Midwestern USA. *Remote Sensing, 12*(7), 1111. **[Crossref]**

Gopal, P. S. M., & Bhargavi, R. (2019). A novel approach for efficient crop yield prediction. *Computers and Electronics in Agriculture, 165*, 104968. **[Crossref]**

Jiya, U., Iliyasu, A., & Ebem, D. U. (2023). Agricultural research and food security under climate change: The place of machine learning models. *Journal of Advanced Mathematics and Computer Science, 11(1).* **[Crossref]**

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146–3154).

Khan, M., Khan, S., & Khan, M. Z. (2021). Optimizing soil fertility through machine learning: Enhancing agricultural productivity and sustainability. *Journal of Agricultural Informatics, 12*(3), 45–58.

Kheir, A. M. S., Negm, A., & El-Bastawesy, M. (2021). Remote sensing and GIS for estimating crop water consumption in dry environments: A case study of the Nile Delta region. *Remote Sensing Applications: Society and Environment, 22*, 100474. **[Crossref]**

Mamatha, & Kavitha, J. C. (2023). Machine learning-based crop growth management in greenhouse environment using hydroponics farming techniques. *Measurement: Sensors, 25*, 100665. **[Crossref]**

Martini, M., Offermann, F., Söder, M., Frühauf, C., & Finger, R. (2022). Machine learning can guide food security efforts when primary data are not available. *Nature Food, 3(9)*, 716–728. **[Crossref]**

Meng, Q., Hou, P., & Li, T. (2021). Integrating random forest and crop modeling improves the crop yield prediction of winter wheat and oilseed rape. *Frontiers in Remote Sensing, 2*, 1010978.

Muhammed I., Khan, S., & Khan, M. Z. (2021). Optimizing Soil Fertility through Machine Learning: Enhancing Agricultural Productivity and Sustainability. *Journal of Agricultural Informatics,* 12(3), 45-58.

Paudel, U., Adhikari, R., & Shrestha, S. (2021). A comparative analysis of machine learning models for rice yield prediction in Nepal. *Heliyon, 7*(3), e06404.

Pedamkar, P. (2020). Random forest in machine learning. *Analytics Vidhya.* Retrieved from https://www.analyticsvidhya.com

Prasad, S., Chawla, I., & Ghosh, S. (2021). Integrating satellite data and machine learning techniques for crop yield prediction: A case study of rice in India. *Remote Sensing Applications: Society and Environment, 22*, 100482.

Ramesh, A., Hebbar, V., Yadav, T., Gunta, A., & Balachandra, A. (2022). CYPUR-NN: Crop yield prediction using regression and neural networks. In *Emerging research in computing, information, communication and applications (ERCICA 2020).* Springer. **[Crossref]**

Seungtaek, & Tarar, S. (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. *Journal of Physics: Conference Series, 1714*(1), 012012. **[Crossref]**

Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S. V., & Laird, D. (2021). A comprehensive review of crop yield prediction using machine learning. *Frontiers in Plant Science, 12*, 616605. **[Crossref]**

Shuaibu, M. N., Muhammad, N., & Abu-Safyan, Y. (2021). Forecasting rice production in Jigawa State, Nigeria using fuzzy inference system. *Dutse Journal of Pure and Applied Sciences, 7(4)*, 203–213.

Singh, A., Kumar, P., & Kumar, A. (2022). Machine learning-based crop yield prediction: A survey. *Journal of King Saud University - Computer and Information Sciences, 34*(5), 1297–1313.

Sun, Y., Wang, S., & Tang, X. (2019). Outlier detection based on clustering by fast search and find of density peaks. *Information Sciences, 480*, 354–364.

United Nations. (2021). *The Sustainable Development Goals Report 2021*. United Nations.

Van Oort, B. G. H., Timmermans, F., Schils, R. L. M., & van Eekeren, N. (2023). Recent weather extremes and their impact on crop yields of the Netherlands. *European Journal of Agronomy, 142*, 126662. **[Crossref]**

Wickramasinghe, R., Weliwatta, P., Ekanayake, P., & Jayasinghe, J. (2021). Modeling the relationship between rice yield and climate variables using statistical and machine learning techniques. *Journal of Mathematics, 2021*, 6646126. **[Crossref]**

World Health Organization. (2021). *World health statistics 2021: Monitoring health for the SDGs, sustainable development goals*. World Health Organization.

Zhang, G. P. (2006). Avoiding pitfalls in neural network research. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, 37*(1), 3–16. **[Crossref]**

Zhang, Z., Wu, R. M. X., Yan, W., Fan, J., Gou, J., & Liu, B. (2022). A comparative analysis of the principal component analysis and entropy weight methods to establish the indexing measurement. *PLOS ONE, 17*(1), e0262261. **[Crossref]**

Zhi, X., Cao, Z., Zhang, T., Qin, L., Qi, L., Ge, A., Guo, X., Wang, C., Da, Y., Sun, W., & Liu, Y. (2022). Identifying the determinants of crop yields in China since 1952 and its policy implications. *Agricultural and Forest Meteorology, 327*, 109216. **[Crossref]**