ajfand Since 2001
AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

Afr. J. Food Agric. Nutr. Dev. 2024; 25(4): 26514-26533 https://doi.org/10.18697/ajfand.141.25330

Date	Submitted	Accepted	Published
	9th September 2024	5th April 2025	9th June 2025

MACHINE LEARNING TECHNIQUES FOR PREDICTION OF RICE YIELD IN RWANDA BASED ON WEATHER AND SOIL PARAMETERS

Mugemangango C[1]*, Nzabanita J[2], Muhoza DN[3] and ND Cahill[4]

Cyprien Mugemangango

*Corresponding author email: mucypro2@yahoo.fr

[1]African Center of Excellence in Data Science (ACE-DS), University of Rwanda

[2]College of Science and Technology, University of Rwanda

[3]College of Business and Economics, University of Rwanda

[4]College of Science, Rochester Institute of Technology, New York, USA

ABSTRACT

Rice is one of the six priority crops chosen by the government of Rwanda to increase agricultural productivity. This crop faces significant challenges from climate variability and soil conditions, affecting yields and food security. The past studies concentrated on building machine learning regression models for predicting the rice productivity with the help of weather parameters as predictors. Few researchers incorporated some soil nutrients to improve the precision of the prediction model. However, most of the farmers in Rwanda rely on agricultural input like inorganic fertilizer to boost the agriculture productivity. This means that the weather parameters are not sufficient to build a robust crop yield prediction model without incorporating the soil parameters. In this paper, rice yield status (favorable, unfavorable) prediction model was optimized using machine learning techniques, including logistic regression, random forest, gradient boosting, extreme gradient boosting, support vector machine, and artificial neural networks. The models were trained on historical weather parameters and soil nutrient data collected from the National Institute of Statistics of Rwanda, Rwanda Agriculture and Animal Resources Development Board and Rwanda Meteorology Agency. Findings indicated that the eXtreme Gradient Boosting (XGBOOST) model outperformed other models under consideration with 91%, 94.1%, and 82.7% of prediction accuracy, sensitivity, and precision, respectively. The above results demonstrate that the XGBOOST is the most robust model in predicting rice yield. The feature importance technique applied in this study also revealed that weather parameters (rainfall, and temperature), phosphorus, sulfur, manganese and carbon were crucial factors to prioritize for rice yield prediction compared to other model predictors under study. The results may provide valuable insights into whether the rice yield will be classified as low or high. This information is essential for helping farmers and decision-makers obtain actionable insights into expected crop yields, enabling them to make informed decisions on agricultural inputs resource allocation to maximize Rwanda's agricultural productivity.

Key words: artificial intelligence, machine learning, precision farming, Smart agriculture, yield prediction

INTRODUCTION

Rice comes from the seeds of the monocot plants *Oryza sativa* and *Oryza glaberrima*. It is a vital crop that plays a crucial role in ensuring food security and economic stability for many developing countries, including Rwanda. Rwanda's Crop Intensification program (CIP) is a land consolidation program aiming at improving agricultural productivity and food security, focusing on six priority crops including rice [1]. The increasing population and rising demand for rice underscore the need for enhanced productivity and sustainable agricultural practices.

Rwanda's unique agro-climatic conditions present opportunities and challenges for rice cultivation, making it an opportunity for further research. The previous researchers revealed that soil parameters including climate variability affect rice yield in Rwanda [2]. Improving rice production can contribute significantly to rural livelihoods, reduce dependency on imports, and support the country's vision for agricultural transformation.

Rwandan farmers face significant challenges in controlling rice yield production due to the heavy dependence on weather conditions, particularly rainfall and temperature. The annual precipitation in Rwanda ranges between 1,200 mm and 1,500 mm, and any fluctuations in these levels can have a profound impact on agricultural output [3]. Both the rainy and dry seasons pose unique challenges to crop production.

Mukamuhirwa *et al.* [4] conducted a study to evaluate the impact of simultaneous drought and temperature stress on seven rice cultivars from Rwanda. The experiment was carried out in climate chambers with two temperature regimes, 23/26°C (night/day) and 27/30°C (night/day), along with single or repeated drought treatments applied during various plant developmental stages. Their findings revealed that drought had a significant effect on plant development and yield, while genotype influenced the quality characteristics. Notably, the combination of high temperature and drought stress during the seedling and tillering stages resulted in the absence of panicle formation in all evaluated cultivars. This study highlights the critical interaction between temperature and drought stress on rice cultivation and the importance of considering both factors in breeding programs aimed at improving rice resilience to climate change. This study provides evidence that the heavy reliance on weather parameters like rainfall and temperature significantly impacts agricultural productivity in Rwanda.

Rwanda's diverse territory encompasses various soil types, each influencing crop production differently. The country is divided into distinct agro-ecological zones, each with unique soil characteristics. These differences, along with general challenges such

Since 2001 ajfand AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

as erosion, soil acidity, and nutrient deficiencies, significantly affect agricultural productivity [5]. In the northern and western provinces, the topography makes these areas particularly susceptible to landslides, which severely impact crop yields [6]. Additionally, the widespread use of fertilizers often fails to consider specific soil characteristics, leading to suboptimal results. For example, NPK 17 17 17 is commonly used across many regions and crops, regardless of the soil's unique needs [7]. To enhance crop yields, it is crucial to understand and manage soil characteristics. Monitoring soil parameters, alongside weather conditions, can provide valuable insights into how these factors influence agricultural productivity. By tailoring agricultural practices to the specific needs of the soil and climate, farmers can better predict and improve crop yields.

Various studies have aimed to develop crop yield prediction models using different technologies. Singh *et al.* [8] developed a machine learning-based approach to predict rice yield categories (bad, medium, high) using macro-nutrients (pH, OC, EC, N, P, K, S) and micro-nutrients (Zn, Fe, Mn, Cu) from data collected at Krishi Bhawan, India. They employed K-Nearest Neighbors (KNN), Naïve Bayes, and Decision Tree models, achieving accuracies of 96.43%, 97.80%, and 93.38%, respectively. The study concluded that Naïve Bayes and Decision Tree classifiers are effective for soil classification and yield prediction. However, the research lacks context on the practical application of these models and does not address the influence of external factors like weather, which, as Mukamuhirwa [4] highlighted, significantly impact rice yield. This suggests that relying solely on nutrient data may not optimize yield prediction in real-world scenarios.

Guo *et al*. [9] investigated the impact of climate change and phenological variables on rice yields. They used multiple linear regression (MLR) and three machine learning (ML) methods: backpropagation neural network (BP), support vector machine (SVM), and random forest (RF), to predict rice yields based on eleven combinations of phenology, climate, and geography data. The results showed that ML methods outperformed MLR, with SVM providing the highest prediction accuracy. Phenological variables significantly improved yield predictions, highlighting the need for their accurate representation in crop models to enhance rice yield predictions under climate change.

Chandraprabha and Dhanaraj [10] applied machine learning algorithms, including Naïve Bayes, Bayes Net, and IbK, to predict suitable crop varieties for soil based on total production and area sown on a district level. They used a soil-based dataset from 32 districts in Tamil Nadu, India, and compared the accuracy of these algorithms. The

Since 2001 ajfand AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

performance of their prediction models was evaluated using metrics such as true positive and false positive values, precision, recall, and F1-score. Their results revealed that Ibk and Random Forest were the most effective models for predicting soil-crop suitability, with high accuracy rates of 97.31% and 97.4%, respectively. In comparison, Bayes Net and Naïve Bayes performed less effectively, achieving significantly lower accuracies of 68.81% and 53.76%. However, the researcher does not provide more information whether the production will be good or bad.

Burdett and Wellen [11] investigated methods to quantify the interplay between soil and topographic characteristics to forecast crop yields using high-resolution data and advanced analytical techniques in South-western Ontario, Canada. Their study considered factors such as pH, soil organic matter (OM) content, cation exchange capacity (CEC), soil test phosphorus, zinc (Zn), potassium (K), elevation, and topographic wetness index. Among the evaluated techniques, random forests emerged as the most effective, achieving R^2 score values of 0.85 for corn and 0.94 for soybeans.

Kuradusenge *et al*. [12] developed a machine-learning model to predict maize and Irish potato yields in the Musanze district using rainfall and temperature as predictors. They found that the Random Forest algorithm was the most effective, with root mean square errors of 510.8 kg/ha for potatoes and 129.9 kg/ha for maize, and R^2 score values of 0.875 and 0.817, respectively. T. van Klompenburg *et al.* [13] reviewed 80 papers on crop yield prediction and noted significant inconsistency in the selection of parameters for inclusion in prediction models. These studies highlight that the accuracy of crop yield predictions depends significantly on the methods and techniques used.

Most previous studies have concentrated on predicting agricultural production using soil type, pH value, area of production, and weather parameters such as temperature, rainfall, precipitation, solar radiation, and relative humidity [13]. Few studies have considered soil nutrients such as nitrogen, potassium, zinc, magnesium, sulfur, boron, calcium, organic carbon, phosphorus, and manganese [8]. However, apart from uncontrolled weather parameters, soil nutrients are the principal components of the fertilizers used by most farmers worldwide to optimize agricultural productivity [14, 15]. Additionally, most literature has focused on developing regression models to predict the quantity of yield to be harvested, rather than building classification models. However, it might be difficult for some farmers to gain deep insights from the predicted values. This is because the predicted values do not indicate whether the production will be high or low, which is crucial for helping farmers make the right decisions about cultivating particular crops.

Since 2001 ajfand AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

The literature has shown that the soil and weather parameters have an undeniable contribution to crop yield. Therefore, there is a need to continue investigating the contribution of both weather and soil nutrient parameters for predicting agricultural production in Rwanda. There is also a need to develop classification prediction models that can inform the general public if agricultural production will be good in a particular season, helping farmers and stakeholders in making data-driven decisions about which crops to cultivate.

The current study aims to optimize the development of the rice yield prediction model by leveraging the power of machine learning techniques such as logistic regression, random forest, gradient boosting, extreme gradient boosting, support vector machine, and artificial neural network, using both historical weather parameters and soil nutrients recorded across different administrative sectors in Rwanda. These machine-learning algorithms were selected because of their popularity and performance in the reviewed literature [8, 16]. Findings from this research provide insight into whether rice yield will be high or low. This information is expected to help farmers and decision-makers to optimize agricultural resources allocation to maximize agricultural production based on facts.

MATERIALS AND METHODS

This section provides a comprehensive and detailed description of the various methodological approaches adopted in this study. The methods encompass data collection, data preprocessing, model implementation and model hyperparameters tuning.

Data Collection

The study utilized panel data gathered from various reliable government institutional databases. The dataset comprises observations from multiple regions (sectors) collected over two agricultural seasons (A and B) across a five-year period (2017 to 2021). The data were consistently gathered from the same regions during each agricultural season. As illustrated in Table 1, several variables were considered for predicting rice crop yield. These include the dependent variable, Rwanda sector-wise crop yield measured in kilograms per hectare sourced from the National Institute of Statistics of Rwanda (NISR), along with independent variables, which cover both quantitative factors such as weather parameters (precipitation, temperature) obtained from the Rwanda Meteorological Agency (Meteo Rwanda) and soil nutrients as well as categorical factors like agricultural season, soil texture, and soil depth sourced from the Rwanda Agriculture and Animal Resources Development Board (RAB). These

ajfand Since 2001 AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

comprehensive datasets ensure a robust foundation for the study, enabling accurate and reliable crop yield predictions. The following section briefly describes the motivation of each variable under consideration.

Crop yield is the output variable of interest, referring to the amount of rice produced per unit area (kg/ha), and is impacted by a combination of various factors such as climatic conditions, soil nutrients, and soil management factors. Precipitation is crucial for providing the necessary water for rice growth, with both the amount and distribution of rainfall impacting soil moisture, nutrient absorption, and overall crop development. While maximum and minimum temperatures are essential for crop growth, extreme temperature values can adversely affect crop development. Conversely, low temperatures may result in cold stress, particularly during the early stages of growth.

Soil nutrients are essential for crop productivity. Phosphorus is an important parameter for root development and energy transfer within the plant. Manganese supports photosynthesis and nitrogen assimilation, while carbon content in the soil, representing organic matter, improves soil fertility and water retention. The Effective Cation Exchange Capacity (ECEC) of the soil indicates its ability to retain and supply essential nutrients, crucial for sustained plant growth. Sodium level is also important to be controlled, excessive sodium can lead to salinity issues, negatively affecting water uptake and causing toxicity. Similarly, magnesium and calcium are essential for chlorophyll synthesis and cell wall stability, respectively, and their deficiencies can lead to poor growth.

Aluminum, commonly found in acidic soils, can impede root development and this indicates the importance of controlling its level to avoid toxicity. The Exchangeable Sodium Percentage (ESP) provides information about soil dispersion and its high concentration can potentially lead to reduced permeability and hindered root development. Iron is essential for crop production; its deficiency can negatively impact productivity since it is required for the synthesis of chlorophyll particularly in damp soil.

Moreover, soil micronutrients like copper and zinc are essential for enzyme activation and protein synthesis, playing essential roles in plant development. Soil depth influences the volume of soil accessible for root growth and nutrient uptake, with deeper soils generally promoting higher yields. Potassium is important for enhancing disease resistance and nutrient transport, both of which are crucial for grain filling. Boron is necessary for reproductive development, and deficiencies can lead to poor pollination. The soil's hydraulic properties, including its ability to retain and transmit water, are essential for ensuring adequate water availability, which is critical for rice growth.

ajfand Since 2001 AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

Electrical Conductivity (EC) measures soil salinity, with higher levels indicating potential salinity issues that can diminish yield. Finally, nitrogen is a key nutrient that supports vegetative growth and grain filling, with its availability having a direct impact on productivity. Soil texture, determined by the ratio of sand, silt, and clay, affects water retention and drainage, with an ideal balance promoting optimal conditions for rice cultivation.

Data Pre-processing

After data collection, the individual datasets were merged into a single comprehensive dataset. Following this, categorical features such as soil depth, soil texture, and agricultural season were transformed into numerical features using the label encoding function in Python. However, for categorical features with more than two values, binary dummy variables were created. The issue of multi-collinearity was addressed by eliminating one of the dummy variables generated, thereby optimizing the prediction process. The crop yield was calculated by dividing the original agricultural production by the area of the plantation. Furthermore, to prevent any bias during the model training process, all dataset features were standardized. This standardization ensured that all features had an equal opportunity to contribute to the model's training.

Categorizing the crop yield

In fulfilling the requirements of a classification machine learning problem, the target variable must consist of discrete classes or categories. To meet this requirement, the crop yield values were transformed into binary values: 1 indicates high yield production, and 0 represents low yield production. This transformation was achieved by computing the minimum, maximum, and mean values for each crop type considered in the study. Any crop yield value greater than or equal to the mean was categorized as high production, while any value less than the mean was categorized as low production.

Model selection and implementation

Prior to the implementation of the models, the predictors and response variables were identified from the original dataset. 75% of the original dataset was utilized as the training set. The remaining 25% of the dataset served as the testing set to validate the predictive models generated. The following section provides a detailed description of each model investigated and outlines the implementation process for each.

a) Random Forest: A random forest is one of the popular supervised machine learning algorithms and it can be used for both classification and regression problems. This algorithm has been selected because of its popularity and performance in different research studies [17]. To implement the random forest classifier, machine learning framework called scikit-learn available in python as a library was used. The following snapshot shows how the model was implemented with a python script.

```
from sklearn.ensemble import RandomForestClassifier # importing algorithm
model=RandomForestClassifier(max_depth=5,random_state=0) # build the model
model.fit(X_train,Y_train) # model training
```

b) **Gradient boosting**: Just like random forests, gradient boosting is a well-known supervised machine learning algorithm that is applicable to both classification and regression tasks. It is favored for its performance and widespread use in various research studies [18]. In order to create a gradient boosting classifier, the scikit-learn machine learning framework was used as demonstrated here.

```
from sklearn.ensemble import GradientBoostingClassifier # importing algorithm
model=GradientBoostingClassifier(max_depth=2,random_state=34) # build the model
model.fit(X_train,Y_train) # model training
```

c) **eXtreme Gradient Boosting:** XGBoost is a powerful machine learning algorithm known for its effectiveness in handling structured data. It belongs to the boosting ensemble methods, improving upon traditional gradient boosting with regularization and parallel processing capabilities. XGBoost iteratively corrects errors made by previous models to minimize a specified loss function, making it scalable and adept at handling large datasets. Its strengths in managing missing data, feature selection, and robust performance across various domains have made it widely used in both academic research and industrial applications for tasks like regression, classification, and ranking [18]. For implementing the model, the following python script was utilized.

```
import xgboost as xgb # importing algorithm
model = xgb.XGBClassifier(max_depth=2,learning_rate=0.1) # build the model
model.fit(X_train,Y_train) # model training
```

d) **Support Vector Machine:** Support Vector Machine (SVM) is a supervised learning model utilized for classification and regression purposes. It operates by identifying an optimal hyperplane in high-dimensional space to differentiate between different data classes [19]. The model was implemented using the python library as shown in the following script

```
from sklearn.svm import SVC # importing algorithm
model=SVC(kernel='rbf',max_iter=950) # build the model
model.fit(X_train,Y_train) # model training
```

e) Artificial Neural Network: Artificial Neural Networks (ANNs) are computational models inspired by biological neural networks. They consist of interconnected nodes organized into layers: input, hidden, and output. ANNs use activation functions to process input signals and adjust weights between neurons through algorithms like backpropagation during training. This process enables ANNs to learn complex patterns and relationships in data, making them versatile for tasks such as pattern recognition, classification, regression, and decision-making [20]. We used a three-layer feedforward neural network of the form:

$$\hat{Y} = \sigma(W_3.ReLU(W_2.ReLU(W_1.X + b_1) + b_2) + b_3,$$

where X is the input vector, W_i is the weight matrix of the i^{th} dense layer, b_i is the bias vector of the i^{th} dense layer, $ReLU(z) = \max(0, z)$ is the nonlinear activation function, and $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. The neural network was trained using dropout, with a fraction of the input units randomly set to zero at each update during training time.

The Tensorflow and Keras deep learning frameworks available in Python were used to implement this model, as demonstrated in the following script.

```
import tensorflow as tf
from tensorflow import keras
#Build model architecture of ANN
model = keras.models.Sequential([keras.layers.Dense(units=64,activation='relu',input_shape=[27]),
                                 keras.layers.Dropout(0.4),#0.2
                                 keras.layers.Dense(units=64,activation='relu'),
                                 keras.layers.Dropout(0.4),
                                keras.layers.Dense(units=1,activation='sigmoid')])
```

Code reproducibility

To enable other researchers to reproduce the results, the code used for data cleaning, preprocessing, transformation, and model development, along with all the datasets, have been uploaded to the GitHub platform. These resources are available at the following GitHub link:https://github.com/Mugemangango/Rice-Prediction-Models.

Model hyperparameters tuning

The model hyperparameters have been optimized using the Grid Search method, which is an exhaustive search method used to find the optimal hyperparameters for the

Since 2001 ajfand AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

machine learning model. The hyperparameters are the parameters that govern the training process and model architecture (max_depth, random state, learning rate, number of neurons in hidden layers of neural network). This was implemented by experimenting different range of values for each hyperparameter and retain the ones which offer the best model prediction accuracy. This method was adopted due to its simplicity in implementation and and guarantee for finding the best parameters combination within the search space [25].

RESULTS AND DISCUSSION

Throughout this study, six machine learning algorithms namely Random Forest, Gradient Boosting, Extreme Gradient Boosting, Logistic Regression, Support Vector Machine and Artificial Neural Network were trained and validated. The evaluation metrics selected include prediction accuracy, precision, and sensitivity, all of which can be computed from the confusion matrix.

Confusion matrix: In binary classification problems, the confusion matrix is a 2×2 matrix with entries describing the number of test samples that were predicted to be positive when they were actually positive (True Positives or TP), predicted to be positive when they are actually negative (False Positives or FP), predicted to be negative when they were actually negative (True Negatives or TN), and predicted to be negative when they were actually positive (False Negatives or FN) [21]. The figure 1 illustrates the confusion matrices of 6 prediction models under investigation in this paper.

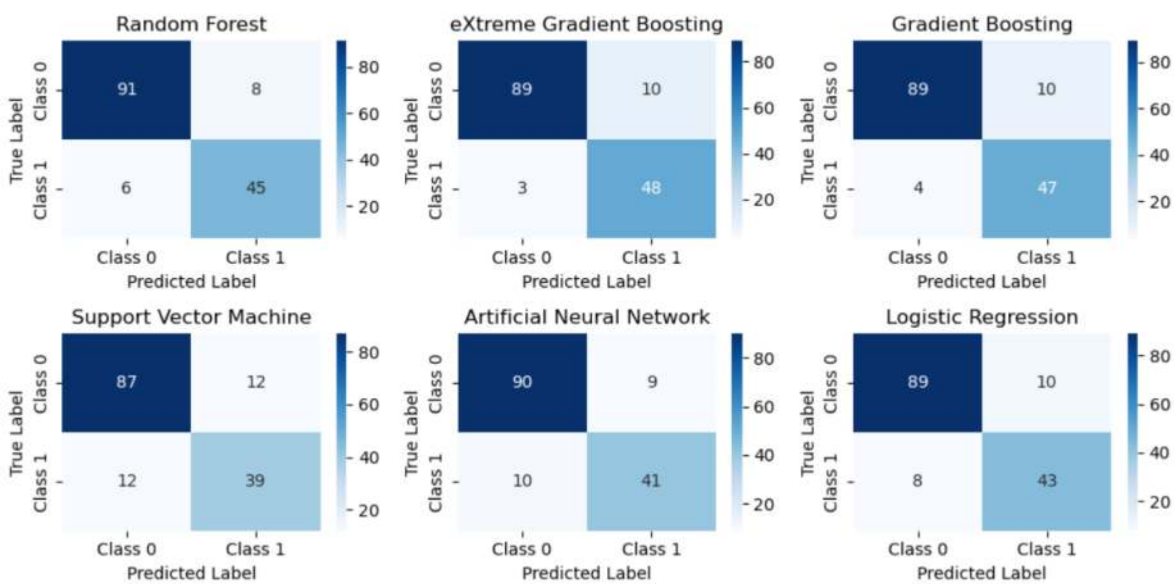

Figure 1: Model Confusion Matrices

Since 2001 ajfand AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

Accuracy: Prediction accuracy is the ratio of the number of correctly classified samples to the total number of the samples [22, 23]. It can be computed from the entries of the confusion matrix via:

Model accuracy = $\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$

Precision: This metric [23] indicates how many positive samples were correctly identified in the positive predictions made by the model. Mathematically, this metric has been computed by taking the ratio of the number of correctly identified positive samples and the total number of samples predicted as positive samples:

Model precision = $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$

Sensitivity: This metric [24] is used for evaluating the model ability to identify the positive samples. This was computed by taking the ratio of correctly predicted positive samples and the total number of positive samples in the validation subset.

Model sensitivity = $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$

As demonstrated in Table 2, which presents the rice yield prediction, the Random Forest, Gradient Boosting, and eXtreme Gradient Boosting models outperformed all other models examined in this study with 90%, 90.7%, and 91% of prediction accuracy, respectively. These results underscore the superior performance of these three machine learning models in predicting rice yields based on soil nutrients and weather conditions.

Overall, the eXtreme Gradient Boosting prediction model demonstrated the highest performance among all models evaluated in this study for rice yield prediction. These findings highlight the distinct advantages of different machine learning models depending on the specific crop being analyzed, with eXtreme Gradient Boosting excelling in rice yield prediction.

As illustrated in Figure 2, which depicts the precision and sensitivity of the rice yield prediction models, the eXtreme Gradient Boosting model achieved the highest precision score at 94.1%. This indicates that 94.1% of the positive predictions made by the eXtreme Gradient Boosting model were accurate, meaning that the model correctly identified 94.1% of the true positive instances of rice yield, while the remaining 5.9% were false positives. This high precision underscores the model's effectiveness in accurately predicting rice yields, reducing the likelihood of incorrect predictions.

Since 2001
ajfand
AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

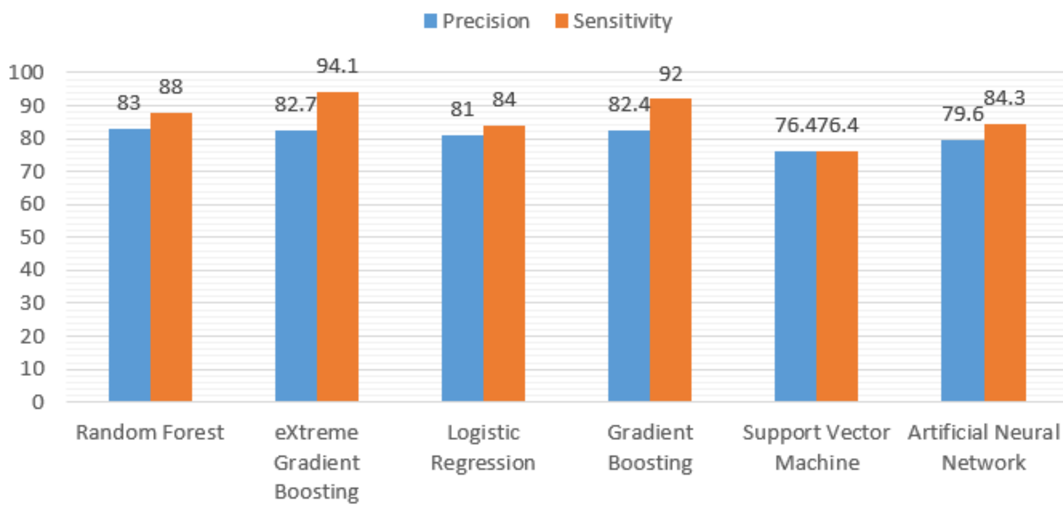

Figure 2: Model Precision and Sensitivity

Additionally, as depicted in Figure 2, the eXtreme Gradient Boosting (XGBoost) prediction model also achieved an impressive sensitivity score of 82.7%. This high sensitivity indicates that the model correctly identified 82.7% of the true positive samples, while only 17.3% were misclassified. These results underscore the potential of utilizing machine learning techniques, specifically XGBoost, in conjunction with weather parameters and soil nutrient data, to accurately predict crop yields, particularly for rice. This finding highlights the robustness and reliability of machine learning models in agricultural applications, paving the way for more precise and efficient crop management practices.

Feature importance

In developing the rice yield prediction model using the eXtreme Gradient Boosting (XGBoost) algorithm, understanding the importance of each feature was crucial to refining the model and enhancing its predictive accuracy. Feature importance provides insights into which variables most significantly influence the model's predictions, enabling better interpretation and more effective decision-making.

To assess feature importance, the XGBoost model was trained on a dataset comprising various weather parameters and soil nutrient levels. The model inherently calculates the importance of each feature by measuring how valuable each feature is in constructing the boosted decision trees. Specifically, XGBoost evaluates feature importance based on three main criteria.

As shown by Figure 3, the analysis revealed that certain features were more influential in predicting rice yield than others.

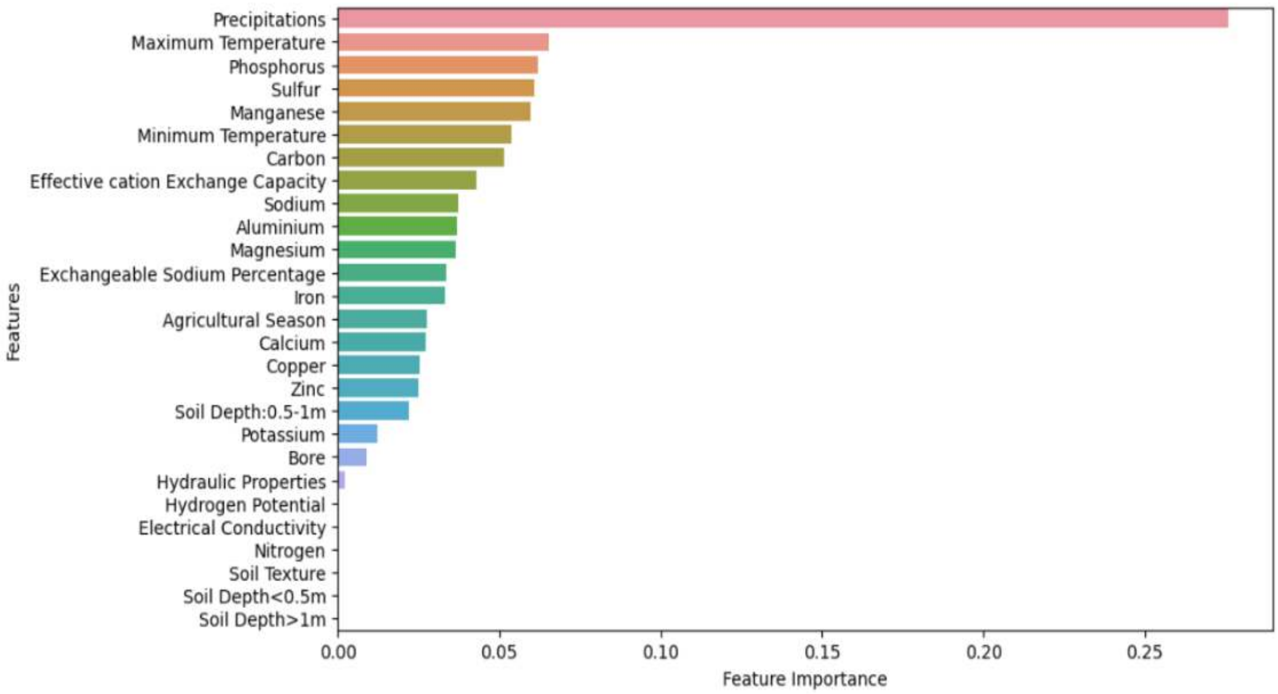

Figure 3: Model Feature Importance

Precipitations: Rainfall emerged as the most critical feature, contributing significantly to the model's predictive power. Adequate rainfall is essential for rice growth, and its variability had a substantial impact on yield predictions.

Temperature: Both minimum and maximum temperatures were important features. Optimal temperature ranges are crucial for the various growth stages of rice, and deviations from these ranges can significantly affect yield.

Soil Nutrients: As depicted in Figure 3, Phosphorus, Sulfur, Manganese and Carbon were identified among the soil nutrients as the critical features that influence the model predictive power.

Agriculture Season: This parameter has been selected also among the features that influence rice productivity.

Understanding the relative importance of these features helps in several ways:

a) **Model Optimization** by focusing on the most influential features, the model can be fine-tuned for more accurate predictions.
b) **Resource Allocation Optimization** where the farmers and agronomists can prioritize resources and management practices on the most impactful factors, such as soil nutrient management and weather parameters.
c) **Policy Formulation** by facilitating the policymakers to design targeted interventions aimed at enhancing key factors that drive yield to overall agricultural productivity.

https://doi.org/10.18697/ajfand.141.25330

The results emphasize the significance of incorporating weather conditions, soil micro-nutrients, and soil macro-nutrients for accurate rice yield prediction. This contrasts with other studies that focus exclusively on a subset of these features. For instance, Guo *et al.* [9], Mukamuhirwa *et al.* [4], and Kuradusenge *et al*. [13] concentrated solely on weather conditions, while Singh *et al.* [8] and Burdett and Wellen [11] primarily incorporated soil micro- and macro-nutrients. Such approaches may lead to sub-optimal performance in prediction models

CONCLUSION AND RECOMMENDATIONS FOR DEVELOPMENT

This study highlights the importance of accurate rice yield status prediction in Rwanda. A variety of machine learning techniques, such as logistic regression, random forest, gradient boosting, XGBOOST, support vector machine, and artificial neural networks have been considered for predicting if the future rice yield will be favorable or unfavorable based on weather conditions and soil parameters. The XGBOOST predictive model has been identified as the most effective tool for predicting rice yields, achieving high accuracy, sensitivity, and precision.

The implications of these findings are significant, as they contribute to enhancing food security and promoting economic stability in Rwanda. By optimizing agricultural production through data-driven strategies, stakeholders can better navigate the challenges posed by climate variability and soil conditions, ultimately leading to more sustainable and resilient agricultural practices. Future research could expand on this work by integrating additional variables, such as crop satellite images for monitoring crop growth, pest and disease prevalence, to further refine yield predictions and support comprehensive agricultural planning in Rwanda.

ACKNOWLEDGEMENTS

The authors extend their gratitude to Rwanda Meteorology Agency, Rwanda Agriculture and Animal Resources Development Board, and National Institute of Rwanda for supplying the data.

Since 2001 ajfand AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

Table 1: Variables under consideration

Variable name	Role	Type	Unit
Yield	Dependent	Continuous	kg/ha
Precipitation	Independent	Continuous	mm
Maximum Temperature	Independent	Continuous	℃
Minimum Temperature	Independent	Continuous	℃
Phosphorus	Independent	Continuous	ppm
Manganese	Independent	Continuous	mg/kg
Carbon	Independent	Continuous	%
Effective Cation Exchange Capacity	Independent	Continuous	$cmol_{(+)}/kg$
Sodium	Independent	Continuous	ppm
Magnesium	Independent	Continuous	$cmol_{(+)}/kg$
Aluminium	Independent	Continuous	ppm
Exchangeable Sodium Percentage	Independent	Continuous	%
Iron	Independent	Continuous	mg/kg
Agricultural Season	Independent	Categorical	
Calcium	Independent	Continuous	$cmol_{(+)}/kg$
Copper	Independent	Continuous	mg/kg
Zinc	Independent	Continuous	mg/kg
Soil depth	Independent	Categorical	mm
Potassium	Independent	Continuous	$cmol_{(+)}/kg$
Boron	Independent	Continuous	ppm
Hydraulic properties	Independent	Continuous	mm/m
Electrical conductivity	Independent	Continuous	mS/m
Nitrogen	Independent	Continuous	%
Soil Texture	Independent	Categorical	%

Since 2001
ajfand
AFRICAN JOURNAL OF FOOD, AGRICULTURE, NUTRITION AND DEVELOPMENT
SCHOLARLY, PEER REVIEWED
Volume 25 No. 4
May 2025
PUBLISHED BY AFRICAN SCHOLARLY SCIENCE COMMUNICATIONS TRUST
ISSN 1684 5374

Table 2: Rice Crop yield prediction accuracy

Model	Training accuracy	Testing accuracy
Random Forest	92%	90 %
eXtreme Gradient Boosting	91%	91%
Logistic Regression	87%	88%
Gradient Boosting	92.3%	90.7%
Support Vector Machine	88.1%	84%
Artificial Neural Network	88.9%	87.3%

REFERENCES

1. **Kathiresan A** Strategies for sustainable crop intensification in Rwanda. Shifting focus from producing enough to producing surplus, 2011: 1-59. http://197.243.22.137/gakenke/fileadmin/templates/DOCUMENT_Z_ABAKOZI/abakozi/CIP_Strategies_2011.pdf *Accessed September 2024.*

2. **Nzungize G, Hakizimana N, Muvunyi JD and L Nahayo** Implication of Climate Change on Rice Production in Nyanza District of Rwanda. *American Journal of Environment and Sustainable Development*, 2020; **5(2):** 24-29.

3. **IFRC.** Rwanda - Floods and Landslides. Disaster Response Emergency Fund (DREF) operation. https://reliefweb.int/report/rwanda/rwanda-floods-and-landslides-dref-operation-ndeg-mdrrw022 *Accessed March 2024*.

4. **Mukamuhirwa A, Persson Hovmalm H, Bolinsson H, Ortiz R, Nyamangyoku O and E Johansson** Concurrent drought and temperature stress in rice—a possible result of the predicted climate change: effects on yield attributes, eating characteristics, and health promoting compounds. *International Journal of Environmental Research and Public Health*, 2019; **16(6):** 1043.

5. **Uwiragiye Y, Ngaba MJ, Zhao M, Elrys AS, Heuvelink GB and J Zhou** Modelling and mapping soil nutrient depletion in humid highlands of East Africa using ensemble machine learning: A case study from Rwanda. *Catena,* 2022; **217**: 106499.

6. **Bizimana H and O Sönmez** Landslide occurrences in the hilly areas of Rwanda, their causes and protection measures. *Disaster Science and Engineering*, 2015; **1(1):** 1-7.

7. **Rwanda Environment Management Authority (REMA)**. Effectiveness and efficiency of fertilizer use in Rwanda, 2016. Available at: https://rema.gov.rw/fileadmin/templates/Documents/rema_doc/publications/Fertilizer-Report.pdf *Accessed July 2024.*

8. **Singh V, Sarwar A and V Sharma** Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach. *International Journal of Advanced Research in Computer Science*, 2017; **8(5).**

9. **Guo Y, Fu Y, Hao F, Zhang X, Wu W, Jin X, Bryant CR and J Senthilnath** Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecological Indicators*, 2021; **120**: 106935.

10. **Chandraprabha M and RK Dhanaraj** Soil based prediction for crop yield using predictive analytics. *In 2021 3rd international conference on advances in computing, communication control and networking (ICAC3N)*, 2021: 265-270. IEEE.

11. **Burdett H and C Wellen** Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. *Precision Agriculture*, 2022; **23(5)**: 1553-74.

12. **Kuradusenge M, Hitimana E, Hanyurwimfura D, Rukundo P, Mtonga K, Mukasine A, Uwitonze C, Ngabonziza J and A Uwamahoro** Crop yield prediction using machine learning models: Case of Irish potato and maize. *Agriculture*, 2023; **13(1):** 225.

13. **Van Klompenburg T, Kassahun A and C Catal** Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 2020; **177**:105709.

14. **Gomes L, Simões SJ, Dalla Nora EL, de Sousa-Neto ER, Forti MC and JP Ometto** Agricultural expansion in the Brazilian Cerrado: Increased soil and nutrient losses and decreased agricultural productivity. *Land*, 2019; **8(1):** 12.

15. **Syers JK** Managing soils for long-term productivity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 1997; **352(1356):** 1011-21.

16. **Halder M, Datta A, Siam MK, Mahmud S, Sarkar MS and MM Rana** A Systematic Review on Crop Yield Prediction Using Machine Learning. *In The International Conference on Intelligent Systems & Networks*, 2023: 658-667. Springer, Singapore.

17. **Parmar A, Katariya R and V Patel** A review on random forest: An ensemble classifier. *In International conference on intelligent data communication technologies and internet of things (ICICI)* 2018, 2019: 758-763. Springer International Publishing.

18. **Bentéjac C, Csörgő A and G Martínez-Muñoz** A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 2021; **54**: 1937-67.

19. **Suthaharan S and S Suthaharan** Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 2016: 207-35.

20. **Zhang Z** A gentle introduction to artificial neural networks. *Annals of translational medicine*, 2016; **4(19).**

21. **Marom ND, Rokach L and A Shmilovici** Using the confusion matrix for improving ensemble classifiers. *In 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, 2010: 000555-000559. IEEE.

22. **Huang J and CX Ling** Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 2005; **17(3):** 299-310.

23. **Rainio O, Teuho J and R Klén** Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 2024; **14(1):** 6086.

24. **Naidu G, Zuva T and EM Sibanda** A review of evaluation metrics in machine learning algorithms. *In Computer Science On-line Conference*, 2023: 15-25. Cham: Springer International Publishing.

25. **Belete DM and MD Huchaiah** Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, 2022; **44(9):** 875-86.