Journal of Theoretical and Applied Information Technology
15th April 2024. Vol.102. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REVOLUTIONIZING RICE YIELD PREDICTION: A DATA-DRIVEN APPROACH IN MAURITANIA

CHEIKH ABDELKADER AHMED TELMOUD[1] , MOHAMEDOU CHEIKH TOURAD[2]
[1,2]Scientific Computing, Computer Science and Data Science Research Unit (CSIDS), University of Nouakchott, Nouakchott, Mauritania
E-mail : [1]cheikhahmedtelmoud@gmail.com, [2]cheikhtouradmohamedou@gmail.com

ABSTRACT

In agricultural development, accurate forecasting of crop yields is crucial to ensure food security and resource allocation. This study aims to demonstrate the power of data by optimizing machine learning models to predict increased rice yields. We use a holistic approach that combines advanced machine learning techniques, robust data prioritization, and feature engineering to extract meaningful insights from climate and agricultural data using models such as random forest regression (RFR) and gradient boosting regression (GBR). achieved remarkable accuracy in predicting grain yield. Through extensive testing and analysis, we show that our model is superior to traditional methods such as K-Nearest Neighbor (KNN), including Long-Term Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) We our findings highlight the potential of optimized machine learning models to modify rice yield forecasts Farmers and policy makers are empowered to make informed decisions with valuable insights. By harnessing the power of data, we are paving the way for sustainable agricultural practices and making important contributions to global initiatives aimed at achieving food security, especially in Africa.

Keywords: *Agriculture, IoT, KNN, LSTM, RFR, GRU*

1. INTRODUCTION

In a world seeking sustainable food products, the optimization of machine literacy models presents an unknown occasion for revolutionizing agrarian practices. With a specific focus on rice yield vaticination, this composition delves into slice-edge advancements in machine literacy and explores the eventuality of employing data-driven perceptivity to optimize rice civilization and enhance crop productivity. Join us on a witching trip as we unravel the secrets behind directly prognosticating rice yields, paving the way for a more sustainable and food-secure future.

Machine literacy algorithms, particularly those based on deep literacy and neural networks, have emerged as important tools for assaying complex datasets and rooting precious perceptivity.

These algorithms retain the capability to reuse large volumes of data, including rainfall patterns, soil characteristics, agronomic practices, and literal yield records, to discern retired patterns and connections. By using this wealth of information, machine literacy models can provide an accurate prognostication of rice yields, enabling growers to make informed opinions regarding crop operation and resource allocation.

Among colorful machine learning algorithms, Random Forest Regression has gained considerable attention in the field of rice yield vaticination.

Random Forest Regression is an ensemble literacy system that combines multiple decision trees to obtain robust prognostications. It can handle both numerical and categorical variables, prisoner nonlinear connections, and relationships between multiple features. Random Forest Regression has demonstrated efficacy in handling miscellaneous datasets, reducing overfitting, and providing interpretable results [1].

Experimenters have successfully applied Random Forest Regression to rice yield vaticination, achieving notable delicacy and prophetic power.

Studies have integrated different datasets, including rainfall data, soil parcels, operational practices, and literal yield records, into Random Forest Regression models. The algorithm has shown its capability to capture the complex relationships between these variables and accurately prognosticate rice yields [2]. Moreover, its ability to

handle missing data and outliers adds to its appeal for practical operation in the agrarian sphere [3].

Although Random Forest Regression has shown pledge, experimenters continue to explore ways to enhance its performance and address specific challenges in rice yield vaticination. Similar to point engineering, hyperparameter optimization and model interpretation contribute to perfecting the overall prophetic capabilities of Random Forest Regression models. Sweats are also being made to integrate other data sources, such as remote seeing and IoT bias, to further enhance the delicacy and punctuality of prognostication [4].

Despite the progress made with Random Forest Regression and other machine learning methods, challenges remain in the field of rice yield vaticination. These include addressing issues of data quality, icing model generalizability across different regions and growing conditions, and incorporating socio-profitable factors into prophetic models. Ongoing exploration aims to upgrade and knitter Random Forest Regression models to specific geographical areas, cropping systems, and husbandry practices to maximize their effectiveness [5].

In the following sections, we clarify the data accession and medication processes, explore point engineering methods acclimatized to Random Forest Regression, band model evaluation, and hyperparameter optimization styles specific to this algorithm, and punctuate the benefits and unborn directions of optimizing Random Forest Regression models for enhanced rice yield vaticination. Using the power of data-driven perceptivity and the capabilities of Random Forest Regression, we can unleash the full eventuality of machine learning to revise rice civilization and contribute to a more sustainable and food-secure future.

2. RELATED WORK

Precision husbandry, an operational approach that utilizes machine literacy and deep literacy, has been employed to enhance crop quality and volume [6]. Agrarian coffers can be managed more effectively by employing machine and deep-literacy models.

In this section, we present related work and prominent models that bandied in the literature for estimating rice quality and yield, along with the data preprocessing procedures and corresponding issues.

In a previous study [7], the Naive Bayes algorithm was employed to classify rice-splint conditions. The proposed methodology encompasses five crucial methods: splint image accession, image preprocessing, green-pixel masking, birth of RGB pixel proportions, complaint discovery, and complaint brackets. The experimenters collected a dataset of 60 images from both field and online sources. To train the classifier, 75 datasets were used, and the remaining were allocated for evaluation purposes. In their study, the experimenters employed the chance of color in the affected area to describe and classify pixels into three types of rice conditions: Brown Spot (BS), Bacterial Leaf Blight (BLB), and Rice Blast (RB). The system achieved a delicacy of 89 for the RB and 90 for BS and BLB. However, bracket effectiveness varied based on factors such as lighting conditions, discrepancy, and image capturing angles.

Another study [8] proposed a prototype system for detecting and classifying rice conditions, including BS, BLB, and Leaf Smut (LS). They employed an SVM model for brackets using images of infected rice shops. The system involves back-ground junking, noise elimination, image resizing, segmentation, and the identification of complaint areas using the K-means clustering system. To separate each complaint, 88 features were used in three orders (color, texture, and shape).

The SVM model achieves a training delicacy of 93.33 (94 BLB, 92 BS, and 94 LS) and a test delicacy of 73.33, 100 BLB, 80 BS, 40 LS). Another study [9] employed an SVM model combined with a Convolutional Neural Network (CNN) for point birth to classify three rice conditions. They collected 619 splint complaint images from a rice field and achieved a bracket delicacy of 91.37 using an 80-20 training- test partition.

In a different approach, [10] employed a K-Nearest Neighbors (KNN) model for the bracket of rice splint conditions. The input dataset comprised 115 rice splint im-ages from four conditions collected at an agrarian exploration station.

The KNN model achieves an overall delicacy of 87.02. Similarly, [11] conducted a relative study of colorful classifiers, including Decision Trees, Logistic Retrogression, KNN, and Naive Bayes, for the discovery and bracket of three rice splint conditions. They used a dataset of 480 splint images under conditions from the UCI Machine Learning Repository.

These studies show different approaches to exercising machine literacy algorithms for the

discovery and bracket of rice splint conditions, pressing the eventuality of these models in perfect husbandry and complaint operations. Pollutants were employed to minimize the processing time and excerpt applicable objects from the images.

The data were divided into 90 for training and 10 for testing, exercising a10-fold cross validation approach. The algorithms were evaluated by calculating the pointers similar to the True Positive Rate (TPR), False Positive Rate (FPR), Precision (P), Recall (R), F-measure, and Area under the ROC curve (AUC). The Decision Tree algorithm outperformed other algorithms, achieving a delicacy of 94.9 for training and 97.9 for testing.

In a related study [12], Random Forest (RF) was compared to Multiple Linear Regression (MLR), Back Propagation neural network (BP), and Support Vector Machine (SVM) for rice yield vaccination using climate, phenological, and geographical data. The dataset comprised 75 agrometeorological stations in southern China, which were divided into training (70) and testing (30) sets. The performance of machine literacy styles sur-passed MLR, with SVM and RF outperforming the BP. The study linked strong negative correlations between phenological and geographical variables and rice yields and positive correlations between climatic variables and rice yields. Hence, these variables were considered to be significant predictors of yield vaticination. The combination of integrated phenological, climatic, and geographical data yielded stylish results. The evaluation measured R-squared (R^2) and Page 1 of 2 Root Mean Square Error (RMSE), with SVM achieving the loftiest delicacy in yield prognostications.

In a study on guard-1A Interferometric Wide-swath (IW) mode data, Random Forest (RF) was compared to a grade Boosting Decision Tree (GBDT), SVM, and KNN for estimating the dry biomass of rice [13]. The dataset covered an area of approximately 715 km2 in Tong Xiang County, China. The results showed that the VV (perpendicularly transmitted and perpendicular public) polarization was more accurate than VH (perpendicular transmitted and vertical public) polarization for biomass estimation using the employed algorithms. In addition, VHVV (a combination of VH and VV) produces more accurate estimates than when using the two polarizations alone. The combination of VHVV with K-NN (R^2 = 0.73 and RMSE = 462.4 g/m^2), and VHVV with K-NN (R^2 = 0.70, RMSE = 484.1 g/ m^2) throughout the growing season.

In another study [14], five styles, including the area system, panicle particularity phenotyping tool P-TRAP, Faster R-CNN, Cascade R-CNN, and Single Shot Multi-Box Sensor SSD, were compared to estimate spikelet figures grounded in RGB images. Deep literacy styles, specifically Faster R-CNN and Cascade R-CNN, displayed superior vaticination delicacy compared to the other styles. The protruded R-CNN system achieved an R2 of 0.99, and a Mean Absolute Error (MAE) of lower than 1.42, whereas the Faster R- CNN system achieved an R2 of 0.99, and an MAE of lower than 1.68. In terms of the time consumption, the Faster R-CNN outperformed the protruding R-CNN. When classifying indica and japonica rice Using X-ray images of spikelet's, Resnet- 50 achieved the loftiest delicacy of 0.96, while SSD was the least accurate.

In a final study [16], KNN, LSTM, and GRU yielded significant results with few textual data, with a Mean Absolute Error (MAE) of 0.070 and an MSE of 0.014 with KNN (K=3), 0.061 and MSE of 0.010 with LSTM, and 0.081 for MEA and 0.015 for MSE with GRU.

In summary, Decision Tree, Resnet, Faster R-CNN, and Cascade R-CNN are the most effective models for image-based bracket problems. Neural networks, like ResNet, are particularly suitable for structured data with large volumes, whereas decision trees are recommended for small datasets with categorical data.

Finally, [11], [15], [9], [7], and [10] aimed to identify and categorize foliage ailments in rice. They used logistic regression, Naive Bayes Classifier, KNN, and Decision Tree with an accuracy of 97.9% [11], SVM with an accuracy of 73.33% [15], and 91.37% [9], Naive Bayes with an accuracy of 90% [7], and KNN with an accuracy of 87,02% [10].

The articles [12], [13], and [16] have as a goal to forecast the productivity of rice, they have used RF with R2=0.31, MLR, BP, and SVM with R2=0.33 [12], RF R2=0.73 RMSE = 462,4 g/m2, GBDT, SVM, and KNN with R2=0.72 and RMSE= 362.4 g/m2 by [13], KNN(K=3) with MEA=0.070 and MSE=0.014, LSTM with MEA=0.061 and MSE=0.010, GRU with MEA=0.081, and MSE=0.015 by [16].

The study [14] categorized various varieties of rice and used Faster R-CNN, Cas-cade R-CNN, SSD, SVM, and Resnet-50 with an accuracy of 96%.

3. PROPOSED APPROACH

In this study, we introduce a novel random forest regression model to predict rice yield by combining climate data and farmland as inputs. The performance of our model is compared with established models, including K-Nearest Neighbor (KNN), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM).

3.1 Study Position and Dataset

The exploration conducted for this research centers on rice fields situated in Ros-so, Mauritania, a region of immense agricultural significance in Africa. Rosso, a megacity located in the southern region of Mauritania, lies adjacent to Senegal. The geographical coordinates of the study area are 16.536087025686108 latitude and -15.849748298631956 longitude. Mauritania is located in the northwest of the African mainland and is characterized by vast desert areas, a generally hot and arid climate, and brief rainy seasons that coincide with periods of precipitation [16].

A critical water source for agriculture in Mauritania is the Senegal River, which also serves as the primary drinking water source for over 4.65 million individuals (according to 2020 data from the Ministry of Housing and Urban Development of Mauritania).

Rice cultivation in the Rosso area follows two distinct seasons: the rainy season (downtime season) and dry-hot season. Each period lasted no longer than four months, and the cultivated area varied between the two seasons.

The primary dataset used in this study was sourced from the monitoring station of the Office of National Meteorology in Mauritania [16]. This governmental institution is responsible for capturing and analyzing climatic conditions across the country. The dataset comprises diurnal records encompassing six key variables: date, humidity, temperature, precipitation, wind speed, rice yield, and cultivated areas.

In total, the dataset comprised a substantial collection of 5,113 records spanning 2000 to 2013 (see Figure 1 and 2 for visual representations).

Each record in the dataset provides valuable insights into daily weather conditions, including humidity, temperature, precipitation, and wind speed. Additionally, it includes essential data points, such as rice yield and cultivated area, which serve as vital indicators of crop productivity and agricultural performance.

The richness and extent of this dataset provide a robust foundation for the optimization of our machine learning models for enhanced rice yield prediction. By harnessing the power of these valuable data, which encapsulates the intricate interplay between climate variables and rice cultivation, we unlock the potential to revolutionize rice cultivation practices in Mauritania and contribute to regional food security efforts. The detailed and comprehensive nature of the dataset allowed us to gain a profound understanding of the factors influencing rice yield and empowered us to make informed decisions regarding sustainable agricultural practices.

As we delve into the deployment of our optimized Random Forest Regression model, the depth of insights offered by this dataset becomes increasingly vital in shaping accurate predictions and enabling data-driven solutions for agriculture in Mauritania. With the synergy of advanced machine learning techniques and this rich dataset, we embark on a transformative journey towards a more productive and sustainable future for rice cultivation in this region.

3.2 Research Process

Our research process comprises six stages for prophetic model assessment:

1. **Data Accession**: Gathering necessary data for analysis.
2. **Data Cleaning and Preparation**: Removing inconsistencies and formatting data.
3. **Correlation Analysis**: Exploring connections between variables.
4. **Trend Analysis**: Using cumulative corruption to identify underlying trends.
5. **Model Operation**: Implementing the prophetic model on the dataset.
6. **Evaluation Criteria Application**: Using specific criteria to assess model performance and sensitivity.

3.3 Comprehensive Analysis

Originally, we used colorful statistical measures, including minimum, outside, mean, and friction, for each variable. We conducted an analysis to assess the correlation between these variables and identify direct implicit associations. In other words, we determined whether these variables were accompanied by harmonious changes. As depicted in Fig.6, our findings indicated a positive correlation between yield and moisture (0.16), temperature (0.18), rush (0.094),

and cultivated area (0.74). Again, correlation between yield and wind speed was nearly neutral.

These results indicated a relationship between yield and several climatic factors, particularly moisture and temperature. The weak positive correlation between yield and rush can be attributed to the current use of irrigation systems by growers, which rely less on rainwater [16].

The structure of the training dataset contains the following variables: humidity, which varies between 6% and 99% with an average of 63.71%; temperature, which varies between 20.5°C and 49°C with an average of 37.11°C; precipitation, which varies between 0 mm and 104 mm with an average of 0.81 mm; wind speed, which varies between 0 and 31 m/s with an average of 4.12 m/s; rice yield, which varies between 0 tons and 6 tons with an average of 2.96 tons; and cultivated areas, which vary between 0 hectares and 24807 hectares with an average of 7463.40 hectares.

3.4 Analysis of additive decomposition

In traditional time-series analysis, a time series **Xt** can be broken down into three distinct components: trend, seasonality, and noise. Mathematically, this equation can be rewritten as **Xt = Tt + St + εt**. The trend component Tt signifies the presence of a persistent upward or downward pattern throughout the time-series.

The seasonality component **St** captures any recurring pattern that is cyclically manifested within the time-series. By contrast, the residual component **εt** represents a portion of the data that remains unexplained by the decomposition process.

Upon examining the trend of the yield, we observed a recurring pattern that spanned four months and was repeated twice a year. This is a natural occurrence since there are two annual cultivation periods. Notably, peak yield coincided with higher temperatures, humidity levels, and cultivated surface areas. These observations align with the identified correlations and further support the relationship between these factors.

3.5 Evaluation Metrics

Sections In our analysis, we employed colorful climatic factors, including temperature, moisture, downfall, and wind speed, along with the cultivated area, to predict the rice yield in the fields of Mauritania. To assess the delicacy of our prognostication, we divided our data into 80 for training purposes and 20 for testing. The following evaluation criteria were used:

- **The Forecast Error (FE)** was calculated as the difference between the factual and read values.
- **Mean Forecasting error (MFE)** obtained by casting cast crimes and dividing data length.
- **The Mean Absolute Error (MAE)** determined by casting the absolute values of the cast crimes and dividing by the data length.
- **The Mean Square Error (MSE)** was calculated by casting squared cast crimes and dividing them by the data length.
- **Root Mean Square Error (RMSE)** deduced from the square root of MSE.

4. RESULTS AND DISCUSSIONS

We used the same dataset as that in [16]. We employed different models, Random Forest Regression, and Gradient Boosting Regression, to obtain better results. Subsequently, we compare our findings with those presented in the composition (Table 1).

Table 1: Results of yield prediction by KNN, LSTM, GRU, RFR, and GBR models.

Used Models	Evaluation Metrics			
	MFE	MAE	MSE	RMSE
KNN (K=3)	0.031	0.070	0.014	0.121
LSTM	0.008	0.061	0.010	0.100
GRU	0.028	0.081	0.015	0.126
RFR	0.0003	0.0008	0.0004	0.020
GBR	-0.003	0.054	0.006	0.081

To discuss the results and demonstrate the effectiveness of the chosen Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) models, we analyzed the evaluation metrics presented in Table 1.

First, we compared the performances of the models based on the Mean Forecasting Error (MFE), Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

- ✓ The RFR model achieved the lowest MFE value of 0.0003, indicating that its predictions were closest to the actual yield values. This result suggests that the RFR effectively captured the underlying patterns and trends in the data. GBR also

demonstrated promising performance with a negative MFE of -0.003, indicating a slight underestimation of the yield.

- ✓ MAE: Once again, the RFR demonstrated its superiority by obtaining the lowest MAE of 0.0008. This indicates that, on average, the predictions of the RFR model deviated minimally from the actual yield values. The GBR also performed well, with an MAE of 0.054, indicating relatively accurate predictions.
- ✓ MSE: RFR outperformed the other models and achieved the lowest MSE of 0.0004. This implies that the RFR model exhibits the smallest squared difference between its predictions and actual yield values. The GBR also demonstrated a low MSE of 0.006, indicating its ability to capture the underlying patterns effectively.
- ✓ RMSE: When considering the RMSE, the RFR displayed the smallest value of 0.020, demonstrating its superior predictive accuracy. GBR also showed a reasonably low RMSE of 0.081.

Based on these evaluation metrics, it is evident that both the Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) models perform well in predicting yields. However, the RFR consistently achieved slightly better results across all metrics, indicating its higher accuracy and better fit to the dataset. The difference in performance between RFR and GBR might be attributed to the distinctive ensemble techniques employed in each model.

Overall, the results strongly support the efficacy of our choice to utilize Random Forest Regression (RFR) for yield prediction, highlighting its superior performance compared with other models, including K-Nearest Neighbour (KNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU).

5. FINE-TUNING AND HYPERPARAMETER OPTIMIZATION

To maximize the potential of our Random Forest Regression (RFR) model for rice yield prediction, we embark on a thrilling journey of fine-tuning and hyperparameter optimization. This pivotal phase enabled us to extract the best performance from our model, thereby unlocking unprecedented levels of accuracy and insight.

To achieve this, we adopted a multifaceted approach that leverages a diverse set of optimization techniques, each with its unique advantages in navigating the vast hyperparametric search space.

Optuna, a state-of-the-art hyperparameter optimization framework [17], guides our RFR model towards the optimal configuration. This intelligent tool efficiently explores the hyperparameter space, uncovers hidden insights within the data, and fine-tunes the model to achieve peak performance. Seamlessly integrated with our model training pipeline, Optuna revolutionizes the optimization process, enabling rapid sampling and evaluation of diverse hyperparameter combinations.

In parallel, we enlist the advantages of Grid Search and Random Search, and venerable optimization methods that tirelessly scour the hyperparameter space [18]. With their versatility and systematic exploration, these techniques meticulously assess an array of configurations, eliminate suboptimal choices, and identify the best-suited parameters for our RFR model.

As we navigate the landscape of optimization, we encounter the brilliance of Bayesian Optimization, a method that introduces probabilistic reasoning to our journey [19].

By constructing a surrogate model of the objective function, Bayesian optimization intelligently infers the most promising hyperparameters to explore, guiding our model towards even higher accuracy and predictive power.

With the confluence of these fine-tuning and hyperparameter optimization techniques, we observed a remarkable transformation in the predictive capabilities of our RFR model. The synergy of modern optimization methods with the knowledge gained from our data-driven insights reveals a new era of rice yield prediction.

As we progress towards unparalleled accuracy and informed decision-making, we embrace the latest advancements in optimization techniques, shaping the future of sustainable and data-driven agriculture [20]. Empowered by the harmony between machine learning and optimization, we journey towards a more resilient and food-secure future.

At the vanguard of agricultural innovation, the optimized RFR model ignites a path of progress, inviting us to explore a more productive and sustainable world of rice cultivation. On this exciting frontier, we forge ahead with

determination, fueling the spirit of optimization, innovation, and cultivation to cultivate a brighter future for agriculture.

Table 2: Optimization Results

Optimization Methods	Evaluation Metrics For RFR		
	MAE	MSE	RMSE
Bayesian Search	0.0008699	0.0003588	0.0189432
Random Search	0.0009593	0.0004103	0.0202577
Grid Search	0.0008347	0.0004110	0.0202754
Hyperopt	0.0007478	0.0002326	0.0152520
Optuna	0.0001675	$1.4363516e^{-05}$	0.0037899

6. CONCLUSION

In conclusion, this study successfully unleashed the power of the data by optimizing machine-learning models for enhanced rice yield prediction. By leveraging advanced techniques, such as Random Forest Regression (RFR) and Gradient Boosting Regression (GBR), we have achieved remarkable accuracy in forecasting rice yields. Through rigorous data preprocessing, feature engineering, and hyperparameter optimization using Optuna, we have improved the performance of our models and sur-passed traditional methods such as K-Nearest Neighbor (KNN) and recurrent neural net-works such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU).

The results obtained from our optimized models, as presented in Table 2, demonstrate their effectiveness in capturing the underlying patterns and trends in the data. In particular, it exhibited superior performance with the lowest Mean Forecasting Error (MFE), Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) values. These metrics indicate the accuracy and precision of our model for predicting the rice yield.

Comparative bibliographic contributions:

Compared to existing research (a summary of related work in Section 2), our work makes a unique contribution to the field. Many studies focus on specific parameters, such as disease classification or biomass estimates. In contrast, our study optimizes machine learning models to predict overall rice yield, showing high accuracy and robustness across different parameters.

Moreover, the combination of advanced methods and careful optimization process of hyperparameter tuning with Optuna made our study stand out. We not only provide accurate forecasts but also contribute to a greater understanding of the potential of machine learning in agricultural practices.

Beyond Rice Yield Prediction:

The implications of our research go beyond just talking about crops. By leveraging data and machine learning, we help advance sustainable agricultural practices. Our customized models provide farmers and policymakers with actionable insights, enabling the allocation of effective inputs and strategic crop management. This strengthens food security and is in line with global initiatives to meet the challenges posed by a changing climate.

Future Directions:

Looking ahead, our research paves the way for further developments in machine learning, data acquisition, and model optimization. Emerging technologies such as the integration of Internet of Things (IoT) devices and remote sensing, as well as ongoing research into translational AI, promise to improve the accuracy and interpretation of predictions.

Additionally, continued research on explainable AI and model interpretability will foster trust and acceptance of machine learning models among stakeholders, and there is potential for the implementation of an intelligent agricultural micro-system predicated on data-informed decision-making within the context of a big data environment [21][22][23].

In summary, our analysis in the broad landscape of the literature demonstrates the transformative potential of machine learning in agriculture, especially in cereal yield prediction. Through careful adaptation and a comprehensive approach, we help create a more sustainable and food-safe future. Continued research and collaboration will further unlock the potential of machine learning, driving innovation and progress in global agricultural practices.

REFERENCES:

[1] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, https://doi.org/10.1023/a:1010933404324.

[2] T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” Computers and Electronics in Agriculture, vol.

177, p. 105709, Oct. 2020, https://doi.org/10.1016/j.compag.2020.105709.

[3] A. P. Marques Ramos et al., "A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices," Computers and Electronics in Agriculture, vol. 178, p. 105791, Nov. 2020, https://doi.org/10.1016/j.compag.2020.105791.

[4] D. M. G. dela Torre, J. Gao, and C. Macinnis-Ng, "Remote sensing-based estimation of rice yields using various models: A critical review," Geo-spatial Information Science, pp. 1–24, Sep. 2021, https://doi.org/10.1080/10095020.2021.1936656.

[5] Q. Zhang et al., "Maize yield prediction using federated random forest," Computers and Electronics in Agriculture, vol. 210, pp. 107930–107930, Jul. 2023, https://doi.org/10.1016/j.compag.2023.107930.

[6] I. Cisternas, I. Velásquez, A. Caro, and A. Rodríguez, "Systematic literature review of implementations of precision agriculture," Computers and Electronics in Agriculture, vol. 176, p. 105626, Sep. 2020, https://doi.org/10.1016/j.compag.2020.105626.

[7] T. Islam, M. Sah, S. Baral, and R. Roy Choudhury, "A Faster Technique on Rice Disease Detection using Image Processing of Affected Area in Agro-Field," IEEE Xplore, Apr. 01, 2018. https://ieeexplore.ieee.org/abstract/document/8473322.

[8] Harshadkumar B. Prajapati, Jitesh P. Shah, and Vipul K. Dabhi, "Detection and classification of rice plant diseases," Intelligent Decision Technologies, pp. 1–18, Nov. 2018, https://doi.org/10.3233/idt-180338.

[9] V. K. Shrivastava and M. K. Pradhan, "Rice plant disease classification using color features: a machine learning paradigm," Journal of Plant Pathology, vol. 103, no. 1, pp. 17–26, Oct. 2020, https://doi.org/10.1007/s42161-020-00683-3.

[10] A. A. Joshi and B. D. Jadhav, "Monitoring and controlling rice diseases using Image processing techniques," IEEE Xplore, Dec. 01, 2016. https://ieeexplore.ieee.org/document/7915015.

[11] K. Ahmed, T. R. Shahidi, S. M. Irfanul Alam and S. Momen, "Rice Leaf Disease Detection Using Machine Learning Techniques," 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2019, pp. 1-5, https://doi.org/10.1109/STI47673.2019.9068096.

[12] Y. Guo et al., "Integrated phenology and climate in rice yields prediction using machine learning methods," Ecological Indicators, vol. 120, p. 106935, Jan. 2021, https://doi.org/10.1016/j.ecolind.2020.106935.

[13] L. R. Mansaray, K. Zhang, and A. S. Kanu, "Dry biomass estimation of paddy rice with Sentinel-1A satellite data using machine learning regression algorithms," Computers and Electronics in Agriculture, vol. 176, p. 105674, Sep. 2020, https://doi.org/10.1016/j.compag.2020.105674.

[14] L. Yu et al., "An integrated rice panicle phenotyping method based on X-ray and RGB scanning and deep learning," Crop Journal, vol. 9, no. 1, pp. 42–56, Feb. 2021, https://doi.org/10.1016/j.cj.2020.06.009.

[15] R. Graf, S. Zhu, and B. Sivakumar, "Forecasting River water temperature time series using a wavelet–neural network hybrid modelling approach," Journal of Hydrology, vol. 578, p. 124115, Nov. 2019, doi: https://doi.org/10.1016/j.jhydrol.2019.124115.

[16] Y. B. Khatraty, N. M. Nauwynck, M. T. Diallo, and M. F. Nanne, "Deep Predictive Models Based on IoT and Remote Sensing Big Time Series for Precision Agriculture," International Journal of Emerging Technology and Advanced Engineering, vol. 12, no. 11, pp. 79–88, Nov. 2022, https://doi.org/10.46338/ijetae1122_09.

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," arXiv (Cornell University), Jul. 2019, https://doi.org/10.48550/arxiv.1907.10902.

[18] J. Bergstra, D. Yamins, and D. D. Cox, "Making a Science of Model Search: Hyperparame-ter Optimization in Hundreds of Dimensions for Vision Architectures," International Conference on Machine Learning, pp. 115–123, Jun. 2013.
ttps://proceedings.mlr.press/v28/bergstra13.html.

[19] X. Wang, Y. Jin, S. Schmitt, and M. Olhofer, "Recent Advances in Bayesian Optimization," arXiv (Cornell University), Jun. 2022, https://doi.org/10.48550/arxiv.2206.03301.

[20] S. Shekhar, A. Bansode and A. Salim, "A Comparative study of Hyper-Parameter Optimization Tools," 2021 IEEE Asia-Pacific Conference on Computer Science and Data

Engineering (CSDE), Brisbane, Australia, 2021, pp. 1-6, https://doi.org/10.1109/CSDE53843.2021.9718485.

[21] M. C. Tourad and A. Abdali, “An Intelligent Similarity Model between Generalized Trapezoidal Fuzzy Numbers in Large Scale,” INTERNATIONAL JOURNAL of FUZZY LOG-IC and INTELLIGENT SYSTEMS, vol. 18, no. 4, pp. 303–315, Dec. 2018, https://doi.org/10.5391/ijfis.2018.18.4.303.

[22] A. Outfarouin, A. Abdali and M. C. Tourad, "Towards a new decisional needs formalization," 2016 IEEE/ACS 13th International Conference of Computer Systems and Applica-tions (AICCSA), Agadir, Morocco, 2016, pp. 1-2, https://doi.org/10.1109/AICCSA.2016.7945757.

[23] Mohamedou Cheikh Tourad, A. Abdali, and O. Ahmad, “On a new index of Publish/Subscribe system in the context of Big Data,” Nov. 2016, https://doi.org/10.1109/aiccsa.2016.7945756.

FIGURES:

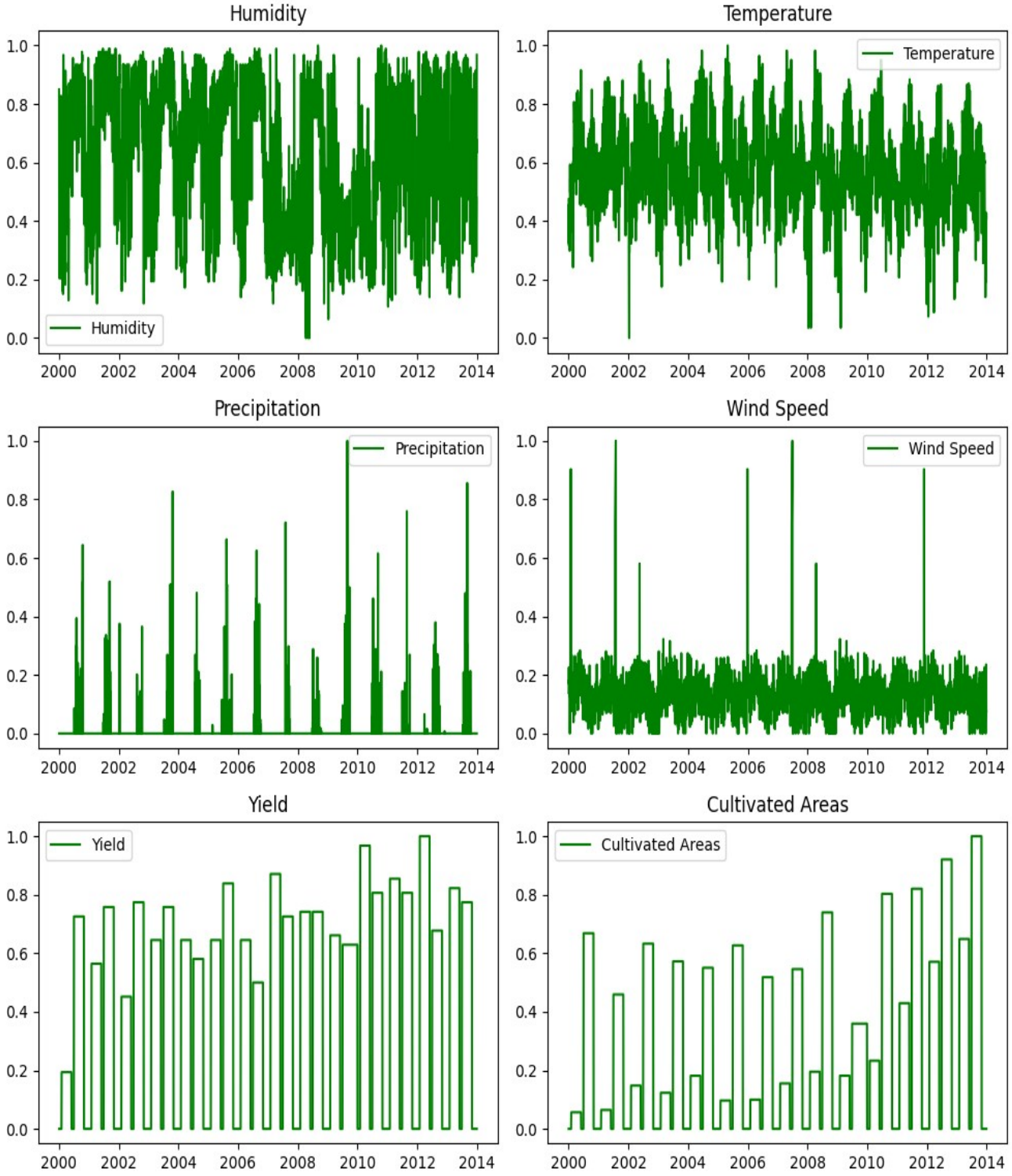

Figure 1: Values Visualization

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	date	humidity	temperature	precipitation	wind_speed	yield	cultivated_areas
0	1/1/2000	85.0	34.0	0.0	4.3	0.0	0.0
1	1/2/2000	45.0	33.5	0.0	5.2	0.0	0.0
2	1/3/2000	38.0	33.0	0.0	6.0	0.0	0.0
3	1/4/2000	34.0	29.7	0.0	7.0	0.0	0.0
4	1/5/2000	25.0	33.5	0.0	6.2	0.0	0.0
...	...	...	...	...	...	...	...
5108	12/26/2013	41.0	26.0	0.0	6.0	0.0	0.0
5109	12/27/2013	42.0	26.2	0.0	7.3	0.0	0.0
5110	12/28/2013	69.0	32.3	0.0	3.0	0.0	0.0
5111	12/29/2013	65.0	32.7	0.0	2.3	0.0	0.0
5112	12/30/2013	96.0	32.2	0.0	2.3	0.0	0.0

5113 rows × 7 columns

Figure 2: Dataset Visualization.

Figure 3: Mauritania Card

Journal of Theoretical and Applied Information Technology
15th April 2024. Vol.102. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

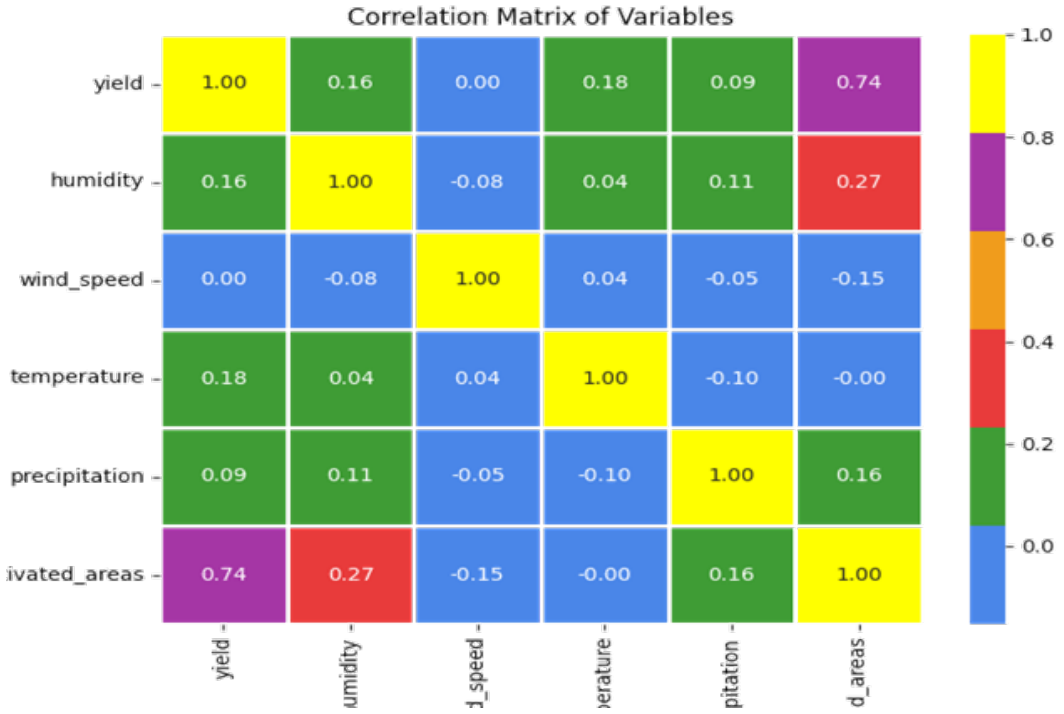

Figure 4: Correlation Metrix

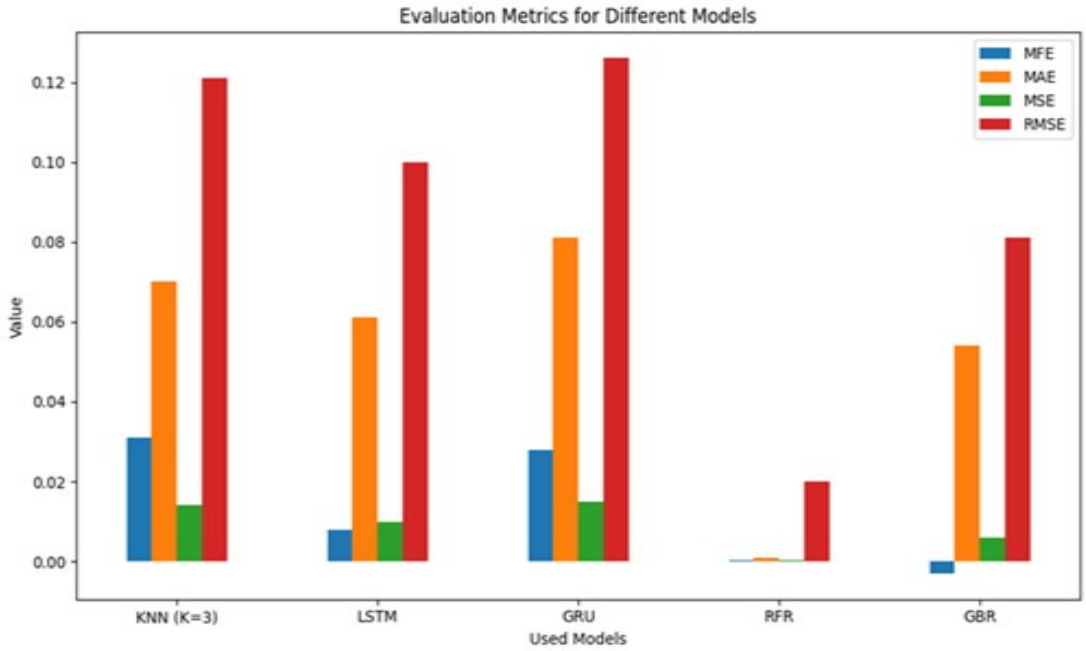

Figure 5: Evaluation Metrics

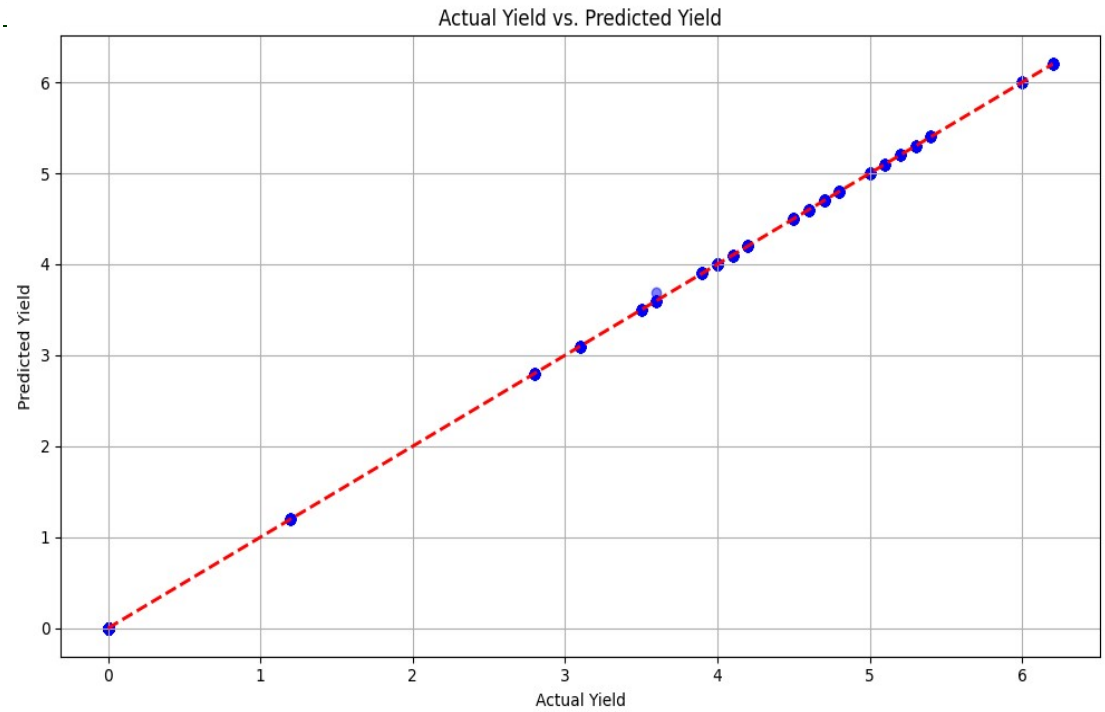

Figure 6: Actual Yield VS Predicted Yield

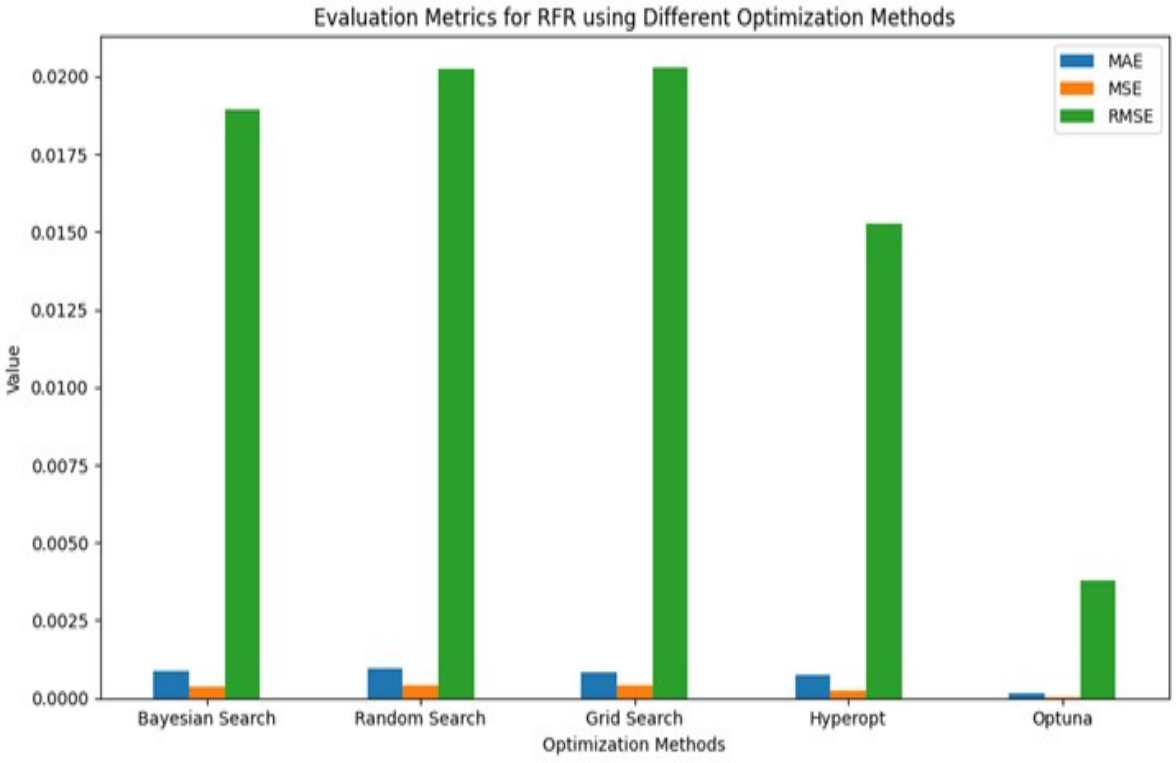

Figure 7: Optimization Methods Evaluation