E-ISSN: 2964-6219, P-ISSN: 2964-6200
Volume 5 Number 1 February 2027.
pp: 99-107. DOI: https://doi.org/10.58477/cj.v5i1.569.

Computer Journal

RESEARCH ARTICLE Open Access

Improving Rice Production Prediction Using XGBoost with Feature Engineering of Weather and Land Characteristics

Syafril Chairiansyah [1], **Sartika Lina Mulani Sitio** [2*]
[1,2*] Department of Informatics Engineering, Universitas Pamulang, South Tangerang City, Banten Province, Indonesia.

Correspondence email:
dosen02681@unpam.ac.id

Article History
Received: 7 October 2026.
Revised: 9 October 2026.
Accepted: 2 November 2026.
Available Online: 5 December 2026
Published: 1 February 2027.

Funding Information
Universitas Pamulang.

Abstract
Rice production prediction is needed to support the use of weather information and land characteristics in estimating production yields more measurably. This study aims to develop a rice production prediction model using XGBoost by utilizing weather factors and land characteristics engineering features. The dataset used consisted of 243 observations with 84 initial variables, which were divided into training and testing data with an 80:20 ratio. Feature engineering produced derivative features in the form of aggregation of rainfall, temperature, humidity, solar radiation, as well as nutrient ratios and soil texture. Furthermore, feature selection using XGBoost selected the 30 most relevant features to build an improved model. Hyperparameter optimization was performed using Optuna, while model performance was evaluated using MAE, RMSE, and R^2. The test results showed that the baseline model obtained an MAE of 0.9049, an RMSE of 1.1576, and an R^2 of 0.5843. The improved model produced an MAE of 0.9717, an RMSE of 1.2729, and an R^2 of 0.4974, indicating no performance improvement compared to the baseline. SHAP analysis showed that temp_Dec, N_to_K_Ratio, bulk density, and TMIN_Mean were the features that contributed most to the prediction.

Keywords: Rice Production; XGBoost; Feature Engineering; Feature Selection; SHAP.

Abstrak
Prediksi produksi padi diperlukan untuk mendukung pemanfaatan informasi cuaca dan karakteristik lahan dalam memperkirakan hasil produksi secara lebih terukur. Penelitian ini bertujuan mengembangkan model prediksi produksi padi menggunakan XGBoost dengan memanfaatkan *feature engineering* faktor cuaca dan karakteristik lahan. Dataset yang digunakan terdiri atas 243 observasi dengan 84 variabel awal, yang dibagi menjadi data pelatihan dan pengujian dengan rasio 80:20. *Feature engineering* menghasilkan fitur turunan berupa agregasi curah hujan, temperatur, kelembapan, radiasi matahari, serta rasio unsur hara dan tekstur tanah. Selanjutnya, *feature selection* menggunakan XGBoost memilih 30 fitur yang paling relevan untuk membangun model *improved*. Optimasi *hyperparameter* dilakukan menggunakan Optuna, sedangkan kinerja model dievaluasi menggunakan MAE, RMSE, dan R^2. Hasil pengujian menunjukkan bahwa model *baseline* memperoleh MAE 0,9049, RMSE 1,1576, dan R^2 0,5843. Model *improved* menghasilkan MAE 0,9717, RMSE 1,2729, dan R^2 0,4974, sehingga belum menunjukkan peningkatan kinerja dibandingkan *baseline*. Analisis SHAP menunjukkan bahwa temp_Dec, N_to_K_Ratio, *bulk density*, dan TMIN_Mean merupakan fitur yang paling berkontribusi terhadap prediksi.

Kata Kunci: Produksi Padi; XGBoost; Feature Engineering; Feature Selection; SHAP.

© The Author(s) 2027. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

1. Introduction

Rice production is a key component of food-system sustainability, because changes in yield affect food availability, distribution planning, and decision-making in the agricultural sector. Production is shaped not only by cultivation practices but also by environmental conditions during the growing period, as temperature, rainfall, solar radiation, and soil properties interact in determining crop development and yield. Studies on the response of rice yield to meteorological conditions indicate that minimum and maximum temperatures and solar radiation relate to yield differently across growth phases (Wirakusuma et al., 2025). These findings show that predicting rice production is a complex problem, since the relationship between environmental variables and output is not always linear. The development of machine learning provides an alternative for modeling such complex relationships. Unlike conventional statistical approaches, which generally require assumptions about the shape of relationships between variables, machine-learning algorithms can learn nonlinear patterns directly from the data. In crop-yield modeling, meteorological variables such as temperature and precipitation are important predictors, and feature selection is necessary because environmental variables are often highly intercorrelated (Lischeid et al., 2022). In rice production specifically, machine learning has shown potential to produce pre-harvest predictions; for example, a study using Sentinel-2 imagery and several machine-learning algorithms demonstrated that time-series data can predict rice yields at the field level before harvest (Son et al., 2022). Extreme Gradient Boosting (XGBoost) is a tree-based ensemble learning algorithm that builds predictive models through gradient boosting, sequentially adding decision trees to improve performance. Tree-ensemble methods such as XGBoost perform strongly on structured tabular data, making them suitable for regression problems with heterogeneous predictors (Shwartz-Ziv & Armon, 2022). In agricultural applications, XGBoost has been used to estimate crop yields from environmental and remote-sensing data (Huber et al., 2022), to forecast annual rice production (Noorunnahar et al., 2023), and in studies that emphasize the importance of dataset characteristic analysis and parameter tuning to obtain a well-fitted model (Wijayanti et al., 2024).

Although XGBoost can capture nonlinear relationships, the quality of data representation remains a decisive factor in modeling. Feature engineering forms new variables that describe data characteristics more informatively, for example through aggregation, transformation, or the construction of ratios between variables. In crop-yield modeling, combining meteorological features with static features such as soil characteristics has been applied to capture yield variation arising from different environmental factors (Liu et al., 2022). Research on head-rice-yield prediction in Australia likewise shows that dataset construction, aggregation of meteorological conditions, pre-processing, and feature selection can strongly affect XGBoost performance (Clarke et al., 2024). Together, these findings indicate that the same algorithm does not necessarily deliver the same performance when the underlying data representation differs. In rice production, soil characteristics must be considered alongside weather factors, because plants derive their growing conditions from the interaction between the atmospheric environment and the soil medium. Research that integrates meteorological, agro-chemical, and soil-physiographic factors shows that the variables affecting yield response differ across rice ecotypes and can be modeled with XGBoost (Ahmed et al., 2022), while studies combining SAR, optical, and meteorological data show that integrating multiple information sources can improve yield predictability (Yu et al., 2023). More recent work on climate- and remote-sensing-based yield prediction confirms that machine-learning approaches perform well, but that success still depends on data characteristics and the combination of predictors used (De Clercq & Mahdi, 2024). A further challenge lies in interpreting the model. Ensemble tree-based models such as XGBoost produce accurate predictions, yet the decision process behind a single prediction is difficult to explain from the output value alone—a concern in agriculture, where users need not only yield estimates but also information about the variables that drive those estimates. SHAP (SHapley Additive exPlanations) offers a framework for attributing each feature's contribution to the model output Based on these gaps, this study analyzes the ability of XGBoost to predict rice production by applying feature engineering to weather factors and land characteristics. The baseline XGBoost model is compared with an improved model that incorporates engineered features and feature selection. Hyperparameter optimization was performed using Optuna, and model performance was evaluated using MAE, RMSE, and R^2. SHAP was then used to interpret feature contributions and identify the factors that dominate the prediction. In this way, the study provides an empirical assessment of whether adding representations of weather and land characteristics yields additional predictive information for XGBoost on the dataset used.

2. Methodology

This study uses a rice production dataset with Yield Smart-Valleys as the target variable. The initial dataset consists of 243 observations and 84 columns: 63 baseline predictor features (15 soil and 48 weather), one target variable, and 20 additional columns excluded from the baseline predictors. Data preprocessing includes structural and quality checks, handling of missing values, detection of infinite and invalid values, and removal of duplicate records, with missing-value imputation parameters estimated using the training data only. The dataset is divided into training and test sets at an 80:20 ratio, yielding 194 training and 49 test observations under a fixed random seed of 42. Two XGBoost models are developed—a baseline model using the original predictor features and an improved model incorporating engineered features and feature selection—as boosting ensemble methods are widely applied to prediction tasks (Sitio et al., 2026). Hyperparameter optimization is performed using Optuna to minimize the mean validation RMSE rather than the training error. To reduce overfitting and prevent data leakage, a nested five-fold cross-validation protocol is applied within the training set, in which the outer folds assess model performance and the inner folds drive hyperparameter optimization (Jayanthi et al., 2026). Feature selection and data-dependent preprocessing are carried out separately within each training fold so that validation observations never influence model development. Search-based hyperparameter tuning has likewise been applied to agricultural prediction, such as soil-nutrient models, demonstrating its benefit for model performance (Dada et al., 2025). Feature engineering generates derived variables from weather factors and soil characteristics, including statistical summaries of rainfall, temperature, humidity, and solar radiation, as well as nutrient indices, nutrient ratios, and soil-texture indicators; comparable feature-engineering strategies have been used for early rice-yield prediction (Mugemangango et al., 2026). The number of engineered features and the resulting predictor count were cross-checked against the implemented code to ensure consistency. XGBoost feature importance is then used to select the 30 highest-ranked features within each relevant training fold, following feature-selection practices established for yield modeling (Sánchez et al., 2025). Optuna performs 50 optimization trials over the following search spaces: n_estimators (100–800), max_depth (2–8), learning_rate (0.01–0.20, logarithmic scale), subsample (0.60–1.00), colsample_bytree (0.60–1.00), min_child_weight (1–10), gamma (0–1), reg_alpha (0.0001–10, logarithmic scale), and reg_lambda (0.001–10, logarithmic scale). A fixed seed of 42 is used for reproducibility, including the Optuna sampler. After model development, the selected model is refitted on the complete training set and evaluated on the held-out test set using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2), as is standard for gradient-boosting models in yield prediction (Mahesh & Soundrapandiyan, 2024); the test set is excluded from preprocessing parameter estimation, feature selection, and hyperparameter optimization. Model interpretation is then carried out using SHAP (SHapley Additive exPlanations) to identify each feature's contribution to the predictions, with global contribution summarized by mean absolute SHAP values and the direction of each contribution examined through a beeswarm plot, an approach also used in rice-yield forecasting (Brinkhoff et al., 2024). These interpretations describe model behavior and should not be read as evidence of causal relationships between the predictors and rice production. Based on the SHAP results, features are grouped into weather factors and land characteristics to determine which group contributes most to the predictions. The overall workflow is summarized in Figure 1.

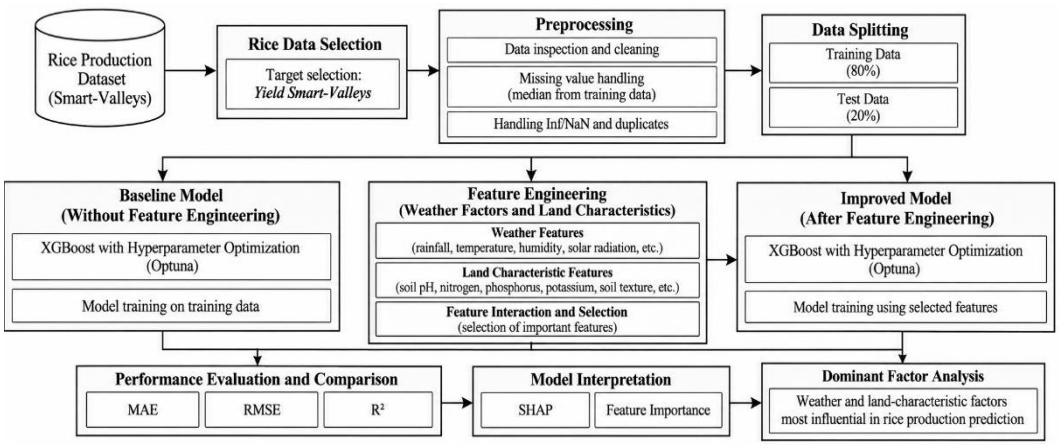

Figure 1. Research Methodology

3. Results and Discussion

The dataset used in this study consisted of 243 observations and 84 columns in the initial data structure. Of these, 63 features were used as baseline predictors, comprising 15 soil characteristic features and 48 weather features, with Yield Smart-Valleys as the target variable. The remaining columns were excluded from the baseline predictor set because they were not part of the selected soil and weather predictors. The dataset was divided into training and testing sets using an 80:20 ratio, resulting in 194 training observations and 49 testing observations. The structural composition of the dataset is summarized in Table 1.

Table 1. Dataset Characteristics

Component	Quantity
Number of observations	243
Initial dataset columns	84
Soil predictor features	15
Weather predictor features	48
Baseline predictor features	63
Target variable	Yield Smart-Valleys
Training observations (80%)	194
Testing observations (20%)	49
Engineered features added	20
Predictor features after feature engineering	83
Features selected for the improved model	30

3.1 Feature Engineering

Feature engineering was applied to derive additional features from weather variables and soil characteristics. A total of 20 new features were generated, resulting in 83 predictor features for the improved model after duplicate columns were removed. The derived features represent rainfall and temperature statistics, solar radiation, humidity, nutrient ratios, and soil texture characteristics. The complete list of engineered features is presented in Table 2.

Table 2. Feature Engineering Results

Groups	Resulting features
Rainfall	Rainfall_Total, Rainfall_Mean, Rainfall_Std, Rainfall_Max
Temperature	Temperature_Mean, Temperature_Std, Temperature_Max, Temperature_Min
Minumum and Maximum temperature Aggregation	TMIN_Mean, TMAX_Mean
Variasi Temperatur	Temperature_Range
Solar Radiation	SolarRadiation_Mean, SolarRadiation_Total
Kelembapan	Humidity_Mean, Humidity_Std
Soil Nutrients	NPK_Index, N_to_P_Ratio, N_to_K_Ratio
Soil Texture	Soil_Texture_Total, Clay_Sand_Ratio
Total engineered features	20

The engineered features provide additional representations of weather conditions and soil characteristics. Weather-related features summarize the distribution and variation of environmental measurements, while soil-related features capture nutrient relationships and texture composition. Following feature engineering, XGBoost-based feature selection was applied to identify the 30 most relevant predictors for the improved model.

3.2 Feature Selection

Feature selection was applied to reduce the number of predictors and identify features that contributed most to the XGBoost model. The improved dataset contained 83 predictor features after feature engineering and the removal of duplicate columns. The 30 most important features were selected based on XGBoost feature importance scores computed within each training fold, ensuring that validation and test data did not influence feature ranking. Table 3 presents the ten highest-ranked features and their corresponding

importance scores.

Table 3. Top Ten Features of Feature Selection Results

Rank	Feature	Importance
1	TMIN_Mean	0.155690
2	SRAD_Nov	0.125297
3	temp_Dec	0.068757
4	TMIN_Nov	0.068245
5	temp_Nov	0.053365
6	TMAX_Jul	0.031399
7	SRAD_Jul	0.031249
8	temp_Oct	0.030119
9	RAIN_Nov	0.028587
10	Solar radiation season	0.025417

Among the selected features, TMIN_Mean received the highest importance score, followed by SRAD_Nov, temp_Dec, TMIN_Nov, and temp_Nov. These results indicate that temperature and solar radiation features dominated the fitted XGBoost model. However, feature importance reflects the model's predictive reliance on these variables and does not establish a causal relationship with rice yield. The 30 selected features served as predictors for the improved model.

3.3 Baseline vs. Improved Model Performance

Performance comparison was conducted to determine the effect of feature engineering and feature selection on XGBoost's ability to predict rice production. The baseline model used the original 63 predictor features, while the improved model used the 30 features selected after engineering. Evaluation was carried out on the held-out test set using MAE, RMSE, and R^2, as presented in Table 4.

Table 4. Comparison of XGBoost Baseline and Improved Performance

Model	MAE	RMSE	R^2
XGBoost Baseline	0.9049	1.1576	0.5843
XGBoost Improved	0.9717	1.2729	0.4974

The baseline model produced lower prediction errors than the improved model across all three metrics on the reported test set. Because MAE and RMSE are expressed in the same unit as the target variable, their absolute magnitude depends on the measurement scale of Yield Smart-Valleys The R^2 difference (0.5843 vs. 0.4974) indicates that the engineered feature subset captured less variance in the target than the full baseline set. These findings suggest that, under the current experimental configuration, the additional feature representations and the reduced feature subset did not improve predictive performance.

3.4 Actual vs. Predicted Values

The comparison between actual and predicted values was used to assess how closely the improved model's predictions track observed production and to identify systematic deviation patterns. Figure 2 presents this relationship for the improved XGBoost model.

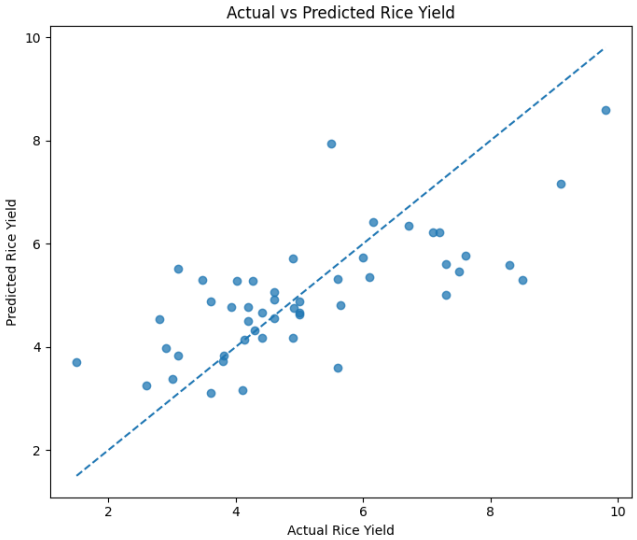

Figure 2. Actual vs. Predicted Rice Yield (Improved XGBoost Model)

Points closer to the diagonal indicate predictions that approximate the actual value, while greater distance from the line reflects larger prediction errors. The scattering pattern in Figure 2 complements the MAE and RMSE values reported in Table 4, confirming that prediction accuracy is uneven across the range of observed yields.

3.5 Residual Analysis

Residual analysis was performed to evaluate the prediction error patterns generated by the improved XGBoost model. The residual was calculated as the difference between the actual and predicted values and plotted against the predicted value, with a horizontal line at zero residual serving as reference. A positive residual indicates that the actual value exceeds the prediction, while a negative residual indicates overprediction. The results are shown in Figure 3.

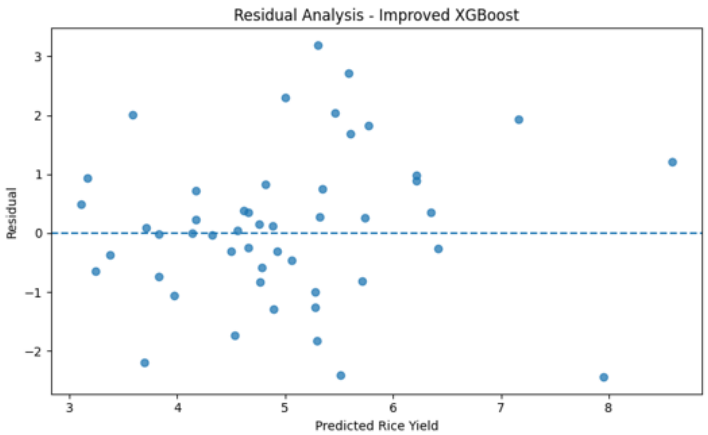

Figure 3. Residual Analysis — Improved XGBoost Model

The residuals in Figure 3 are scattered on both sides of the zero line, with most observations falling within the predicted-value range of approximately 4–6. This spread indicates that the model produces a mix of underprediction and overprediction rather than a consistent directional bias. Some observations exhibit fairly large positive residuals (approximately 3.2), while the most negative residual reaches around −2.4. Although the residuals are not entirely skewed to one side, their wide dispersion shows that the model has not achieved consistent prediction errors across all observations.

3.6 Feature Influence Analysis Based on SHAP

Feature influence was analyzed using SHAP to identify the contribution and direction of each feature's effect on rice production prediction. The mean absolute SHAP value measured the magnitude of each

feature's global contribution, while the beeswarm visualization displayed the distribution of SHAP values and the relationship between feature values and their directional effect on predictions. Figure 4 presents the top features ranked by mean absolute SHAP value, and Figure 5 displays the beeswarm pattern.

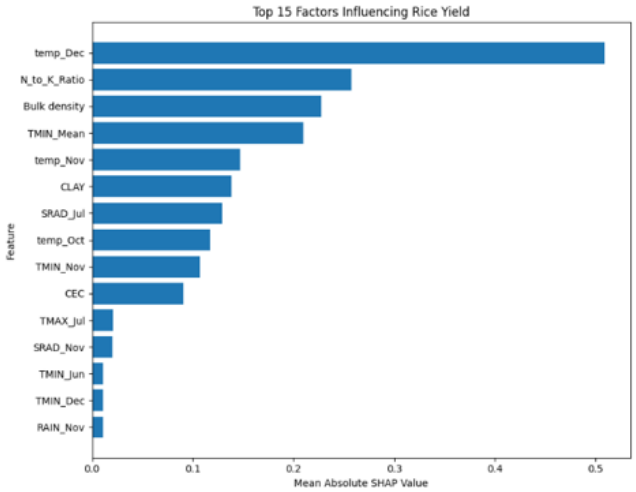

Figure 4. Model Interpretation Based on Mean Absolute SHAP Values

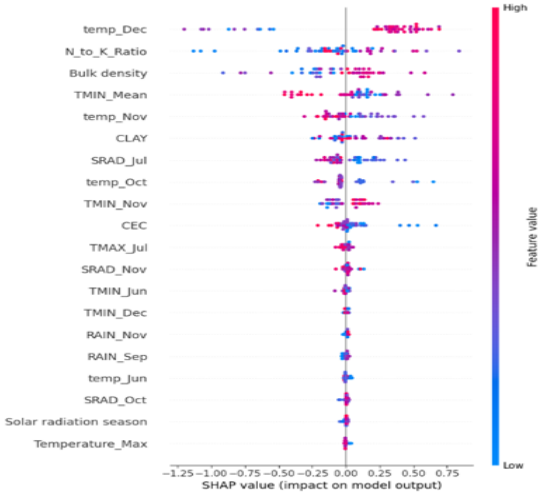

Figure 5. SHAP Beeswarm Plot of Features Influencing Rice Production Prediction

According to Figure 4, temp_Dec has the highest mean absolute SHAP value (0.509191), followed by N_to_K_Ratio (0.257290), bulk density (0.227797), and TMIN_Mean (0.209645). These features contribute more to the model's predictions than the remaining variables. In the beeswarm plot (Figure 5), red points indicate relatively high feature values and blue points indicate relatively low feature values. The position of each point relative to the zero line indicates the direction of its SHAP contribution: positive values push the prediction higher, while negative values push it lower.

3.7 SHAP Values per Feature Group

To determine which broad category of predictors dominates the model, features were grouped into weather factors and soil/land characteristics based on their SHAP contributions. The aggregated mean absolute SHAP values for each group are presented in Table 5.

Table 5. Mean Absolute SHAP Value per Feature Group

Group/Factor	Mean Absolute SHAP Value
Weather	1.371637
Soil/Land	0.714946

The weather factor group obtained a substantially larger aggregate contribution (1.371637) than the soil/land group (0.714946), indicating that weather-related features dominated the model under the analysis performed. Nevertheless, several land-characteristic features such as N_to_K_Ratio, bulk density, CLAY, and CEC retained relatively high individual contributions, showing that the model still utilizes a combination of weather and land information. This finding should be interpreted within the scope of the dataset and the fitted model and does not establish a causal relationship between weather conditions and rice production.

4. Conclusion

This study evaluated XGBoost for rice production prediction using weather factors and land characteristics, with feature engineering and feature selection applied to the same dataset. On the held-out test set, the baseline model outperformed the improved model, achieving an MAE of 0.9049, an RMSE of 1.1576, and an R^2 of 0.5843, compared with an MAE of 0.9717, an RMSE of 1.2729, and an R^2 of 0.4974 for the improved model. This indicates that the engineered features and the reduced feature subset did not improve predictive performance under the current experimental configuration. SHAP analysis showed that both weather-related and soil/land features contributed to the predictions, with weather factors dominating

in aggregate while several land-characteristic features remained individually prominent. These results describe the behavior of the fitted model and should not be read as evidence of causal relationships between the predictors and rice production. Given the modest predictive performance and the limited dataset, the model is best regarded as an exploratory approach rather than a validated decision-support tool for agricultural management. Its practical use requires evaluation on larger and more geographically diverse data, validation across different growing seasons, and independent testing under field conditions. Future work should align feature construction with rice growth stages, examine the stability of feature importance across validation folds, and benchmark alternative models under a rigorous validation protocol. Together, these steps would clarify whether the approach can reach the reliability needed for agricultural planning and decision support.

References

Ahmed, M. S., Tazwar, M. T., Khan, H., Roy, S., Iqbal, J., Rabiul Alam, M. G., Hassan, M. R., & Hassan, M. M. (2022). Yield response of different rice ecotypes to meteorological, agro-chemical, and soil physiographic factors for interpretable precision agriculture using extreme gradient boosting and support vector regression. Complexity, 2022, https://doi.org/10.1155/2022/5305353

Brinkhoff, J., Clarke, A., Dunn, B. W., & Groat, M. (2024). Analysis and forecasting of Australian rice yield using phenology-based aggregation of satellite and weather data. Agricultural and Forest Meteorology, 353, 110055. https://doi.org/10.1016/j.agrformet.2024.110055

Clarke, A., Yates, D., Blanchard, C., Islam, M. Z., Ford, R., Rehman, S., & Walsh, R. (2024). The effect of dataset construction and data pre-processing on the eXtreme Gradient Boosting algorithm applied to head rice yield prediction in Australia. Computers and Electronics in Agriculture, 219, 108716. https://doi.org/10.1016/j.compag.2024.108716

Dada, B. A., Nwulu, N. I., & Olukanmi, S. O. (2025). Bayesian optimization with Optuna for enhanced soil nutrient prediction: A comparative study with genetic algorithm and particle swarm optimization. Smart Agricultural Technology, 12, 101136. https://doi.org/10.1016/j.atech.2025.101136

De Clercq, D., & Mahdi, A. (2024). Feasibility of machine learning-based rice yield prediction in India at the district level using climate reanalysis and remote sensing data. Agricultural Systems, 220, 104099. https://doi.org/10.1016/j.agsy.2024.104099

Erlin, Yunianta, A., Wulandhari, L. A., Desnelita, Y., Nasution, N., & Junadhi. (2024). Enhancing rice production prediction in Indonesia using advanced machine learning models. IEEE Access, 12, 151161–151177. https://doi.org/10.1109/ACCESS.2024.3478738

Huber, F., Yushchenko, A., Stratmann, B., & Steinhage, V. (2022). Extreme gradient boosting for yield estimation compared with deep learning approaches. Computers and Electronics in Agriculture, 202, 107346. https://doi.org/10.1016/j.compag.2022.107346

Jayanthi, S., Suvarna Kumari, T., Inturi, S., Nathan, B., Josephine Sathya, M. A., & Karmakonda, K. (2026). Comparative evaluation of hyperparameter tuning methods in crop yield prediction for enhancing food security. International Journal of Safety and Security Engineering, 16(3), 479–489. https://doi.org/10.18280/ijsse.160301

Lischeid, G., Webber, H., Sommer, M., Nendel, C., & Ewert, F. (2022). Machine learning in crop yield modelling: A powerful tool, but no surrogate for science. Agricultural and Forest Meteorology, 312, 108698. https://doi.org/10.1016/j.agrformet.2021.108698

Liu, Q., Yang, M., Mohammadi, K., Song, D., Bi, J., & Wang, G. (2022). Machine learning crop yield models based on meteorological features and comparison with a process-based model. Artificial Intelligence for the Earth Systems, 1(4), 1–14. https://doi.org/10.1175/aies-d-22-0002.1

Mahesh, P., & Soundrapandiyan, R. (2024). Yield prediction for crops by gradient-based algorithms. PLoS ONE, 19(8), 1–20. https://doi.org/10.1371/journal.pone.0291928

Mugemangango, C., Nzabanita, J., Ndaruhuye Muhoza, D., & Cahill, N. D. (2026). Leveraging machine learning and feature engineering techniques for early rice yield prediction in Rwanda using vegetation indices. African Journal of Applied Research, 12(5), 361–380. https://doi.org/10.26437/3qt42n09

Noorunnahar, M., Chowdhury, A. H., & Mila, F. A. (2023). A tree-based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. PLoS ONE, 18(3), 1–15. https://doi.org/10.1371/journal.pone.0283452

Sánchez, J. C. M., Mesa, H. G. A., Espinosa, A. T., Castilla, S. R., & Lamont, F. G. (2025). Improving wheat yield prediction through variable selection using support vector regression, random forest, and extreme gradient boosting. Smart Agricultural Technology, 10, 100791. https://doi.org/10.1016/j.atech.2025.100791

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Sitio, S. L. M., (2026). Comparative analysis of bagging and boosting models in ensemble learning for graduation prediction. 11(3), 979–986. https://doi.org/10.33480/jitk.v11i3.7579

Son, N.-T., Chen, C.-F., Cheng, Y.-S., Toscano, P., Chen, C.-R., Chen, S.-L., Tseng, K.-H., Syu, C.-H., Guo, H.-Y., & Zhang, Y.-T. (2022). Field-scale rice yield prediction from Sentinel-2 monthly image composites using machine learning algorithms. Ecological Informatics, 69, 101618. https://doi.org/10.1016/j.ecoinf.2022.101618

Wijayanti, E. B., Setiadi, D. R. I. M., & Setyoko, B. H. (2024). Dataset analysis and feature characteristics to predict rice production based on eXtreme Gradient Boosting. Journal of Computing Theories and Applications, 1(3), 299–310. https://doi.org/10.62411/jcta.10057

Wirakusuma, K. A., (2025). Optimasi model prediksi Extreme Gradient Boosting dengan genetic algorithm untuk produksi dan produktivitas padi. STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer, 4(4),https://doi.org/10.55123/storage.v4i4.6633

Yu, W., Yang, G., Li, D., Zheng, H., Yao, X., Zhu, Y., Cao, W., Qiu, L., & Cheng, T. (2023). Improved prediction of rice yield at field and county levels by synergistic use of SAR, optical, and meteorological data. Agricultural and Forest Meteorology, 342, 109729. https://doi.org/10.1016/j.agrformet.2023.109729