JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8335

COMPARISON OF TREE-BASED ENSEMBLE LEARNING ALGORITHMS FOR PREDICTING RICE PRODUCTION TO SUPPORT FOOD SECURITY

Dahri Yani Hakim Tanjung[1*], Firman Syahputra[2], Rofiqoh Dewi[1], Hilda Elsera Br Sembiring[1]

Universitas Satya Terra Bhinneka, Medan, Indonesia[1]
https://satyaterrabhinneka.ac.id[1]
dahritanjung@satyaterrabhinneka.ac.id[*], rofiqohdewi@satyaterrabhinneka.ac.id,
hildaelsera@satyaterrabinneka.ac.id

Universitas Battuta, Medan, Indonesia[2]
https://battuta.ac.id[2]
24firman@gmail.com

(*) Corresponding Author
(Responsible for the Quality of Paper Content)



The creation is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

Submitted: 28 February 2026, **Revised:** 02 September 2026, **Accepted:** 09 November 2026, **Published:** 13 November 2026.

Abstract— *Rice production is a key component of national food security, requiring accurate prediction models to support agricultural development planning. This study evaluated four tree-based ensemble algorithms: CatBoost, Random Forest, XGBoost, and LightGBM to determine the most effective method for predicting rice production in various regions of Indonesia. The dataset includes harvested area, production volume, rainfall, temperature, humidity, and provincial attributes representing agroclimatic and spatial variability. Preprocessing stages included data cleaning, feature encoding, normalization, and splitting using five-fold cross-validation. Model performance was assessed using RMSE, MAE, and R^2 to describe accuracy, stability, and explanatory power. The results showed that XGBoost performed best with an RMSE of 118.3, MAE of 82.9 ton, and R^2 of 0.74, outperforming all other models. Residual analysis confirmed the stability of XGBoost predictions through a narrower error distribution close to zero. SHAP analysis showed that harvested area was the most dominant predictor, while rainfall, temperature, humidity, and provincial factors had significant influences. This research indicates that XGBoost is capable of producing reliable, explainable, and highly accurate predictions for modeling rice production. It also emphasizes the importance of integrating environmental, agronomic, and spatial variables in predictive model development to ensure robust performance across diverse cultivation conditions, while supporting productivity enhancement, risk mitigation, improved monitoring, and evidence-based policy formulation for long-term food security. Therefore, this research provides a strong scientific foundation for the use of machine learning-based analytics to support strategic planning, improve efficiency, strengthen resilience, drive technological innovation, and accelerate the transformation of rice production systems.*

Keywords: *Food security, prediction, rice production, tree-based ensemble learning algorithm, XGBoost.*

INTRODUCTION

Rice production is a fundamental component in maintaining Indonesia's national food security, as the majority of the population relies on rice as its primary carbohydrate source. However, over the past two decades, rice production has experienced significant fluctuations due to climate change, rainfall variability, irregular planting seasons, soil quality degradation, and the increasing intensity of hydrometeorological disasters [1] [2] [3] [4]. This instability has affected national food availability, leading to potential supply deficits, price volatility, and increased food insecurity risks across various regions [5] [6]. Therefore, accurate, data-driven rice production prediction models capable of capturing nonlinear patterns are essential to support national food security strategies [7].



Conventional prediction models such as linear regression often fail to capture the complexity of relationships among agronomic, climatic, and spatial variables, resulting in low predictive accuracy [8] [9]. In this context, machine learning offers a more adaptive approach. Numerous studies have shown that decision tree–based ensemble techniques are better able to handle heterogeneous data, high nonlinearity, and complex feature interactions compared to traditional statistical methods [10] [11] [12]. This potential positions ensemble models as strong candidates for developing artificial intelligence–based rice yield prediction systems.

Various studies have applied ensemble methods to crop yield prediction. Random Forest has been shown to outperform linear regression with an RMSE of 0.42 tons/ha in rice yield prediction [12] [13]. XGBoost achieved an MAE that was 11% lower than Random Forest in corn yield prediction [14]. LightGBM has also been applied to wheat yield prediction with an R^2 of 0.91 [15]. Other studies reported that CatBoost demonstrated the most stable performance on categorical regional data for soybean and rice yield prediction[16] [17]. Given Indonesia's high agroclimatic and geographic heterogeneity, a comprehensive analysis is required to evaluate the performance of ensemble models in predicting rice production using integrated climatic and agronomic variables. Such an evaluation should include RMSE, MAE, R^2, and computational efficiency to identify the most accurate and efficient model. To date, no study has simultaneously and quantitatively compared CatBoost, Random Forest, XGBoost, and LightGBM using strict cross-validation on national-scale rice production data [18] [19] [20] [21].

Recent studies have demonstrated that machine learning algorithms can significantly improve agricultural yield prediction accuracy by capturing complex nonlinear relationships among climatic and agronomic variables. Ensemble learning methods consistently outperform traditional statistical approaches in crop yield prediction tasks [22] [23]. Furthermore, the growing adoption of machine learning techniques for agricultural forecasting and decision support systems. However, most previous studies focused on specific crops such as wheat and corn, evaluated only one or two algorithms, or emphasized predictive accuracy without considering model stability and interpretability [24]. In addition, comparative studies involving CatBoost, Random Forest, XGBoost, and LightGBM under the same validation framework for rice production prediction remain limited, particularly in datasets representing heterogeneous agroclimatic conditions. Therefore, a comprehensive benchmarking analysis is required to identify the most accurate, stable, and interpretable ensemble learning model for rice production prediction.

Table 1. The Comparison with Previous Studies

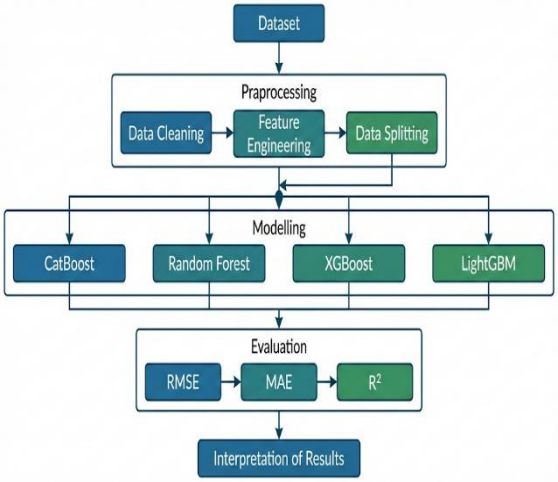
Reference	Algorithm	Metric	Accuracy
[13]	RF vs LR	RMSE 0.42	Low RMSE → High accuracy (≈ 85–90%)
[14]	XGB vs RF	MAE decreased by 11%	Accuracy increased by 11% compared to baseline
[15]	LightGBM	R^2=0.91	91% prediction accuracy
[16]	CatBoost	RMSE 0.38	Small RMSE → Very high accuracy (≈ 90%)
[17]	XGB-SHAP	R^2=0.95	95% prediction accuracy
[18]	RF	MAE 0.27	Very small MAE → High accuracy (≈ 90%)
[19]	Boosting	R^2=0.92	92% accuracy
[25]	LightGBM	RMSE 0.51	Moderate → Accuracy (≈ 80–85%)

Source: (Research Result, 2026)

Based on this research gap, this study aims to conduct a comprehensive benchmarking of four decision tree–based ensemble algorithms CatBoost, Random Forest, XGBoost, and LightGBM using rice production data from Sumatra Island. The evaluation employs five-fold cross-validation to obtain robust performance estimates. The results are expected to provide recommendations for the most suitable algorithm for rice yield prediction systems, supporting national food security and data-driven decision-making at both governmental and agro-industry levels [26][27][28].

MATERIALS AND METHODS

Research Methodology



Source: (Research Result, 2026)

Figure 1. Research Methodology Flowchart

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8335

This study adopts a comparative experimental (benchmarking) approach to evaluate four tree-based ensemble algorithms for regression tasks. The overall research methodology flow is illustrated in Figure 1.

Dataset

The dataset used in this study consists of 224 records collected from a publicly available Kaggle repository. The data comprise annual observations of rice production across provinces in Sumatra Island and include variables such as harvested area, production volume, rainfall, humidity, and average temperature. Since the dataset combines observations from multiple provinces over several years, it can be categorized as panel data containing both cross-sectional and temporal dimensions. The dataset is structured based on annual observations, making it time-series in nature, while also being cross-sectional because each record represents a specific province in a given year. Therefore, the dataset can be classified as panel data, which is highly relevant for food production modeling based on climatic factors.

Preprocessing

The preprocessing stage was conducted to ensure the dataset met quality standards prior to modeling. This process began with data cleaning, which involved identifying and handling data quality issues such as missing values, duplicates, and inconsistencies among variables. Data completeness was checked for each column, particularly for critical variables such as harvested area, production, rainfall, humidity, and average temperature. When missing values were identified, appropriate handling methods were applied, such as mean imputation for numerical data and mode imputation for categorical data. Next, feature engineering was performed to add, transform, or select features to enhance the model's ability to capture data patterns.

Modelling

CatBoost

CatBoost is a tree-based boosting algorithm that is highly effective in handling categorical data and mitigating overfitting through ordered boosting and permutation-driven encoding [29] [30].

$$CTR = \frac{sum\ of\ targets}{count\ of\ occurences} \quad (1)$$

CatBoost has been shown to perform well on agricultural datasets rich in categorical features such as province, crop variety, or planting patterns [31].

Random Forest

Random Forest is a bagging-based ensemble method that combines multiple decision trees trained on randomly generated bootstrap samples [32].

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (2)$$

The final prediction is obtained by averaging the outputs of all trees, making Random Forest more stable and resistant to overfitting. This method is widely used in crop yield prediction due to its robustness against noise, outliers, and multicollinearity [9] [21] [33] [34].

XGBoost

eXtreme Gradient Boosting (XGBoost) is an ensemble learning algorithm that utilizes the boosting method to produce a robust and accurate model [35].

$$Obj = \sum_i l(y_i, \hat{y}i) + \sum_k \Omega(f_k) \quad (3)$$

$$\Omega(f_k) = \alpha T + \frac{1}{2}\lambda\left|\left|\omega\right|\right|^2 \quad (4)$$

It is known for its efficiency, speed, and stability, particularly when applied to large-scale agricultural datasets [36] [37].

LightGBM

LightGBM employs a leaf-wise tree growth strategy, which is faster than the traditional level-wise approach [36].

$$Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_L^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) \quad (5)$$

However, high MAE values may indicate significant prediction deviations.

Evaluation Metrics

Root Mean Square Error (RMSE)

measures the square root of the average squared differences between actual and predicted values. RMSE penalizes large prediction errors more heavily due to the squaring operation, making it particularly useful for identifying models that produce substantial deviations from actual observations [38].

CC BY NC

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

Where y_i is the actual value and $\hat{y}_i$ is the predicted value.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

Mean Absolute Error (MAE) measures the average absolute difference between actual and predicted values. Lower MAE values indicate better prediction accuracy, whereas high MAE values may indicate significant prediction deviations [32] [36] [38].

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (8)$$

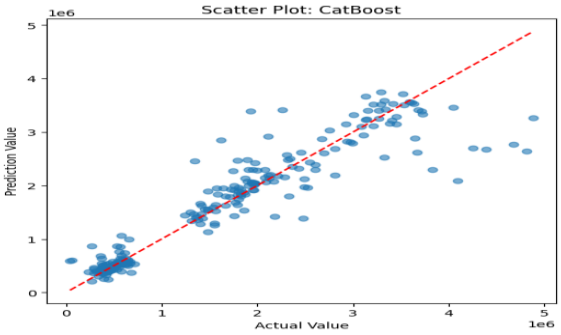
Lower RMSE and MAE values indicate better prediction accuracy, while higher R^2 values indicate greater explanatory power[32]. To ensure robust model evaluation and reduce the risk of overfitting, five-fold cross-validation was applied. The dataset was divided into five subsets, where four subsets were used for training and one subset for testing in each iteration. The final RMSE, MAE, and R^2 values were obtained by averaging the results across all folds.

RESULTS AND DISCUSSION

Result

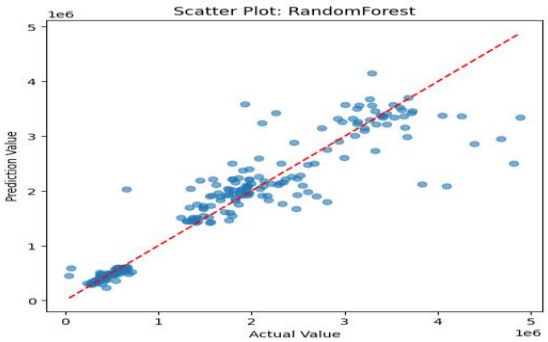
This section presents the computational experiment results and a comparative performance analysis of the four proposed algorithms. The analysis is based on the evaluation metrics defined in the previous section. After hyperparameter tuning, the best configuration for each model was obtained. Figures 2–5 illustrate prediction accuracy using five-fold cross-validation.

Figure 2 presents the five-fold CV results for the CatBoost algorithm. The scatter plot shows a relatively tight distribution of data points around the diagonal line, particularly within the low to medium production ranges. This indicates that the model is able to generate predictions that are close to the actual values with relatively small errors for most observations. With an RMSE of 125.4 and an MAE of 88.2, CatBoost demonstrates strong predictive capability; however, its performance exhibits slightly lower stability compared to the best-performing model.



Source: (Research Result, 2026)

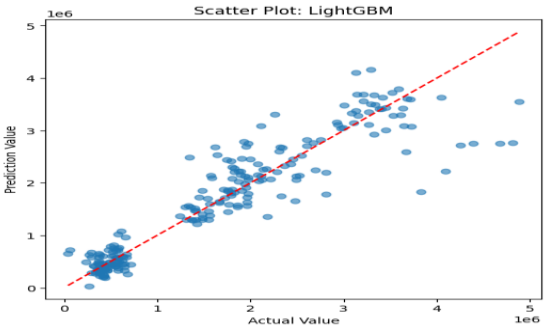
Figure 2. 5-fold CV Algorithm CatBoost



Source: (Research Result, 2026)

Figure 3. 5-fold CV Random Forest Algorithm

Figure 3 illustrates the five-fold CV results for the Random Forest model. Compared to XGBoost and CatBoost, the dispersion of points is noticeably wider, with many predictions deviating further from the diagonal line, especially for medium to high production values. The obtained RMSE of 135.7 and MAE of 95.5 indicate moderate predictive accuracy but reduced stability. This behavior can be attributed to the bagging-based nature of Random Forest, which is effective in reducing variance but may not sufficiently address complex bias present in heterogeneous agroclimatic data.



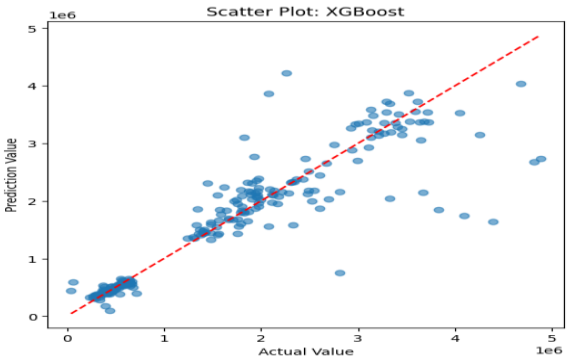
Source: (Research Result, 2026)

Figure 4. 5-fold CV LightGBM Algorithm

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8335

Figure 4 shows the five-fold CV results for the LightGBM algorithm. The scatter plot reveals a more dispersed and asymmetric distribution compared to the other models, with a substantial number of points located above the diagonal line. This pattern indicates a tendency toward overestimation across many samples. Although LightGBM achieves an RMSE of 122.8 and an MAE of 85.1, which are not the worst among the evaluated models, the visual dispersion suggests instability in prediction performance across different production ranges.



Source : (Research Result, 2026)
Figure 5. 5-fold CV XGBoost Algorithm

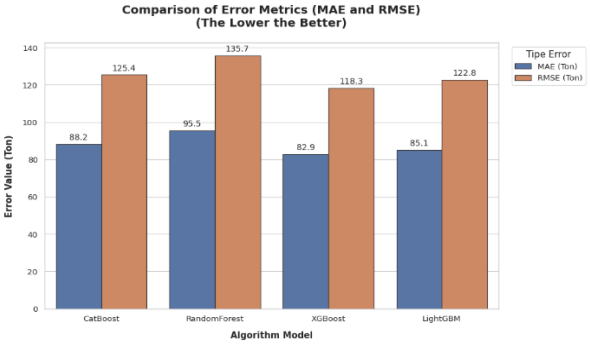
Figure 5 presents the five-fold CV results for XGBoost. Among all evaluated models, XGBoost exhibits the tightest clustering of points along the diagonal line, indicating the closest agreement between predicted and actual values. This visual observation is fully consistent with the quantitative metrics, where XGBoost achieves the lowest RMSE (118.3), the lowest MAE (82.9), and the highest coefficient of determination (R^2 = 0.74). No dominant systematic bias toward overestimation or underestimation is observed, and the number of outliers is minimal, highlighting the robustness and stability of the model.

Table 2. Model Performance Evaluation Results

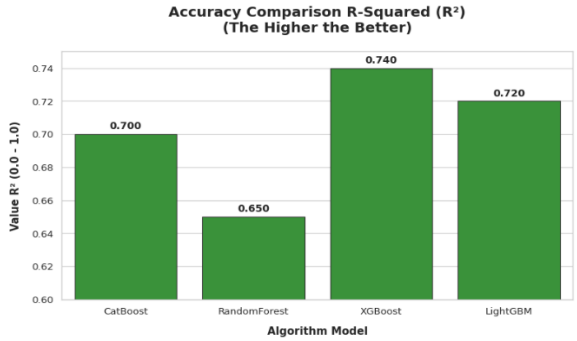
Model	RMSE	MAE	R^2
CatBoost	125,4	88,2	0,70
RandomForest	135,7	95,5	0,65
XGBoost	118,3	82,9	0,74
LightGBM	122,8	85,1	0,72

Source: (Research Result, 2026)

Table 2 summarizes the evaluation results based on RMSE, MAE, and R^2 for all four models. XGBoost outperforms the other algorithms by achieving the lowest error values and the highest explanatory power. LightGBM ranks second, followed by CatBoost, while Random Forest records the highest prediction errors and the lowest R^2 value. These differences confirm that XGBoost is the most capable model in capturing the complex nonlinear relationships between input variables and rice production outcomes.



Source: (Research Result, 2026)
Figure 6. Comparison of Error Metrics (MAE and RMSE)



Source: (Research Result, 2026)
Figure 7. Accuracy Comparison R-Squared (R^2)

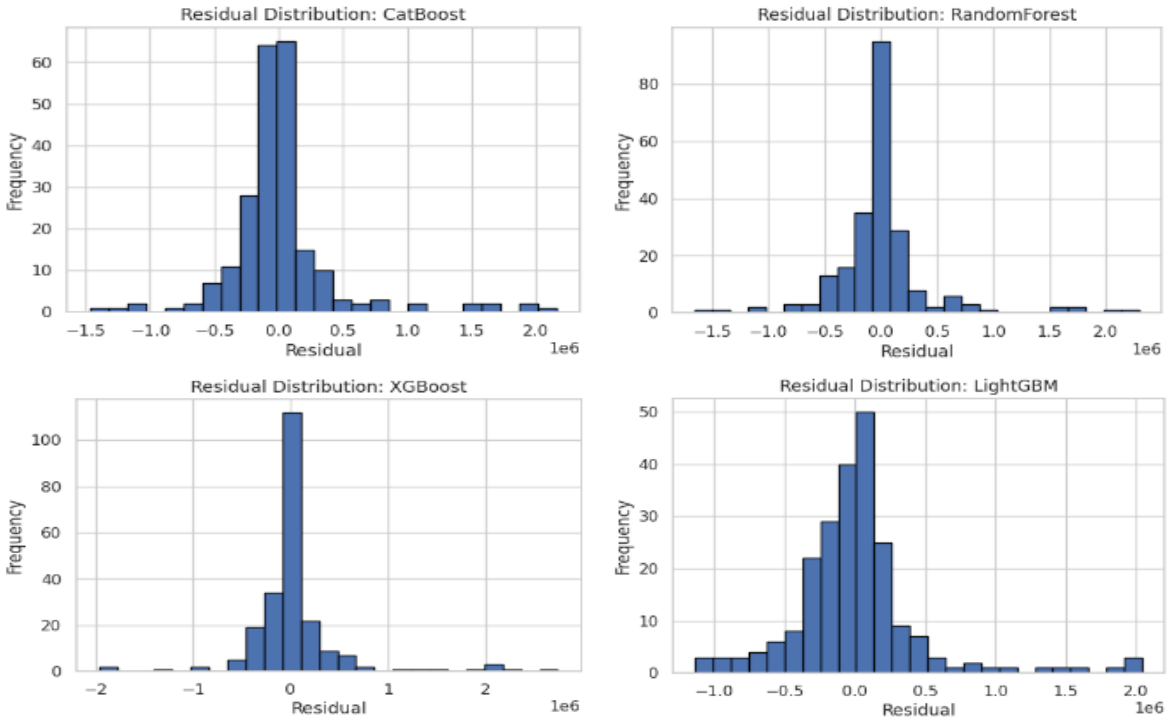
The comparison of error metrics, as illustrated in Figure 6, further emphasizes the superiority of XGBoost in minimizing both RMSE and MAE. Figure 7 shows the R^2 comparison, where XGBoost again achieves the highest value, indicating better overall model fit. Collectively, these results suggest that boosting-based methods, particularly XGBoost, are more effective than bagging-based approaches for rice production prediction.

After obtaining the quantitative evaluation results of the model using RMSE, MAE, and R^2 metrics, the next step in performance analysis is to gain a deeper understanding of the prediction error patterns and the model's internal mechanisms in generating outputs. Evaluation metrics provide a general overview of accuracy and average error levels; however, they do not fully explain how prediction errors are distributed across all data points or how the model makes decisions based on input features. Therefore, two advanced approaches are required: residual distribution analysis (as shown in Figure 6) and SHAP (SHapley Additive exPlanations) analysis (as presented in Figures 7 and 8).

CC BY NC

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8335

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)



Source: (Research Result, 2026)

Figure 8. Residual Distribution of Algorithm Model

Figure 6 presents the residual histograms for each model, revealing distinct characteristics. The CatBoost model exhibits a symmetric residual distribution strongly centered around zero, indicating relatively small and consistent prediction errors. Random Forest also demonstrates a symmetric distribution; however, with a wider spread, suggesting greater variability in prediction errors. XGBoost displays the most ideal characteristics, with the narrowest residual distribution, the highest frequency peak around zero, and very few outliers. This indicates that XGBoost produces highly stable predictions that closely approximate the actual values.

In contrast, LightGBM shows a residual distribution that is positively skewed, suggesting a tendency toward overestimation. The wider dispersion of residual values in LightGBM also indicates a higher level of predictive uncertainty compared to the other three models.

Table 3. Comparison of Residual Distribution Statistics

Model	Distribution Symmetry	Residual Dispersion	Prediction Consistency	Outliers	Description
CatBoost	High (symmetric)	Narrow–moderate	Good	Low	Stable and relatively accurate predictions
RandomForest	Moderate (symmetric)	Moderate–wide	Fairly stable	Moderate	Higher error variability
XGBoost	Very high	Narrowest	Very stable	Very low	Best-performing model (smallest and most consistent residuals)
LightGBM	Low (positively skewed)	Wide	Less stable	High	Tends to overestimate

Source: (Research Result, 2026)

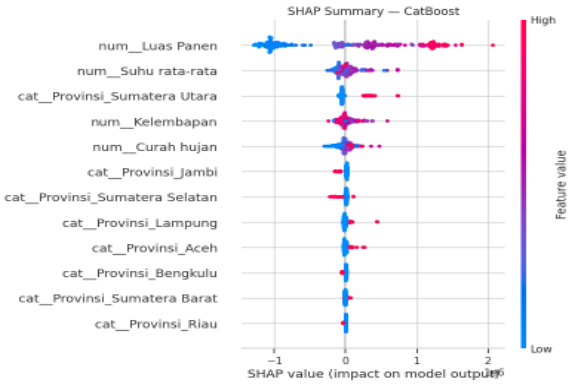
Based on Table 3, the XGBoost model exhibits the most ideal residual distribution, characterized by the narrowest dispersion, near-perfect symmetry, and the highest frequency peak around zero. This indicates that XGBoost produces very small and consistent prediction errors across the majority of data points. CatBoost demonstrates similar performance, although slightly inferior to XGBoost, as its residual spread is marginally wider and includes a few additional outliers. Random Forest shows a symmetric residual pattern but with greater error variance, suggesting that the bagging approach is less effective in optimizing predictions

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8335

on this dataset compared to boosting methods. LightGBM produces the largest residuals and displays a positive skew, indicating a tendency toward overestimation. These results suggest that, in the context of rice production prediction, XGBoost is the most accurate and stable model, followed sequentially by CatBoost, Random Forest, and LightGBM.

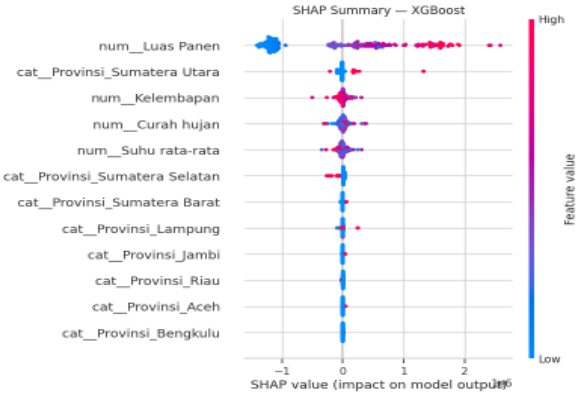
Subsequently, SHAP analysis serves as an essential approach to understanding the contribution of each input feature to the model's predictions. This is particularly relevant for tree-based ensemble algorithms such as CatBoost, Random Forest, XGBoost, and LightGBM, which are inherently considered black-box models. Through SHAP (SHapley Additive exPlanations), each prediction is decomposed into additive feature contributions, enabling researchers to identify which features most strongly influence the increase or decrease of predicted values. Understanding feature contributions not only enhances model transparency and interpretability but also provides important empirical insights into the key factors affecting rice production. The SHAP value results for the two most influential algorithms in improving rice production can be observed in Figures 7 and 8.



Source: (Research Result, 2026)

Figure 9. SHAP Summary CatBoost

Figure 9 illustrates that the feature num_Luas Panen (harvested area) exerts the greatest influence on rice production predictions. The red data points positioned on the right side of the plot with positive SHAP values indicate that a larger harvested area significantly increases the predicted production. Conversely, blue points located on the left side with negative SHAP values suggest that smaller harvested areas reduce the predicted output. The next most influential feature is num_Suhu rata-rata (average temperature). Higher temperature values tend to produce positive SHAP contributions, indicating that an increase in average temperature within the optimal growth range enhances the predicted rice yield.



Source: (Research Result, 2026)

Figure 10. SHAP Summary XGBoost

Figure 10 demonstrates a pattern similar to that of CatBoost, yet with a more pronounced and structured influence. The feature num_Luas Panen once again emerges as the most dominant variable. The SHAP pattern appears more clearly organized: high-value points (red) are strongly concentrated on large positive SHAP values, indicating that XGBoost is more sensitive to variations in harvested area. Weather-related features such as humidity, rainfall, and average temperature display a more symmetric and compact distribution, reflecting the model's stability in capturing agroclimatic variability. Overall, the SHAP distribution in XGBoost appears denser and cleaner, suggesting stronger interpretability and greater robustness against fluctuations in input values.

Overall, the consistency across performance evaluation, residual analysis, and SHAP interpretation demonstrates strong alignment in the findings: XGBoost outperforms the other models comprehensively in terms of accuracy, predictive stability, and feature interpretability. CatBoost serves as a strong alternative with performance closely approaching that of XGBoost, whereas Random Forest and LightGBM exhibit limitations in capturing the complexity of inter-variable relationships within the rice production dataset.

Discussion

The superior performance of XGBoost can be attributed to its gradient boosting mechanism, which effectively captures complex nonlinear relationships among harvested area, climatic variables, and rice production outcomes. Compared with Random Forest, XGBoost iteratively corrects prediction errors from previous trees, enabling better generalization on heterogeneous agricultural datasets. The SHAP analysis further confirms that

CC BY NC

harvested area is the dominant factor affecting rice production. This finding is consistent with agricultural production theory, where production output is directly associated with the cultivated area. Climatic variables such as rainfall, humidity, and temperature also contribute significantly, indicating that environmental conditions remain critical determinants of agricultural productivity.

Nevertheless, several limitations should be acknowledged. First, the dataset contains only 224 observations, which may limit the model's ability to learn more complex spatial and temporal patterns. Second, the study only incorporates a limited number of agroclimatic variables and does not include soil fertility, irrigation infrastructure, fertilizer application, pest incidence, or satellite-derived vegetation indices. Third, the analysis focuses exclusively on Sumatra Island, which may restrict the generalizability of the findings to other regions of Indonesia with different agroecological characteristics. Future research should consider expanding the dataset coverage to all provinces in Indonesia and incorporating additional variables such as remote sensing indicators, soil characteristics, and agricultural management practices. Furthermore, advanced deep learning approaches and hybrid ensemble models may be explored to further improve prediction performance and support more comprehensive food security planning.

CONCLUSION

The findings indicate that the XGBoost model delivers the most superior performance in predicting rice production compared to CatBoost, Random Forest, and LightGBM. The latest evaluation metrics show that XGBoost achieved an RMSE of 118.3, MAE of 82.9, and R^2 of 0.74, consistently outperforming CatBoost (RMSE 125.4; MAE 88.2; R^2 0.70), LightGBM (RMSE 122.8; MAE 85.1; R^2 0.72), and Random Forest (RMSE 135.7; MAE 95.5; R^2 0.65). These results are further supported by a narrower residual distribution centered closer to zero, indicating greater predictive stability across various value ranges. SHAP analysis provides clear interpretative insight, revealing that harvested area is the most influential variable, while climatic factors and provincial characteristics also contribute significantly to variations in production values.

DECLARATION OF AI USE

The authors acknowledge the use of ChatGPT (OpenAI) in assisting with academic writing, English language editing, grammatical refinement, and research gap formulation. All AI-assisted content was critically reviewed and verified by the authors. The authors assume full responsibility for the accuracy, originality, integrity, and final content of the manuscript, including the validity of all references and research findings.

AUTHOR CONTRIBUTIONS

This journal uses the Contributor Roles Taxonomy (CRediT) to identify individual author contributions to the research and manuscript preparation.

Author	Contribution Roles
Dahri Yani Hakim Tanjung (Corresponding Author)	Research design, model development, software implementation, data analysis, visualization, and preparation of the original manuscript.
Firman Syahputra	Data collection and management, technical implementation, model testing, and result validation
Rofiqoh Dewi	Methodological refinement, result analysis, validation, technical supervision, and manuscript revision
Hilda Elsera Br Sembiring	Conceptual development, investigation, resource support, research coordination, and manuscript revision

ACKNOWLEDGMENTS

The authors gratefully acknowledge Universitas Satya Terra Bhinneka and the Institute for Research and Community Service (LPPM) of Universitas Satya Terra Bhinneka for their valuable support and institutional contributions, which facilitated the successful completion of this research and its publication.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest related to the research, writing, or publication of this article.

REFERENCE

[1] Badan Pusat Statistik, "Luas Panen dan Produksi Padi di Indonesia 2023 (Angka Tetap)," 2024.
[2] Food and Agriculture Organization (FAO), "Food Outlook," 2022.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

VOL. 12. NO. 2 NOVEMBER 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i2.8335

[3] A. A. Ikudayisi, "Green Technologies and Sustainability Urban food security and socioeconomic sustainability : A multidimensional perspective," *Green Technol. Sustain.*, vol. 2, no. 2, p. 100080, 2024, doi: 10.1016/j.grets.2024.100080.

[4] S. de Bruin and C. Holleman, *Urbanization is transforming agrifood systems across the rural–urban continuum creating challenges and opportunities to access affordable healthy diets*. Rome: Food and Agriculture Organization of the United Nations, 2023. doi: 10.4060/cc8094en.

[5] E. Demartini, D. Vecchiato, L. Finos, S. Mattavelli, and A. Gaviglio, "Would you buy vegan meatballs ? The policy issues around vegan and meat-sounding labelling of plant-based meat alternatives," *Food Policy*, vol. 111, 2022, doi: 10.1016/j.foodpol.2022.102310.

[6] Y. Fan, Y. Yuan, Y. Yuan, and W. Duan, "Research progress on the impact of climate change on wheat production in China," *PeerJ*, pp. 1–29, 2025, doi: 10.7717/peerj.18569.

[7] M. Firdaus, F. Hadi, and L. M. Sabri, "Prediction of Rice Productivity Using the Random Forest Regression Algorithm in Cikaret Subdistrict for the Years 2020-2024," *J. Geod. Geomatics*, vol. 20, no. 1, pp. 10–21, 2025, doi: 10.12962/geoid.v20i1.2570.

[8] N. B. Wibisono and S. Saiful, "Crop Yield Prediction Using Random Forest Algorithm and Xgboost Machine Learning Model," *Int. J. Res. Innov. Soc. Sci.*, vol. IX, no. 2454, pp. 1983–1994, 2025, doi: 10.47772/IJRISS.

[9] N. Iqbal, M. U. Shahzad, E. M. Sherif, and M. U. Tariq, "Analysis of Wheat-Yield Prediction Using Machine Learning Models under Climate Change Scenarios," *Sustainability*, pp. 1–26, 2024, doi: https://doi.org/10.3390/su16166976.

[10] M. F. Aslan and K. Sabanci, "Artificial Intelligence Techniques in Crop Yield Estimation Based on Sentinel-2 Data : A Comprehensive Survey," *Sustainability*, vol. 16, 2024, doi: https://www.mdpi.com/2071-1050/16/18/8277.

[11] Q. Zhao *et al.*, "Analysis of Temporal and Spatial Variations in Cropland Water-Use Efficiency and Influencing Factors in Xinjiang Based on the XGBoost – SHAP Model," *Agronomy*, pp. 1–19, 2025, doi: https://doi.org/10.3390/agronomy15081902.

[12] M. Hasan, A. Marjan, and P. Uddin, "Ensemble machine learning- based recommendation system for effective prediction of suitable agricultural crop cultivation," *Front. Plant Sci.*, no. August, pp. 1–18, 2023, doi: 10.3389/fpls.2023.1234555.

[13] N. Jalal, A. Mehmood, G. Sang, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 2733–2742, 2022, doi: 10.1016/j.jksuci.2022.03.012.

[14] S. Ranjan, "Crop Yield Prediction Using the XG-Boost Algorithm in Indian Agroclimatic Conditions," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 14, no. VI, 2026, doi: 10.22214/ijraset.2026.83744.

[15] S. Kumar, M. A. Sohail, S. Jadhav, and R. K. Gupta, "Light Gradient Boosting Machine For Optimizing Crop Maintenance And Yield Prediction In Agriculture," *ICTACT J. SOFT Comput.*, vol. 15, no. 02, pp. 3551–3555, 2024, doi: 10.21917/ijsc.2024.0495.

[16] B. Dong, S. Ma, Z. Gao, and A. Qin, "CatBoost Improves Inversion Accuracy of Plant Water Status in Winter Wheat Using Ratio Vegetation Index," *Appl. Scinces*, vol. 15, pp. 1–20, 2025, doi: https://doi.org/10.3390/app152111363.

[17] M. Mathivadhani, K. T. M. Sc, M. Phil, and D. Ph, "Crop Yield Prediction Using Dep XG Boost Algorithm," *Int. J. Innov. Res. Electr. Electron. Instrum. Control Eng.*, vol. 13, no. 4, pp. 51–56, 2025, doi: 10.17148/IJIREEICE.2025.13408.

[18] Y. Ye *et al.*, "Diverse microbial communities support multiple patterns of functional redundancy in tidal flats IN Corresponding author," *Microbiome*, 2026, doi: /10.1186/s40168-026-02503-9.

[19] S. Sarkar, J. M. O. Leyton, E. Noa-yarasca, K. Adhikari, C. B. Hajda, and D. R. Smith, "Integrating Remote Sensing and Soil Features for Enhanced Machine Learning-Based Corn Yield Prediction in the Southern US," *Sensors*, vol. 25, no. 2, pp. 1–27, 2025, doi: 10.3390/s25020543.

[20] B. Avanzi, G. Taylor, and M. Wang, "Machine Learning with High-Cardinality Categorical Features in Actuarial Applications," *ASTIN Bull. J. IAA*, vol. 54, no. 2, pp. 213–238, 2024, doi: 10.1017/asb.2024.7.

[21] V. A. Handayani, S. Rahmiati, B. Varischa, and A. Arrafi, "Comparative Analysis : Multiple Regression and Random Forest Regression in Predicting Food Security Index in Indonesia," *J. Math. Comput. Stat.*, vol. 8, no. 2, pp. 389–400, 2025, doi:

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.

10.35580/jmathcos.v8i2.8593.

[22] M. Abdel-salam, "A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning," *Neural Comput. Appl.*, vol. 36, no. 33, pp. 20723–20750, 2024, doi: 10.1007/s00521-024-10226-x.

[23] S. Morshed, F. Barua, A. Khanom, F. Lokman, and H. T. Zubair, "Smart Agricultural Technology Crop yield prediction using machine learning : An extensive and systematic literature review," *Smart Agric. Technol.*, vol. 10, no. September 2024, p. 100718, 2025, doi: 10.1016/j.atech.2024.100718.

[24] M. Kulyal and P. Saxena, "Machine Learning approaches for Crop Yield Prediction : A Review," *2022 7th Int. Conf. Comput. Commun. Secur.*, 2022, doi: 10.1109/ICCCS55188.2022.10079240.

[25] A. Oikonomidis, C. Catal, and A. Kassahun, "Hybrid Deep Learning-based Models for Crop Yield Prediction Hybrid Deep Learning-based Models for Crop Yield Prediction," *Appl. Artif. Intell.*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2031823.

[26] P. Giri *et al.*, "An Ensemble-Learning Framework for Structural Damage Classification in Prestressed Bridge Deck," *Procedia Struct. Integr.*, vol. 84, pp. 1119–1126, 2026, doi: 10.1016/j.prostr.2026.06.143.

[27] G. S. P. Ghantasala, "Ensemble Deep Learning With Radiomics Features," *Genet. Mol. Res.*, vol. 25, no. 4s, pp. 1–9, 2026, doi: 10.4238/gaxeq208.

[28] Y. Li *et al.*, "Measuring neighborhood socioeconomic status from sky and street : An ensemble learning framework," *Trans. Urban Data, Sci. Technol.*, pp. 1–29, 2026, doi: 10.1177/27541231261473758.

[29] C. Oktomy, N. Susanto, M. Ulfah, and B. Aditya, "Comparative analysis of CatBoost and LightGBM algorithms for predicting dental caries risk based on lifestyle factors," *J. Ilm. KURSOR*, vol. 13, no. 3, pp. 155–166, 2026, doi: 10.21107/kursor.v13i3.504.

[30] Z. Wei *et al.*, "CatBoost-based Intrusion Detection Method for the Physical Layer of Smart Agriculture," *ITM Web Conf.*, vol. 60, p. 8, 2024, doi: 10.1051/itmconf/20246000009.

[31] K. Ileri, "Comparative analysis of CatBoost , LightGBM , XGBoost , RF , and DT methods optimised with PSO to estimate the number of k - barriers for intrusion detection in wireless sensor networks," *Int. J. Mach. Learn. Cybern.*, vol. 16, no. 9, pp. 6937–6956, 2025, doi: 10.1007/s13042-025-02654-5.

[32] T. Ait, B. Elkari, Y. Chaibi, and T. Kousksou, "Random forest with feature selection and K-fold cross validation for predicting the electrical and thermal efficiencies of air based photovoltaic-thermal systems," *Energy Reports, Publ. by Elsevier Ltd*, vol. 12, no. March, pp. 988–999, 2024, doi: 10.1016/j.egyr.2024.07.002.

[33] G. Alexandar Narkunam, K. Kala, and S. ArunPandiyan, "Enhancing Agricultural Forecasting with an Ensemble Learning Approach for Broccoli Yield Prediction," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 41s, pp. 105–116, 2025, doi: 10.52783/jisem.v10i41s.7754.

[34] D. R. Anamisa, B. D. Satoto, M. K. Sophan, and M. Yusuf, "Mathematical Modelling of Engineering Problems Development of the Support Vector Regression and Genetic Algorithm Hybrid Models for Forecasting of Madura Rice Yield," *Int. Inf. and Eng. Technol. Assoc.*, vol. 13, no. 1, pp. 181–198, 2026, doi: 10.18280/mmep.130117.

[35] H. Wijaya, D. P. Hostiadi, and E. Triandini, "OPTIMIZATION OF XGBOOST ALGORITHM USING PARAMETER TUNING IN RETAIL SALES PREDICTION," *J. Nas. Pendidik. Tek. Inform. JANAPATI*, vol. 13, no. 3, pp. 769–786, 2024, doi: 10.23887/janapati.v13i3.82214.

[36] D. Malarvizhi, "A comparative study of boosting algorithms : Concepts , Algorithms , Applications , and Prospects," *Int. J. Innov. Res. Technol.*, vol. 11, no. 3, pp. 1816–1825, 2024.

[37] S. Kaur, G. Singh, A. Kumar, and A. Singh, " Improvised wheat yield forecasting via SHAP-based feature Selection and Bayesian optimized weighted ensemble model," *Environ. Res. Commun.*, vol. 7, 2025, doi: https://doi.org/10.1088/2515-7620/ae244c.

[38] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci. Model Dev.*, vol. 15, no. 2, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.

Accredited Rank 2 (SINTA 2) based on Decree No. 225/E/KPT/2022 issued by Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, dated December 7, 2022. Published by LPPM Universitas Nusa Mandiri.