IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 September 2024, accepted 8 October 2024, date of publication 11 October 2024, date of current version 24 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3478738

RESEARCH ARTICLE

Enhancing Rice Production Prediction in Indonesia Using Advanced Machine Learning Models

ERLIN[1], ARDA YUNIANTA[2], LILI AYU WULANDHARI[3], YENNY DESNELITA[4], NURLIANA NASUTION[5], AND JUNADHI[6]

[1]Department of Informatics Engineering, Faculty of Computer Science, Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru 28127, Indonesia
[2]Department of Information Systems, Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[3]Department of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
[4]Department of Information Systems, Faculty of Computer Science, Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru 28127, Indonesia
[5]Department of Informatics Engineering, Faculty of Computer Science, Universitas Lancang Kuning, Pekanbaru 28265, Indonesia
[6]Department of Informatics Engineering, Universitas Sains dan Teknologi Indonesia, Pekanbaru 28294, Indonesia

Corresponding author: Erlin (erlin@lecturer.pelitaindonesia.ac.id)

This work was supported by the Ministry of Education, Culture, Research, and Technology of Indonesia under Contract 112/E5/PG.02.00.PL/2024; and by Institut Bisnis dan Teknologi Pelita Indonesia.

ABSTRACT This study delves into the application of machine learning techniques for predicting rice production in Indonesia, a country where rice is not just a staple food but also a key component of the agricultural sector. Utilizing data from 2018 to 2023, sourced from the Central Bureau of Statistics of Indonesia and the Meteorology, Climatology, and Geophysics Agency of Indonesia, this research presents a comprehensive approach to agricultural forecasting. The study begins with an Exploratory Data Analysis (EDA) to understand the variability and distribution of variables such as harvested area, production, rainfall, humidity, and temperature. Significant regional disparities in rice production are identified, highlighting the complexity of agricultural forecasting in Indonesia. Five machine learning models— Random Forest, Gradient Boosting, Decision Tree, Support Vector Machine, and Artificial Neural Network—are trained and tested. The Random Forest model stands out for its superior performance, as evidenced by the lowest Mean Squared Error (MSE) of 0.016186 and the highest R-squared (R^2) of 0.850039, compared to the other models, indicating its high predictive accuracy and reliability. Hyperparameter tuning using the GridSearchCV technique was conducted on all five models, resulting in performance improvements across the board. Despite these enhancements, the Random Forest model remained the best, achieving a lower MSE of 0.014162 and a higher R^2 value of 0.911257. This research not only underscores the effectiveness of machine learning in improving rice production predictions in Indonesia but also sets the stage for future research. It highlights the potential of advanced analytical techniques in enhancing agricultural productivity and decision-making, paving the way for further explorations into more sophisticated models and a broader range of data, ultimately contributing to the resilience and sustainability of Indonesia's agricultural sector.

INDEX TERMS Agroclimatic variability, crop yield production, data analysis, Indonesia rice production, machine learning.

I. INTRODUCTION

Indonesia is a country well-known for its agricultural sector, with rice being a keystone of this landscape. Rice is not only a staple food but also a cultural and economic symbol. It serves as a crucial source of livelihood for millions of farmers across the country. The importance of rice in Indonesia is immense, playing a central role in the daily lives and diet of over 270 million people. Indonesians, on average, consume a significant amount of rice every day. In terms of

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

consumption, Indonesia ranks fourth behind China, India, and Bangladesh, consuming approximately 35.2 million metric tons in 2022/2023, according to Statista [1]. As a producer, it ranks similarly, being the fourth largest with a production of 31.54 million metric tons in 2022, as reported by USDA [2].

Predicting rice production in Indonesia is important because it faces several challenges. These include unpredictable weather patterns and climate change, limited arable land, varying humidity levels, and the need to balance production with sustainability [3]. Accurate predictions can help farmers plan their activities and reduce economic risks [4]. Furthermore, reliable predictions assist the government and stakeholders in making timely interventions to control price fluctuations, which are vital for the welfare of farmers and consumers [5]. Additionally, precise predictions of rice production are crucial in ensuring sufficient rice availability, which is essential for national food security, preventing hunger and malnutrition, and maintaining social stability [6].

Machine learning, a subset of artificial intelligence, utilizes algorithms and statistical models to improve computer systems' performance through learning from data. In agriculture, it has become increasingly prominent, enhancing productivity, sustainability, and decision-making. Machine learning can identify correlations between climate variables and harvested area, supporting climate adaptation strategies and driving sustainable agricultural practices [7].

Although significant progress has been made in applying machine learning to agricultural forecasting, to the best of the author's knowledge, there has yet to be any research that comprehensively discusses rice production across all provinces in Indonesia. Specifically, there is a lack of studies focusing on implementing and developing machine learning models tailored to the diverse agroclimatology of Indonesia.

Building upon previous research, which often focused on a singular aspect of agricultural forecasting such as yield prediction based solely on historical production data [8], [9], [10], this study adopts a more holistic approach. We integrate multiple key environmental factors-specifically harvested area, rainfall, humidity, and temperature to predict rice production in Indonesia. This multifaceted model not only considers the direct impact of climatic conditions on crop yields but also acknowledges the intricate relationship these factors have with one another in influencing agricultural output. Furthermore, our research expands its scope beyond the commonly studied regions, encompassing a comprehensive analysis across all provinces in Indonesia. This approach allows for a more accurate and region-specific understanding of rice production, considering the diverse agroclimatic conditions across the country. Additionally, we employ a range of advanced machine learning techniques, including Random Forests, Gradient Boosting, Decision Tree, Support Vector Machines, and Artificial Neural Network, to determine which algorithms are most effective in the Indonesian context. This methodological diversity, along with our extensive geographic and environmental coverage, contributes uniquely to this research, guiding us toward providing more nuanced, accurate, and regionally tailored predictions for rice production in Indonesia.

This study aims to develop and implement advanced machine learning models specifically designed to predict rice production in Indonesia, with a focus on analyzing the complex interplay between harvested area, rainfall, humidity, and temperature. Our objective is to accurately forecast rice yields, leveraging these key climatic and environmental variables, to enhance decision-making for farmers and policymakers, and contribute to the sustainability and efficiency of the agricultural sector.

The paper is organized as follows: Section II presents a comprehensive review of the relevant literature, providing the historical context of previous research. Section III describes the research methods used to predict rice production. Section IV presents the results and discussion. Section V concludes the paper with final remarks and outlines directions for future research.

II. RELATED WORK

Machine learning (ML) has become an essential tool in modern agriculture, significantly improving the accuracy of crop yield predictions and overall agricultural outcomes [11], [12], [13]. Its application in rice production has been particularly transformative, enhancing yield forecasts, optimizing resource utilization, and improving decision-making for farmers and policymakers [14], [15].

A wide range of studies have explored various machine learning algorithms for estimating rice production. Decision tree algorithms, for instance, have been used to forecast yields based on data such as annual crop yield, annual rainfall, area of production, season, and the specific state in India [16]. Their simplicity and user-friendliness make them particularly suitable for agricultural applications. Moreover, the adoption of ensemble learning techniques, like random forests, has been noteworthy for their ability to handle complex, multi-dimensional data, yielding robust and reliable predictions [17].

Support Vector Machines (SVMs) have been primarily implemented for classifying rice production outcomes, employing various kernels [18]. Their ability to handle non-linear data makes them versatile in diverse rice production scenarios. Additionally, deep learning techniques, including artificial neural network, have been utilized to discern correlations between climatic factors and paddy yield, demonstrating their potential to predict future yields more effectively and efficiently [19].

One area of particular interest among researchers is data sources and feature engineering. The extensive use of climate data, such as temperature, rainfall, and humidity, has played a crucial role in modeling the influence of weather conditions on rice yields [20], [21]. This data encompasses both historical weather records and real-time climate information from sensors [22]. Furthermore, the integration of soil characteristics, including pH levels,

nutrient content, and texture, has significantly increased the accuracy of predictive models, highlighting the critical role of soil data in understanding nutrient availability and crop suitability [23].

In addressing regional specificity and crop varieties, researchers have emphasized the importance of accounting for regional variations in rice production. This involves adapting models to different rice varieties, agronomic practices, and climatic conditions, enabling region-specific predictions [24]. Customized machine learning models have been created for various rice types, including hybrid and traditional varieties, each with unique growth patterns and requirements.

Globally, researchers have been predicting rice production using various methods and techniques. A study by [25] focused on developing and evaluating a Machine Learning system for predicting rice yield in Thailand, analyzing historical data from Thai government departments and various ML models including Generalized Linear Model, Feed-Forward Neural Network, Support Vector Machine, and Random Forest. They found the Feed-Forward Neural Network particularly effective for complex data, despite longer training times. Another study estimated rice yield using MODIS 250 m data and compared the performance of three machine-learning regression algorithms: multiple linear regression, SVM, and Random Forest, with the latter showing the most objective results and good statistical correlations [26].

In India, a machine learning model developed for rice yield prediction was trained on unseen data, testing various algorithms including decision tree, random forest, support vector machine, and convolutional neural network. Linear regression performed best in achieving R^2, RMSE, and MAE scores [27]. In China, Zhang et al., [28], conducted a study on mapping the spatial distribution of double-season rice utilizing Sentinel-1 SAR time series data, achieving high classification accuracy. Another study focused on creating an accurate crop yield forecasting system for wheat, maize, and rice using environmental predictors and the Random Forest model, which demonstrated strong performance [29].

Tanaka et al., [30] present a deep-learning approach using RGB images to estimate rice yield, capturing over 22,000 images across six countries in Africa and Japan. Using a convolutional neural network, they effectively predicted yield variation, showing potential for unmanned aerial vehicle scaling. In Vietnam, Hoang-Phi et al., [31] developed a tool for estimating rice yield in the Mekong Delta using Synthetic Aperture Radar (SAR) data from the Sentinel-1 satellites, aligning closely with in-situ yield measurements.

In Indonesia, machine learning applications in agriculture are an expanding field. While existing research by [32], [33] has advanced this area, it often covers limited geographical regions and does not fully consider a broad range of environmental factors. Our research expands on these frameworks, adopting a holistic strategy by integrating key environmental factors and applying our methods to all provinces in Indonesia. Through advanced machine learning algorithms, we aim to provide a more detailed and precise understanding of rice production prediction.

III. RESEARCH METHODS

The methodology adopted for this research is illustrated in the flow diagram presented in Figure 1. The initial phase is data understanding, aiming to gain a deep understanding of the dataset that will be utilized to train the model. The next step is Exploratory Data Analysis (EDA), which is divided into three parts: descriptive statistical analysis, correlation analysis, and data visualization. Following this, the dataset undergoes a pre-processing phase that encompasses data cleaning, data transformation and encoding, normalization and standardization, as well as data splitting. The next phase is model training, during which a variety of machine learning algorithms are evaluated, such as Random Forest, Gradient Boosting, Decision Tree, Support Vector Machines, and Artificial Neural Network.

Random Forest is chosen for its robustness and ability to reduce overfitting by averaging multiple decision trees, supported by its feature importance scores and ability to handle high-dimensional datasets. Gradient Boosting is selected for its high predictive accuracy and ability to handle complex data patterns through the stage-wise addition of weak learners, justified by its performance in minimizing the loss function and reducing bias and variance. Decision Tree is utilized for its simplicity and interpretability, making it a good baseline model, justified by its ease of visualization and understanding of data splits. Support Vector Machine is preferred for its effectiveness in high-dimensional spaces and ability to handle linear and non-linear data with kernel functions, supported by its margin maximization approach and flexibility in tuning parameters. Artificial Neural Network is chosen for its flexibility and capability to model complex, non-linear relationships in the data, justified by its structure inspired by the human brain and the effectiveness of backpropagation and gradient descent in training.

After the training phase, model evaluation is conducted using Mean Squared Error (MSE) and R-squared (R^2) metrics to assess the performance of the model. The procedure advances with Hyperparameter Tuning, where the optimal set of hyperparameters is selected for the model that demonstrates the highest precision, following the prior evaluation stages. The final step is to re-evaluate the selected model to compare the outcomes before and after the hyperparameter selection.

A. DATA UNDERSTANDING

Data Understanding is the initial step in gaining a comprehensive view of the dataset. The dataset comprises seven variables, including one dependent variable (target), which is rice production, and six independent variables: year, province, harvested area, rainfall, humidity, and temperature. One of the variables is categorical (province), while others are numerical. Understanding this data enables researchers to

make informed decisions on how to process it and select the most suitable algorithms.

B. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) plays a crucial role in the process of model development. EDA is the stage of data exploration that aids in understanding the complexity, structure, and relationships present within a dataset. This dataset comprises 204 entries (rows) and 7 columns. At this stage, a descriptive statistical analysis is conducted on all variables, encompassing the analysis of minimum, maximum, median, and mean values. During the EDA phase, an analysis is performed to observe trends in rice production, whether increasing or decreasing over time. It also involves examining the impact of factors such as harvested area, rainfall, humidity, and temperature on rice production. Furthermore, EDA includes the analysis of consistent patterns or relationships among variables within the dataset.

C. PRE-PROCESSING

In the pre-processing stage, several critical steps are undertaken to prepare the data for use in developing rice production prediction models. These steps include cleaning the data by imputing missing values, addressing and rectifying issues with data format mismatches, and encoding categorical variables. Given the observed skewness in the harvested area and production features, a transformation is applied to normalize their distribution. This is achieved using the MinMax Scaler on the numerical data, ensuring all variables were on the same scale. The final step in pre-processing is to divide the dataset into a training set and a test set. The training set is used to train the machine learning model, while the test set is used to evaluate the model's performance.

D. MODEL TRAINING AND TESTING

This stage is dedicated to constructing and optimizing a model capable of making predictions based on data. It involves selecting an algorithm or model architecture that is appropriate for the nature of the research problem (regression) and the type of data at hand. The models used in this study include Random Forest, Gradient Boosting, Decision Tree, Support Vector Machine, and Artificial Neural Network. Each of these models has distinct characteristics and advantages.

1) RANDOM FOREST (RF)

RF is an ensemble method that constructs multiple decision trees during training and outputs the average prediction of the individual trees to improve the model's accuracy and control overfitting. It enhances robustness by reducing the variance of the model through averaging. Random Forest can handle large datasets with higher dimensionality and is less prone to overfitting compared to individual decision trees [34]. It also provides feature importance scores, which can be useful for understanding the influence of different features on the prediction. Equation (1) represents the prediction formula for a Random Forest model. This equation indicates that the final prediction $\hat{f}(x)$ is the mean of the predictions from all B decision trees $T_b(x)$ for a given input (x). Each tree in the forest is trained on a random subset of the data with replacement (bootstrap sample), and a random subset of features is considered for splitting at each node. This process helps in reducing the variance of the model, leading to more robust and accurate predictions.

$$\hat{f}(x) = \frac{1}{B}\sum_{b-1}^{B} T_b(x) \tag{1}$$

where:

$\hat{f}(x)$ is the predicted value obtained by averaging the predictions of all decision trees in the forest.

$T_b(x)$ is the prediction of the b-th decision tree

B is the total number of trees

2) GRADIENT BOOSTING (GB)

GB is another ensemble technique that builds the model in a stage-wise fashion by sequentially adding predictors to an ensemble. Each new predictor corrects the errors made by the previous ones, which focuses on minimizing the loss function. This iterative process helps in reducing both bias and variance, making it particularly effective for handling complex data patterns [35]. Gradient Boosting is known for its high predictive accuracy, but it requires careful tuning of parameters such as the learning rate and the number of trees to avoid overfitting. Equation (2) encapsulates the essence of Gradient Boosting, where each subsequent model incrementally improves the overall predictive performance by addressing the shortcomings of the previous models. Gradient Boosting builds models sequentially to minimize the loss function. Each new model $h_m(x)$ is trained to correct the errors (residuals) of the previous model $F_{m-1}(x)$. The contribution of the new model is scaled by the learning rate γ_m, controlling how much of the new model's prediction is added to the current model.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2}$$

where:

$F_m(x)$ is the current model

$F_{m-1}(x)$ is the previous model

γ_m is the learning rate

$h_m(x)$ is the base learner

3) DECISION TREE (DT)

DT models are simple and interpretable, which makes them easy to visualize and understand. They work by splitting the data into subsets based on the most significant feature at each node, creating a tree-like structure of decisions. While decision trees can handle both numerical and categorical data, they are prone to overfitting, especially with noisy data. Pruning techniques are often employed to mitigate overfitting by removing branches that have little importance. Equation (3) represents the prediction formula for a simple

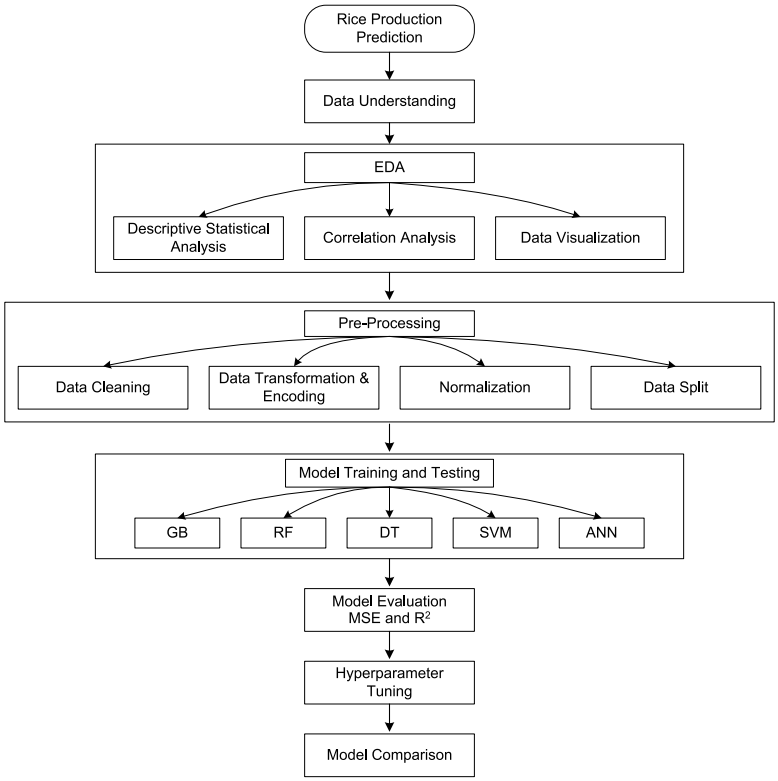

FIGURE 1. Proposed research design.

averaging model, commonly used in the context of a Decision Tree for regression at a leaf node. In the context of a Decision Tree used for regression, the tree splits the data into different regions. Each leaf contains a subset of the training data, and the prediction for any input falling into that leaf is typically the mean of the target values of all the training samples in that leaf. This equation indicates that the predicted value $\hat{f}(x)$ is the mean (average) of the target values y_i of all N observations in the leaf node where the input x falls [36].

$$\hat{f}(x) = \frac{1}{N} \sum_{i-1}^{N} y_i \tag{3}$$

where:

$\hat{f}(x)$ is the predicted value at a leaf node, which is typically the mean of the target values of the training samples that fall into that node

N is the number of observations in the leaf node

4) SUPPORT VECTOR MACHINE (SVM)

SVM is a powerful model particularly well-suited for high-dimensional spaces. For regression tasks, SVM aims to find a hyperplane that best fits the data points within a certain margin of tolerance [37]. SVM is effective for datasets with clear margin separation and can be adapted to nonlinear data using kernel functions (e.g., polynomial, radial basis function). However, SVMs can be computationally intensive and require careful tuning of parameters such as the regularization parameter and kernel type. The prediction formula used in SVM is shown in Equation (4). This equation describes a linear relationship between the input features and the predicted output. It can be used for both regression and classification tasks. The goal is to find the weight vector w and bias term b that minimize the difference between the predicted values and the actual target values (usually by minimizing the mean squared error).

$$\hat{f}(x) = w.x + b \tag{4}$$

where:

$\hat{f}(x)$ is the decision function that predicts the output based on the optimal hyperplane

w is the weight vector

b is the bias

5) ARTIFICIAL NEURAL NETWORK (ANN)

ANN is inspired by the structure of the human brain, consisting of interconnected neurons organized in layers. ANN is highly flexible and capable of capturing complex, non-linear relationships in the data. They are particularly useful for large datasets and can be used for a variety of tasks, including regression. Training ANN involves optimizing weights through backpropagation and gradient descent, which requires significant computational resources. Additionally, ANN requires careful tuning of hyperparameters like the number of layers, neurons per layer, learning rate, and activation functions to achieve optimal performance. Equation (5) captures the essence of how a single neuron in an artificial neural network processes its input. The weights

w_i and bias b are learned during the training process, and the activation function f allows the network to model complex, non-linear relationships in the data. This neuron forms the basic building block of more complex neural network architectures [38].

$$\hat{f}(x) = f(\sum_{i-1}^{n} w_i x_i + b) \tag{5}$$

where:
$\hat{f}(x)$ is the predicted output of the neuron for an input x
f is the activation function
w_i are the weights
x_i are the inputs
b is the bias

Before the commencement of training, a portion of the data is reserved as a test set. This data is not utilized during the training process, ensuring that the model has no prior knowledge of it. The objective is to ascertain that the model possesses robust generalization capabilities and does not merely adapt to the training data, thereby avoiding overfitting.

E. MODEL EVALUATION

Model evaluation is a critical process in the development of machine learning models, involving the measurement and analysis of model performance. This stage occurs after training and testing the model and before its deployment. For this research, the evaluation employs metrics such as Mean Squared Error (MSE) and R-squared (R^2). The Mean Squared Error (MSE) is a widely used metric in regression problems, measuring the average of the squares of the errors. The formula for MSE is presented in Equation (6).

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

where:
$\boldsymbol{y_i}$ is the actual value of the ith sample
$\hat{y}_i$ is the predicted value of the ith sample
$\boldsymbol{n}$ is the number of samples in the dataset

In this equation, $(\boldsymbol{y_i} - \hat{y}_i)^2$ represents the square of the prediction error for each sample, and the Mean Squared Error (MSE) is calculated by averaging all these squared errors. MSE serves as an indicator of how closely the model's predictions align with the actual values. Lower MSE values signify smaller errors, thus indicating a more accurate model.

R-squared (R^2), also known as the coefficient of determination, is a metric used to assess how well a regression model approximates the actual data. The expression for R-squared can be found in Equation (7).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

where:
$\boldsymbol{y_i}$ is the actual value of the ith sample
$\hat{y}_i$ is the predicted value of the ith sample
$\bar{y}$ is the average (mean) of all values in the dataset
$\boldsymbol{n}$ is the number of samples in the dataset

F. HYPERPARAMETER TUNING

Hyperparameter tuning is an important process in the development of machine learning models, aimed at identifying the optimal combination of hyperparameters for the model. These parameter configurations are set before the learning process commences and are not derived from the data itself. Successfully finding the ideal combination of hyperparameters can markedly enhance the model's accuracy and efficacy in making predictions. In this research, the hyperparameter optimization technique employed is GridSearch because it explores each pre-defined parameter combination, thus enabling the discovery of the best parameter combination within a well-defined parameter space.

G. MODEL COMPARISON

In the final stage of the rice production prediction process, model comparison is a critical step to ensure the selection of the most accurate and efficient predictive model. The evaluation metrics used for comparison are Mean Squared Error (MSE) and R-squared (R^2), which provide insights into the accuracy and predictive power of each model. The best model is then chosen for predicting rice production, ensuring that the prediction process is as accurate and reliable as possible. This rigorous comparison ensures that the final model deployed is the most suitable for achieving the predictive goals.

IV. RESULTS AND DISCUSSION

A. DATASET

In this study, data from the Central Bureau of Statistics of Indonesia (BPS) [39] and the Meteorology, Climatology, and Geophysics Agency of Indonesia (BMKG) [40] were combined to create a comprehensive dataset for analysis. The BPS dataset included harvested area and rice production data, while the BMKG dataset provided climatic data such as rainfall, humidity, and temperature. These data were collected annually from 2018 to 2023. The datasets were merged using province and year as key identifiers. Standardization and cleaning processes were applied to ensure consistency, followed by an inner join operation to combine the datasets. The resulting dataset, encompassing variables such as harvested area, rice production, rainfall, humidity, and temperature, was converted to CSV format for further processing.

B. EXPLORATORY DATA ANALYSIS

In the descriptive statistical analysis, we examined key statistical measures such as the minimum, maximum, median, and mean values for each variable in the dataset. This analysis provided insights into the distribution and variability of factors such as harvested area, rainfall, humidity, and temperature, which are critical for predicting rice production.

Table 1 presents a statistical summary of the numerical data in the dataset from 2018 to 2023, covering key variables such as harvested area, rice production, rainfall, humidity,

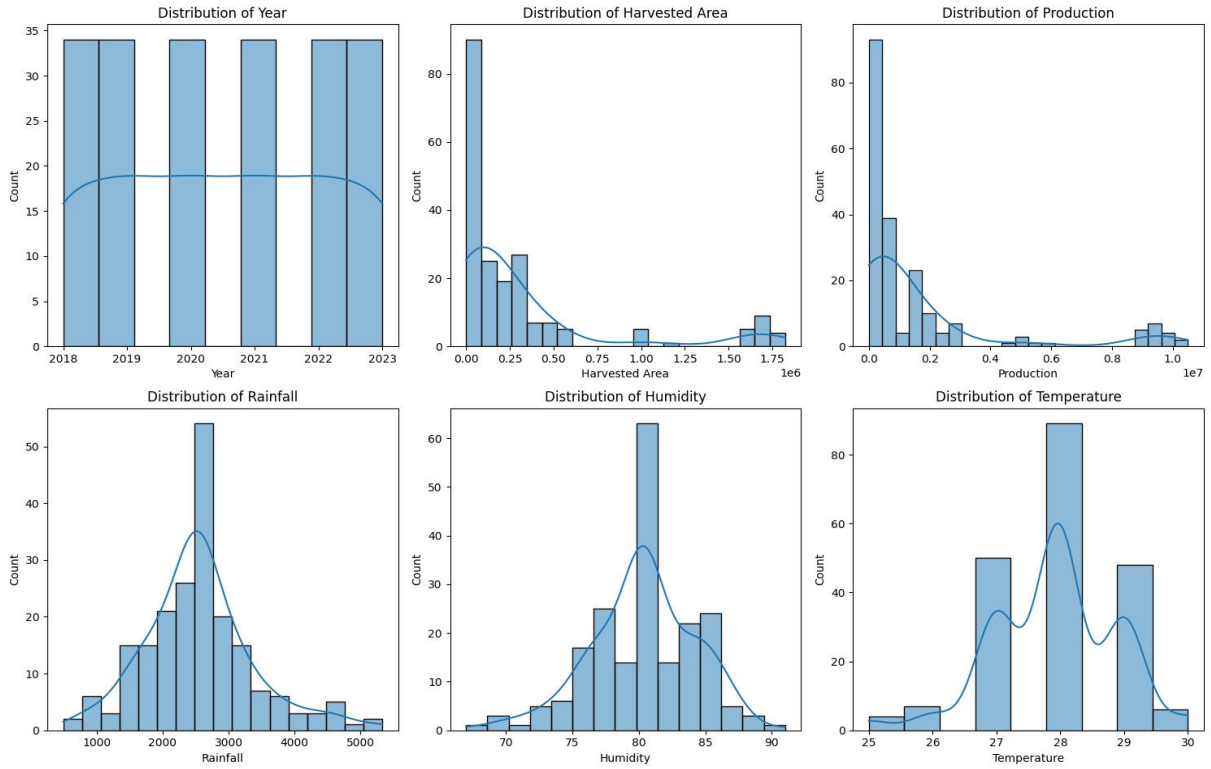

FIGURE 2. Annual distribution of harvest area, production, and climate conditions in Indonesia.

and temperature. The data exhibit substantial variability, particularly in harvested area and rice production, which is reflected in the high standard deviations. This indicates significant disparities across regions in Indonesia, with some provinces showing far greater agricultural output than others. The relatively stable statistics for rainfall and humidity suggest less variation in these climate factors. Understanding this variability is critical, as it directly influences the machine learning models' ability to predict rice yields. The high standard deviation in harvested area and production underscores the complexity of agricultural forecasting in diverse geographic regions, reinforcing the need for robust predictive models that can handle such variability.

Figure 2 illustrates the annual distribution of harvested area, rice production, and climate conditions (rainfall, humidity, and temperature) in Indonesia from 2018 to 2023. The distribution shows considerable variation in harvested area and production, with a right-skewed distribution indicating that a few provinces contribute significantly more than others. Climate variables such as rainfall and humidity exhibit more consistent distributions. This visualization helps identify patterns in the data that may not be immediately apparent from the raw figures, such as the impact of climate variability on production levels. It provides a foundation for exploring relationships between these variables and their influence on rice production, which is critical for the machine learning models' ability to make accurate predictions.

TABLE 1. Statistical summary of the dataset (2018-2023).

	Year	Harvested Area	Production	Rainfall	Humidity	Temperature
count	204.000000	2.040000e+02	2.040000e+02	170.000000	170.000000	170.000000
mean	2020.500000	3.126199e+05	1.623743e+06	2555.252941	80.276471	27.905882
std	1.712026	4.735449e+05	2.682466e+06	915.039696	4.345988	1.044835
min	2018.000000	1.390000e+02	4.230000e+02	490.000000	67.000000	25.000000
25%	2019.000000	5.226425e+04	2.300308e+05	1985.750000	77.000000	27.000000
50%	2020.500000	1.112275e+05	5.310985e+05	2452.500000	80.500000	28.000000
75%	2022.000000	3.126992e+05	1.560927e+06	3051.000000	84.000000	29.000000
max	2023.000000	1.821983e+06	1.049959e+07	5332.000000	91.000000	30.000000

Figure 3 illustrates rice production across various provinces in Indonesia, highlighting considerable disparities in production levels. Notably, provinces such as East Java (Jawa Timur), Central Java (Jawa Tengah), and West Java (Jawa Barat), represented by green bars, emerge as the predominant rice producers, significantly outpacing other regions. In contrast, provinces like the Riau Islands (Kepulauan Riau) and DKI Jakarta exhibit relatively low production levels. The vertical black bars above each column represent the confidence interval or standard error of the mean production. Provinces with longer vertical lines indicate greater uncertainty in their average production, attributable to fluctuations in annual yield.

Figure 4 displays the top 10 rice-producing provinces in Indonesia. East Java, Central Java, and West Java lead in production, with substantially higher yields compared to

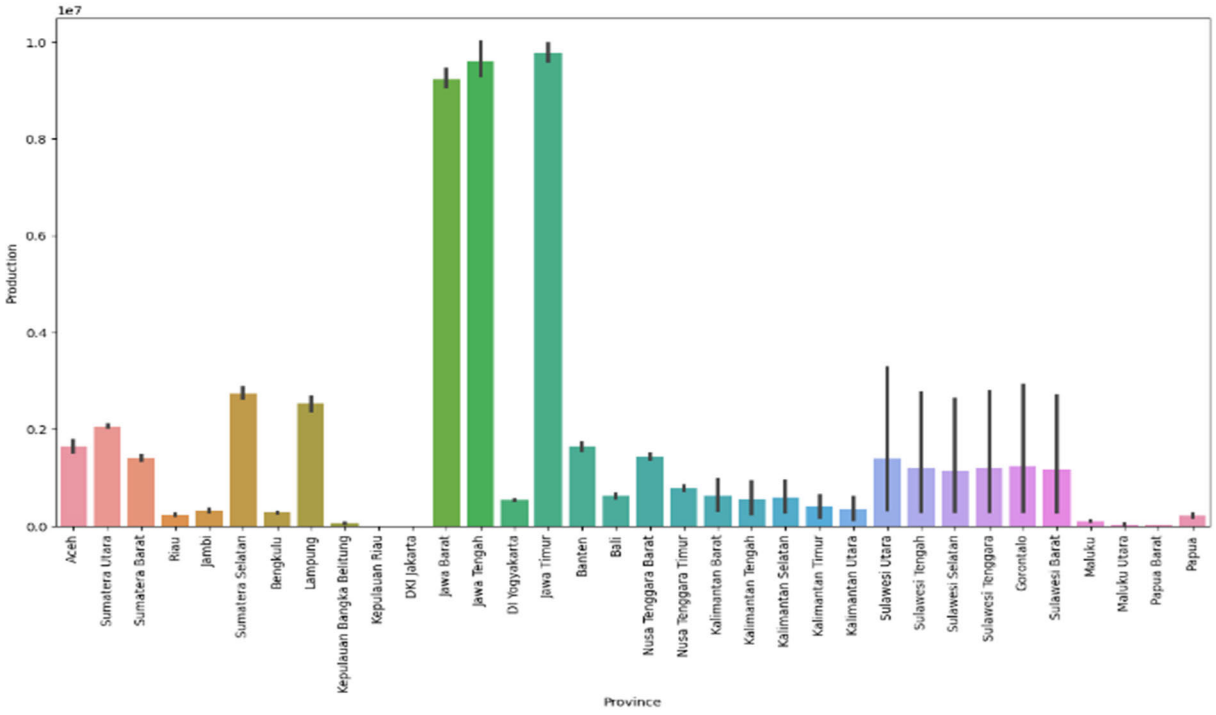

FIGURE 3. Rice production across various provinces in Indonesia.

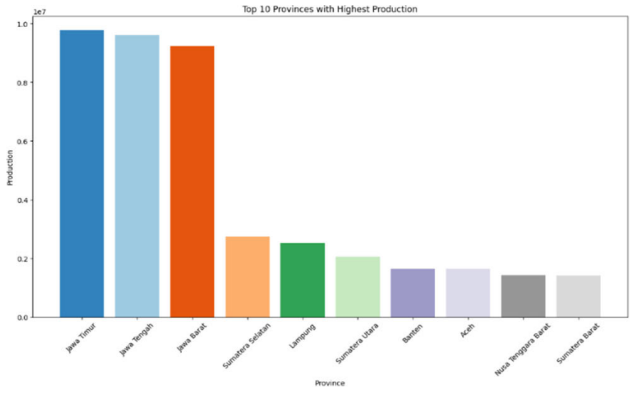

FIGURE 4. Top ten provinces with the highest production.

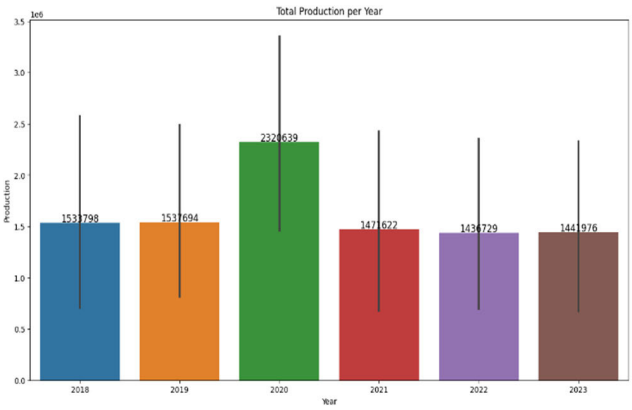

FIGURE 5. The annual trend of total rice production.

other provinces. While South Sumatra (Sumatera Selatan) and Lampung also demonstrate considerable production, their output is not as high as that of the three Javanese provinces. This suggests favorable conditions for rice cultivation in these areas as well. A noticeable decrease in production is evident from the provinces with the largest output to those with smaller yields, indicating various factors influencing the rice production capacity of each province. Notably, production is geographically diverse, spreading across several islands, including Sumatra and Sulawesi, thus showcasing Indonesia's geographic diversification in rice production. The presence of multiple provinces with high production levels is a positive sign for Indonesia's food security, especially considering rice as the country's staple food.

Figure 5 illustrates the annual trend of total rice production from 2018 to 2023. There was a significant surge in production in 2020, marking a notable increase compared to the preceding years. Following the peak in 2020, a decline was observed in 2021, with production stabilizing in 2022 and 2023. This trend may reflect external factors affecting production or a return to normal levels after the exceptional year of 2020. The production in 2018 and 2019 was remarkably consistent, with nearly identical figures. This data is crucial for stakeholders in planning capacity, logistics, and policy-making.

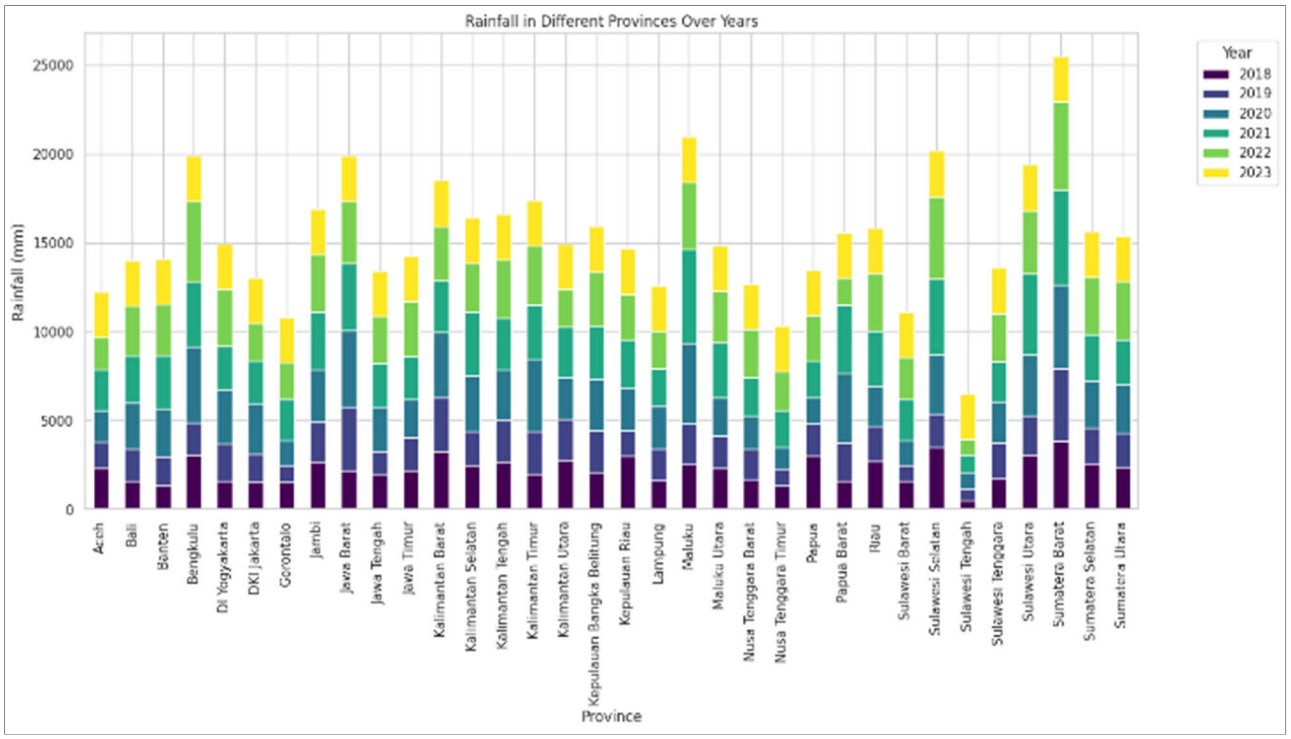

FIGURE 6. Amount of rainfall in various provinces in Indonesia (2018-2023).

Figure 6 presents the rainfall amounts in various provinces from 2018 to 2023. Provinces like Sumatera Barat (West Sumatra) and Maluku exhibit significantly higher rainfall compared to others, such as Sulawesi Tengah (Central Sulawesi) and Nusa Tenggara Timur (East Nusa Tenggara).

A notable trend of either increasing or decreasing rainfall is observed in several provinces over these years. For instance, Kepulauan Bangka Belitung, Bali and Nusa Tenggara Barat (West Nusa Tenggara) have shown a relatively stable amounts, while other provinces have more varied fluctuations such as Aceh and DKI Jakarta. Marked changes in rainfall from one year to the next are evident in certain provinces, possibly due to annual climate phenomena, which significantly affect rainfall patterns. Some provinces consistently record high annual rainfall, suggesting a generally wetter climate, whereas others, with consistently low rainfall, may have a drier climate or be geographically influenced, such as being on the leeward side of mountains which create a rain shadow effect.

A correlation analysis was conducted using Pearson correlation coefficients to identify relationships between variables such as harvested area, production, rainfall, humidity, and temperature. Understanding these correlations can provide valuable insights into how different environmental factors and agricultural metrics are interrelated, which can be useful for decision-making in agricultural planning and management. Table 2 displays the Pearson correlation coefficients calculated to quantify these linear relationships. The analysis revealed a strong positive correlation (r=0.90) between harvested area and production, indicating that larger harvested areas tend to yield higher production. In contrast, environmental factors such as rainfall, humidity, and temperature show weaker correlations with production, suggesting that these factors may not be primary determinants of rice production outcomes in the dataset. The moderate correlations among the environmental factors themselves (e.g., rainfall and humidity) highlight their interconnected nature.

Figure 7 is a visualization of a correlation analysis in the form of a heatmap of the correlation matrix for harvested area, production, rainfall, humidity, and temperature. The heatmap provides a quick and effective visual summary of the relationships between the variables. It visually represents the Pearson correlation coefficients between each pair of variables, with the intensity of the colors indicating the strength and direction of the correlations. The correlation coefficients range from -1 to 1, where values closer to 1 or -1 indicate a stronger positive or negative linear relationship, respectively, and values around 0 indicate no linear relationship. Similar to the observations from the Pearson correlation coefficients, the correlation matrix also shows a strong positive correlation (0.9) between harvested area and production, indicating that an increase in harvested area is strongly associated with an increase in production. Conversely, for example, there is a weak negative correlation (-0.2) between production and humidity, suggesting that higher humidity is slightly associated with lower production levels. These findings help identify the most influential

TABLE 2. Pearson correlation coefficients between harvested area, production, rainfall, humidity, and temperature.

	Harvested Area	Production	Rainfall	Humidity	Temperature
Harvested Area	1.000000	0.903459	0.065607	-0.250314	-0.131761
Production	0.903459	1.000000	-0.012981	-0.199966	-0.179358
Rainfall	0.065607	-0.012981	1.000000	0.344409	-0.136164
Humidity	-0.250314	-0.199966	0.344409	1.000000	-0.518513
Temperature	-0.131761	-0.179358	-0.136164	-0.518513	1.000000

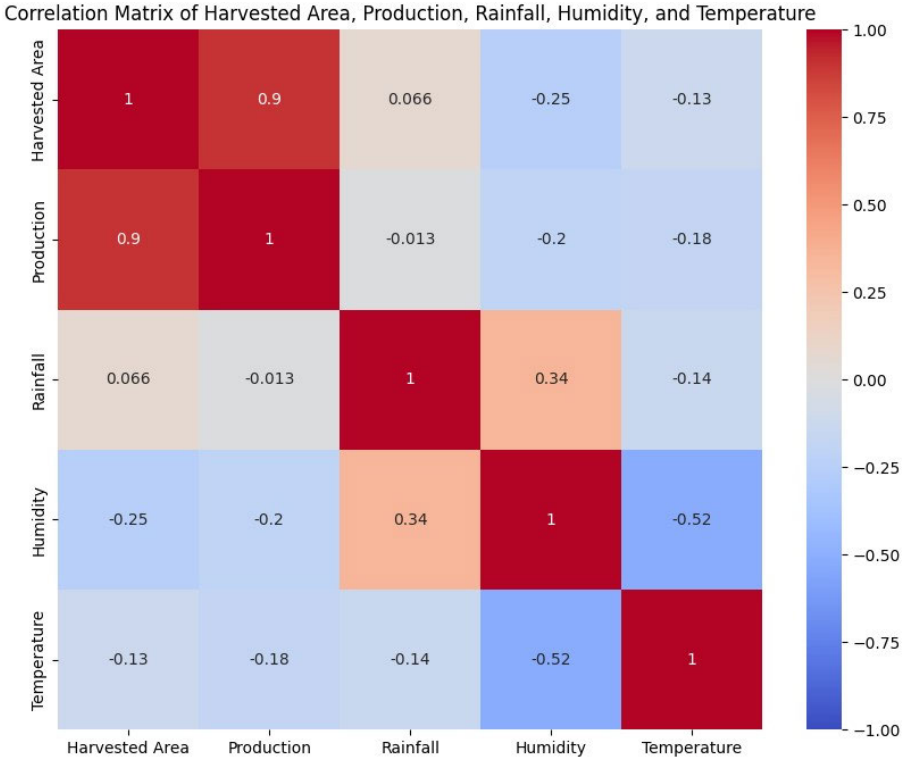

FIGURE 7. Heatmap of a correlation matrix.

TABLE 3. Missing feature values in the dataset.

	Province	Year	Harvested Area	Production	Rainfall	Humidity
199	Papua	2019	54132	235340	1823.0	77.0
200	Papua	2020	52728	166002	1502.0	75.0
201	Papua	2021	64985	286280	2028.0	76.0
202	Papua	2022	49742	193944	2576.0	84.0
203	Papua	2023	49323	200115	NaN	NaN

	Temperature
199	28.0
200	28.0
201	28.0
202	28.0
203	NaN

variables for predicting rice production and guide feature selection for machine learning models.

C. PRE-PROCESSING

After completing the Exploratory Data Analysis (EDA) stage, which involves understanding the data and the correlations between dataset variables, the next step is the pre-processing stage. This stage begins with data cleaning. The dataset contains missing values in the Rainfall, Humidity, and Temperature features (Table 3). Before proceeding with further analysis or modeling, these missing values will be addressed using mean imputation, where empty values are replaced with the column's average. This method is particularly suitable as the features with missing data are numerical [41], [42]. Additionally, this study also checks for data duplication, as duplicates can lead to errors in statistical analysis and impact the performance of the model.

The One-Hot Encoding technique is utilized in data pre-processing to transform categorical variables into a format suitable for machine learning algorithms. Since most of these algorithms perform optimally with numerical data, one-hot encoding is valuable for converting categorical data into binary vectors. In this format, each category is represented by a separate column. Table 4 demonstrates the application of One-Hot Encoding to the dataset, where the categorical Province feature generates new columns. In this arrangement, each data row will have only one column marked with the

TABLE 4. New data frame resulting from the implementation of one-hot-encoding.

	Harvested Area	Production	Rainfall	Humidity	Temperature	Province_Aceh	Province_Bali	Province_Banten	Province_Bengkulu	Province_DI Yogyakarta	...	Province_Papua Barat	Province_Riau
0	329516	1861567	2336	81	28	1	0	0	0	0	...	0	0
1	310012	1714438	1437	82	27	1	0	0	0	0	...	0	0
2	317869	1757313	1790	76	29	1	0	0	0	0	...	0	0
3	297058	1634640	2293	76	29	1	0	0	0	0	...	0	0
4	271750	1509456	1834	76	29	1	0	0	0	0	...	0	0

5 rows × 39 columns

TABLE 5. The dataset was normalized using MinMaxScaler.

Harvested Area	Production	Rainfall	Humidity	Temperature
0.180793	0.177266	0.381247	0.583333	0.6
0.170088	0.163253	0.195580	0.625000	0.4
0.174400	0.167336	0.268484	0.375000	0.8
0.162977	0.155652	0.372367	0.375000	0.8
0.149086	0.143729	0.277571	0.375000	0.8

value '1', denoting that specific category, while all other columns will hold the value '0'. Subsequently, the 'Year' variable is removed from the one-hot encoded data frame as it is not pertinent to production predictions, simultaneously aiding in the reduction of the dataset's dimensions.

The dataset's features vary significantly in scale. For instance, the Humidity and Temperature features have a range of 0-100, whereas the Harvested Area and Production features span from 0 to 10,000,000. To address this disparity, the dataset is normalized using the MinMaxScaler, which adjusts all numerical features to a uniform scale, specifically a range between 0 and 1 as demonstrated in Table 5. This normalization is crucial as features with larger scales can disproportionately influence the model's results, potentially leading to ineffective performance.

The final step in pre-processing involves dividing the dataset into training and test sets, which is achieved using the Train-Test-Split function from the scikit-learn library in Python. Based on common usage in existing research and justification from the well-known Pareto principle, the percentage ratio of 80:20 (which means 80% of the data is for training and 20% for testing) is the most suitable ratio for data splitting [43]. Therefore, this research used a percentage ratio of 80% for training data with 163 samples and 20% for the test data with 41 samples. Furthermore, the random-state parameter has been set to 42 to allow for reproducible results.

D. MODEL TRAINING AND TESTING

In this phase, the prediction algorithm is trained and tested using five different models: Random Forest (RF), Gradient Boosting (GB), Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Networks (ANN). Each model was trained on the same dataset, which had been carefully pre-processed through stages including data cleaning, imputation of missing values, feature normalization, and the handling of categorical variables using one-hot encoding techniques. The purpose of comparing these five models is to assess and understand their performance in predicting rice production in Indonesia, considering factors such as accuracy, and generalization ability. This comparison will yield valuable insights into the effectiveness of various machine learning techniques in practical scenarios, enabling more informed decisions regarding the most suitable algorithm for rice production prediction applications.

Gradient Boosting and Random Forest are ensemble algorithms that are trained by creating and combining a series of weak models to form stronger predictions. In GB, these models are built sequentially, with each new model focusing on correcting the errors made by its predecessor. In contrast, RF trains numerous decision trees in parallel, and their collective results are averaged to produce the final prediction.

Decision Trees are developed by dividing the training set based on features that provide the most effective information gain or uncertainty reduction at each stage. This process continues until either all data in a node have the same label, or a pre-set threshold is reached to prevent overfitting. The Support Vector Machine, or in its regression variant, Support Vector Regression, is trained by determining the hyperplane in a multi-dimensional space that most effectively separates the data, either by maximizing the margin or minimizing error. Artificial Neural Network is trained via a backpropagation process. This involves passing data through the network, computing the errors, and iteratively adjusting the network weights to reduce these errors.

E. MODEL EVALUATION

The performance of each model is measured and analyzed using the Mean Squared Error (MSE) and the coefficient of determination (R-squared) metrics. Evaluating these models is crucial not only for identifying the most suitable one for predicting rice production in Indonesia but also for offering insights into each model's behavior with previously unseen data. This evaluation is key in determining how effectively a model captures patterns in the data and makes accurate predictions, thereby serving as a critical step in verifying the model's overall effectiveness.

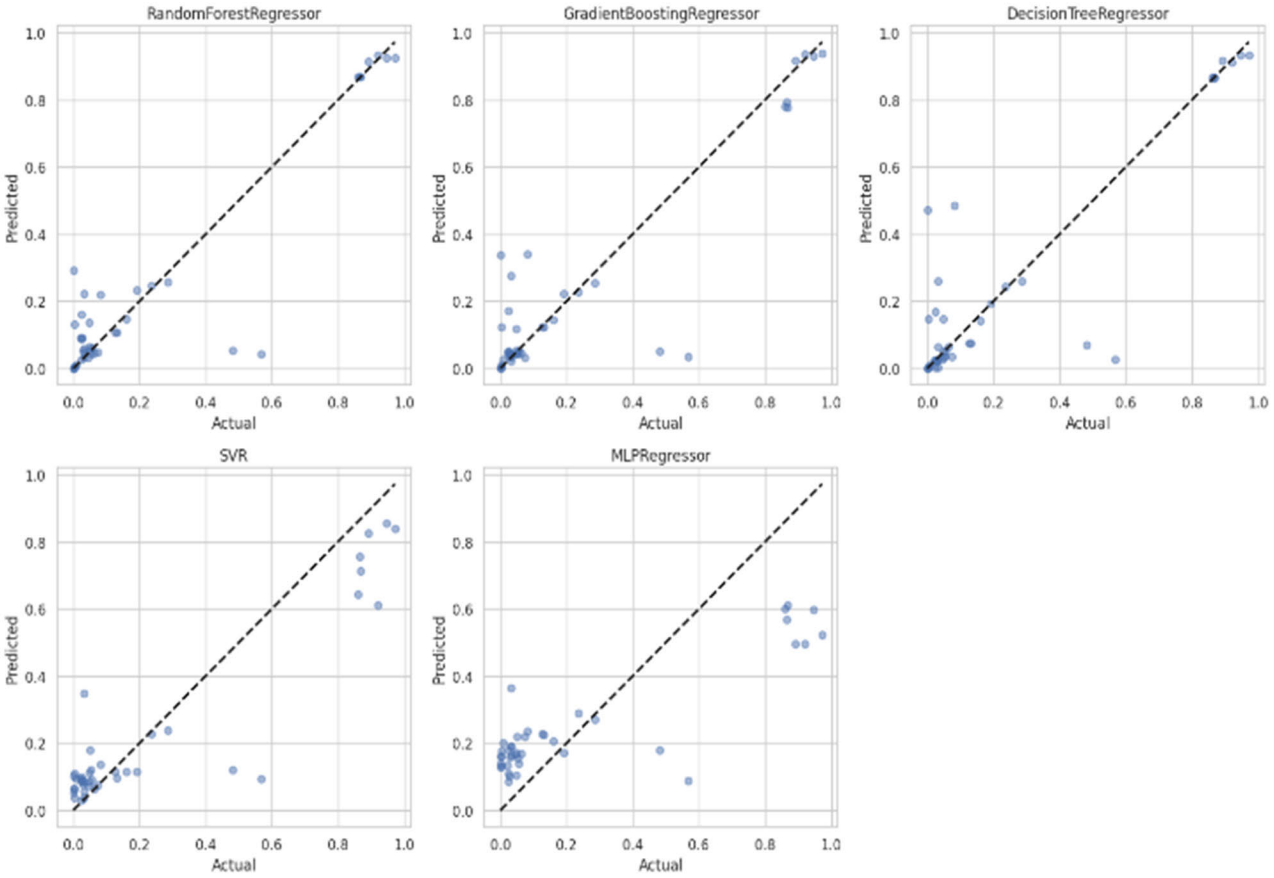

FIGURE 8. Comparison of actual and predicted values by five machine learning models.

Table 6 presents the evaluation results of the five applied models. The Random Forest model exhibits the lowest Mean Squared Error (MSE) at 0.016186, suggesting it has the smallest squared prediction error among the models. Its high R-squared value (0.850039) indicates excellent ability to explain variance in the data and consistent performance in this set. The Gradient Boosting model has a marginally higher MSE (0.018909) compared to Random Forest, implying slightly greater MSE. Despite this, its high R-squared (0.824811) demonstrates effective variance prediction, although not as proficient as Random Forest. The Decision Tree model shows a higher MSE (0.023467) than the preceding models, indicating a tendency for larger prediction errors. Its lower R-squared (0.782577) suggests reduced capability in explaining data variance compared to Random Forest and Gradient Boosting. The Support Vector Machine (SVM) records an MSE (0.018728) close to Gradient Boosting, indicating similar prediction error performance. Its competitive R-squared (0.826485) suggests the model's effectiveness in predicting data variance. The Artificial Neural Network (ANN), with a significantly higher MSE (0.043397), demonstrates less precision in predictions for this dataset. Its considerably lower R-squared (0.597923) suggests that the ANN struggles to adapt to the data and may require more intensive tuning.

TABLE 6. Performance evaluation of the five models.

Model	Mean Squared Error	R^2 Score
Random Forest	0.016186	0.850039
Gradient Boosting	0.018909	0.824811
Decision Tree	0.023467	0.782577
Support Vector Machine	0.018728	0.826485
Artificial Neural Network	0.043397	0.597923

Referring to the data in Table 6, the Random Forest model emerges as the most superior, boasting a low MSE coupled with a high R^2 value. This indicates that Random Forest not only yields predictions closely aligned with actual values but also demonstrates consistency and reliability in its predictions. Gradient Boosting and Support Vector Machine also display commendable results, positioning them as viable alternatives. While Decision Trees are simple and easily interpretable, they fall short in performance compared to the ensemble models. The Artificial Neural Network, on the other hand, registers the lowest performance among the five models in terms of both MSE and R^2 values.

The scatter plots compare the actual versus predicted values from the five models presented in Figure 8. In the

Random Forest plot, the points are closely clustered around the diagonal line, indicating a high level of accuracy in the predictions. This model's predicted values align well with the actual values, reflecting the high R^2 score previously discussed. In Gradient Boosting, the predictions are also close to the diagonal, but with a bit more spread than Random Forest, which is consistent with the slightly higher MSE and lower R^2 score. It indicates a strong predictive performance, albeit with marginally greater errors than the Random Forest. In Decision Tree, the spread of points is wider compared to the ensemble methods above, particularly for higher actual values. This dispersion suggests that while the model captures the general trend, it has larger individual errors, which is in line with the higher MSE and lower R^2 score. In Support Vector Machine, the points here are moderately close to the diagonal line, suggesting good predictive accuracy. However, some points are further away from the line, indicating occasional larger errors. In Artificial Neural Network, the plot shows a more significant spread of points, with several outliers far from the diagonal. This indicates a relatively lower accuracy in the predictions, which corresponds to the highest MSE and the lowest R^2 score among all the models.

Overall, these plots visually reinforce the quantitative findings from the model evaluation figure. The Random Forest, Gradient Boosting, and Support Vector Machine models show the highest level of agreement between actual and predicted values, signifying better performance. The Decision Tree has more variability, and the Artificial Neural Network shows the largest deviations.

F. HYPERPARAMETER TUNING

In this study, hyperparameter tuning was conducted for each machine learning model to enhance their performance in predicting rice production. The tuned hyperparameters and their optimal values are shown in Table 7. Hyperparameter tuning involves finding the best combination of hyperparameters to maximize model performance.

For the RF model, the optimal hyperparameters were a max_depth of 10, min_samples_leaf of 1, min_samples_split of 2, and n_estimators of 200. This combination strikes a balance between bias and variance, preventing overfitting while capturing the data's complexity.

The GB model was optimized with a learning_rate of 0.1, max_depth of 3, min_samples_leaf of 1, min_samples_split of 5, and n_estimators of 100. These settings allow the model to learn at a controlled pace and avoid overfitting by limiting tree depth and the minimum number of samples at leaf nodes.

For the DT model, the criterion was set to friedman_mse, with a max_depth of 10, min_samples_leaf of 2, min_samples_split of 2, and max_features set to sqrt. These hyperparameters manage tree growth, maintaining interpretability and preventing overfitting by restricting tree depth and the number of samples required for splits.

The optimal hyperparameters for the SVM model were C at 10, gamma at 1, and a polynomial kernel. These settings balance the trade-off between minimizing training error and model complexity, with the polynomial kernel and specific values of C and gamma capturing non-linear data relationships.

TABLE 7. Hyperparameter optimization.

Machine Learning Models	Hyperparameters	Best Value
Random Forest	max_depth	10
	min_samples_leaf	1
	min_samples_split	2
	n_estimators	200
Gradient Boosting	learning_rate	0.1
	max_depth	3
	min_samples_leaf	1
	min_samples_split	5
	n_estimators	100
Decision Tree	criterion	friedman_mse
	max_depth	10
	min_samples_leaf	2
	min_samples_split	2
	max_features	sqrt
Support Vector Machine	C	10
	gamma	1
	kernel	poly
Artificial Neural Network	activation	tanh
	alpha	0.001
	hidden_layer	100
	learning_rate	constant
	max_iter	1000
	solver	sgd

The ANN model's optimal hyperparameters were an activation function of tanh, alpha of 0.001, hidden_layer size of 100, learning_rate set to constant, max_iter of 1000, and solver set to sgd. These parameters enable the ANN to learn effectively, with the tanh activation handling non-linearity, alpha controlling regularization, and the learning rate, hidden layers, and solver ensuring a balance between learning speed and convergence.

Hyperparameter tuning is essential for improving the predictive accuracy of machine learning models. The optimal hyperparameters identified in this study ensure that each model is finely tuned to handle the complexities of the rice production dataset. These optimized models, show good performance and reliability in predicting rice production, confirming their suitability for this application.

G. MODEL COMPARISON

At this stage, this research has progressed to a point where it can evaluate and confirm the efficacy of the models selected through a thorough hyperparameter optimization process. Following a comprehensive series of experiments, which included cross-validation and various hyperparameter search strategies, this research has pinpointed the most effective models by comparing the Mean Squared Error (MSE) and R-squared (R^2) scores of the tuned models (results from GridSearchCV) against those of the original models, as detailed in Table 8.

The comparison of the machine learning models' performance before and after tuning reveals significant improvements across all metrics. For the Random Forest model, tuning reduced the MSE from 0.016186 to 0.014162 and

TABLE 8. Comparison of machine learning models' performance before and after tuning.

Model	MSE		R^2	
	Before Tuning	After Tuning	Before Tuning	After Tuning
Random Forest	0.016186	0.014162	0.850039	0.911257
Gradient Boosting	0.018909	0.016205	0.824811	0.883401
Decision Tree	0.023467	0.020997	0.782577	0.845429
Support Vector Machine	0.018728	0.016878	0.826485	0.874312
Artificial Neural Network	0.043397	0.038794	0.597923	0.632014

increased the R^2 from 0.850039 to 0.911257, indicating enhanced accuracy and explanatory power. Similarly, the Gradient Boosting model saw its MSE decrease from 0.018909 to 0.016205 and its R^2 rise from 0.824811 to 0.883401. The Decision Tree model's performance also improved, with the MSE dropping from 0.023467 to 0.020997 and the R^2 increasing from 0.782577 to 0.845429, though it still lagged behind the ensemble methods. The Support Vector Machine (SVM) model showed a decrease in MSE from 0.018728 to 0.016878 and an increase in R^2 from 0.826485 to 0.874312, indicating better predictive capabilities post-tuning. Lastly, the Artificial Neural Network (ANN) had the highest MSE and lowest R^2 both before and after tuning (MSE from 0.043397 to 0.038794 and R^2 from 0.597923 to 0.632014), suggesting it is the least effective model among those evaluated, despite showing some improvement.

Ensemble methods, particularly the Random Forest and Gradient Boosting models, generally perform better than individual models such as Decision Tree, SVM, and ANN, both before and after tuning. This aligns with common findings in machine learning, where ensemble methods often outperform single models. After tuning, the Random Forest model exhibits the lowest MSE (0.014162) and the highest R^2 (0.911257), making it the best-performing model in this comparison. This highlights the importance of model tuning in improving predictive accuracy and explanatory power, as all models show an improvement in performance metrics after tuning.

Despite improvements from tuning, the ANN model lags behind the other models in terms of both MSE and R^2. This suggests that further optimization or perhaps a different architecture might be necessary to enhance its performance. The comparison highlights the significant impact of tuning on the performance of machine learning models. Ensemble methods, particularly the Random Forest model, show superior performance after tuning, making them suitable choices for predictive tasks. However, even models with initially lower performance, such as the ANN, benefit from tuning, demonstrating the value of this step in the modeling process.

This evaluation of the models, especially the best model, provides insights not only into their predictive accuracy for rice production but also assesses their robustness and reliability in real-world scenarios. Consequently, this enables a precise measurement of the models' suitability for practical implementation in real applications.

Table 9 presents a comprehensive performance comparison of various models and techniques used for predicting agricultural outputs, particularly focusing on rice yield estimation. Mongkolnithitada, Nontapun, and Kaewplang utilize machine learning using MODIS, achieving an R^2 of 0.760 and RMSE of 291.7. Their approach integrates machine learning with remote sensing data, enhancing accuracy and reliability. Moreover, the use of multiple models ensures robustness, with Random Forest likely contributing to the highest accuracy. The R^2 value suggests good explanatory power, while the RMSE indicates a moderate prediction error.

Similarly, Li et al. employ a random forest model, achieving an R^2 higher than 0.75 and nRMSE lower than 18.0%, highlighting their development of a within-growing season yield forecast system. The high R^2 and low nRMSE values indicate a strong predictive performance and reliability of the Random Forest model within the growing season. Furthermore, Ansari, Lin, and Lur use the CERES-Rice model from the DSSAT software, achieving an R^2 of 0.89 and NSE of 0.88, effectively simulating rice production. The CERES-Rice model demonstrates high accuracy and efficiency in simulating rice production. The high R^2 and NSE values reflect the model's ability to closely match observed outcomes. Conversely, Han et al. employ a regression model, resulting in an R^2 of 0.462 and an adjusted R^2 of 0.453, focusing on integrating remote sensing imagery with deep learning to enhance agricultural applications. The lower R^2 values indicate that the regression model may not capture the complexity of the data as effectively as other models. However, it provides a baseline for comparison and highlights areas for improvement through the integration of additional data sources or advanced techniques.

Additionally, Longfei et al. utilize a UAV-based multi-temporal feature method, achieving an R^2 of 0.795, RMSE of 0.298, and RRMSE of 0.072, which aids in pre-harvest yield prediction using UAV-based remote sensing. This model leverages UAV-based remote sensing, which provides detailed spatial and temporal data, thereby leading to high accuracy in pre-harvest yield prediction. Manasa et al. achieve an R^2 of 0.87 with multiple linear regression and hybrid feature selection techniques, emphasizing the importance of feature selection in machine learning. The high R^2 indicates that the inclusion of hybrid feature selection techniques significantly enhances model performance, highlighting the importance of selecting relevant features in machine learning algorithms. Meanwhile, Mia et al. combine UAV-based multispectral images and weather data using convolutional neural networks, achieving an R^2 of 0.68 and RMSE of 0.821. Although the R^2 is moderate, the convolutional neural networks (CNN) model combines UAV-based

TABLE 9. Performance comparison of different models.

Model/Technique	Metrics Used	Key Features
Machine Learning using MODIS [26]	$R^2 = 0.760$ RMSE = 291.7	Integrates machine learning with remote sensing data to enhance the accuracy and reliability
Random Forest [29]	$R^2 > 0.75$ nRMSE < 18.0%	Develop a within-growing season yield forecast system with random forest model.
CERES-Rice model [33]	$R^2 = 0.89$ MSE = 0.88	Employs the CERES-Rice model from the Decision Support System for Agrotechnology Transfer (DSSAT) software to simulate rice production
Regression Model [44]	$R^2 = 0.462$ Adjusted $R^2 = 0.453$	Focuses on the innovative integration of remote sensing imagery with deep learning techniques to enhance agricultural applications.
Unmanned Aerial Vehicle-Based Multi-Temporal Feature Method [45]	$R^2 = 0.795$ RMSE = 0.298, RRMSE = 0.072	Pre-harvest yield prediction of ratoon rice using UAV-based remote sensing to aid precision agriculture.
Multiple Linear Regression and Hybrid Feature Selection Techniques [46]	$R^2 = 0.87$	Highlights the crucial role of feature selection in enhancing the productivity and accuracy of machine learning algorithms.
Convolutional Neural Networks [47]	R2 = 0.68 RMSE = 0.821 RMSPE =13	Combines UAV-based multispectral images and weather data for precise yield predictions
Transformer-based Model (Informer) [48]	$R^2 = 0.81$ MSE = 0.41	Integrates time-series satellite data and environmental variables.
Hybrid Deep Neural Networks with Multi-Tasking [49]	$R^2 = 0.64$ RMSE = 344.56	Uses multi-task learning combining classification and regression models with remote sensing data.
Proposed Model		Identifies significant regional disparities, underscoring the complexity of agricultural forecasting in the region of Indonesia.
- Random Forest	$R^2 = 91\%$ MSE = 0.014162	
- Gradient Boosting	$R^2 = 88\%$ MSE = 0.016205	
- Decision Tree	$R^2 = 84\%$ MSE = 0.020997	
- Support Vector Machine	$R^2 = 87\%$ MSE = 0.016878	
- Artificial Neural Network	$R^2 = 63\%$ MSE = 0.038794	

multispectral images and weather data, suggesting potential in capturing complex patterns.

In contrast, Liu et al. use a transformer-based model, achieving an MSE of 0.41 and R^2 of 0.81. The transformer-based model integrates time-series satellite data and environmental variables, showing a high R^2 and relatively low MSE. This indicates strong predictive capability, especially for time-series data. Similarly, Chang et al. use hybrid deep neural networks with multi-tasking, achieving an RMSE of 344.56 and R^2 of 0.64, integrating remote sensing data for multi-task learning. The hybrid deep neural network model, despite a moderate R^2 and relatively high RMSE, benefits from multi-task learning and integration of remote sensing data.

Notably, the proposed random forest model in this study achieves an MSE of 0.014162 and an R^2 of 91%, indicating an exceptional predictive performance. A lower MSE value signifies a higher accuracy of the model's predictions. An MSE of 0.014162 is remarkably low, suggesting that the model's predictions are very close to the actual observed values. An R^2 value of 91% implies that 91% of the variability in rice yield can be explained by the model. This high R^2 value indicates that the model has very strong explanatory power and can reliably predict the outcomes based on the input data. Consequently, this combination of a very low MSE and a very high R^2 demonstrates that the proposed random forest model in this study is highly accurate and superior to other models analyzed in the comparison. The model not only minimizes prediction errors but also explains a significant portion of the variability in rice yield, making it an excellent tool for agricultural research and practical applications in predicting rice production. Additionally, the model's ability to identify significant regional disparities highlights its potential utility in addressing agricultural productivity issues specific to different regions in Indonesia. This level of performance underscores the effectiveness of the random forest approach in handling complex, non-linear relationships inherent in agricultural data.

V. CONCLUSION AND FUTURE WORK

This research has successfully harnessed the potential of machine learning in predicting rice production in Indonesia, a critical aspect of the country's agricultural sector and food security. Utilizing a comprehensive dataset, combining data from the Central Bureau of Statistics of Indonesia and the Meteorology, Climatology, and Geophysics Agency of Indonesia, this study covered a significant period from 2018 to 2023. The exploratory data analysis revealed substantial variability in harvested area, production, and climatic conditions across different provinces, highlighting the complexity of agricultural forecasting in Indonesia.

Through rigorous pre-processing, including data cleaning, imputation of missing values, one-hot encoding, and normalization, the dataset was aptly prepared for effective machine learning applications. The subsequent model training and testing phase evaluated five different machine learning models: Gradient Boosting, Random Forest, Decision Tree, Support Vector Machine, and Artificial Neural Network. Among these, the Random Forest model emerged as the most effective, exhibiting the lowest Mean Squared Error and highest R-squared value, thereby indicating its superior predictive accuracy and reliability. While Gradient

Boosting and Support Vector Machine also showed promising results, the Decision Tree and Artificial Neural Network models displayed limitations in their predictive capabilities. Furthermore, the hyperparameter tuning process, particularly focusing on the Random Forest model, significantly enhanced the model's performance, as evidenced by the improved metrics post-tuning. The best model evaluation reaffirmed the robustness and reliability of the tuned Random Forest model in real-world scenarios, marking it as a suitable choice for practical implementation.

This study not only contributes to the field of agricultural forecasting in Indonesia but also demonstrates the power of machine learning in addressing complex, real-world problems. The results underscore the importance of advanced analytical techniques in improving agricultural productivity and decision-making processes. This research also paves the way for future studies to explore more sophisticated machine learning models and larger datasets, potentially offering even more nuanced insights into agricultural production patterns. By doing so, it is hoped that such advancements will continue to bolster the resilience and sustainability of Indonesia's rice production, ultimately contributing to the nation's food security and economic stability.

REFERENCES

[1] M. Shahbandeh. (2021). *Rice Consumption Worldwide in 2021/2022, By Country (in 1,000 Metric Tons)*. Accessed: Nov. 25, 2023. [Online]. Available: https://www.statista.com/statistics/255971/top-countries-based-on-rice-consumption-2012-2013/#:~:text=Asthemostpopulouscountry,consumptioninthesameperiod

[2] U.S. Department of Agriculture. (2020). *Rice Sector at a Glance 2020/21–22/23*. Accessed: Nov. 25, 2023. [Online]. Available: https://www.ers.usda.gov/topics/crops/rice/rice-sector-at-a-glance/#Global

[3] M. F. Ikhwali, S. Nur, D. Darmansyah, A. M. Hamdan, N. S. Ersa, N. Aida, A. Yusra, and A. Satria, "A review of climate change studies on paddy agriculture in Indonesia," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 1116, no. 1, Dec. 2022, Art. no. 012052.

[4] B. Smerbeck and B. Thompson. (Nov. 21, 2023). *How Accurate is The Old Farmer's Almanac's Weather Forecast?* Almanac. [Online]. Available: https://www.almanac.com/how-accurate-old-farmers-almanacs-weather-forecast

[5] S. K. Purohit, S. Panigrahi, P. K. Sethy, and S. K. Behera, "Time series forecasting of price of agricultural products using hybrid methods," *Appl. Artif. Intell.*, vol. 35, no. 15, pp. 1388–1406, Dec. 2021.

[6] P. Mishra1, "Forecasting of rice production using the meteorological factor in major states in India and its role in food security," *Int. J. Agricult. Environ. Biotechnol.*, vol. 14, no. 1, p. 2021, Mar. 2021.

[7] D. Sihi, B. Dari, A. P. Kuruvila, G. Jha, and K. Basu, "Explainable machine learning approach quantified the long-term (1981–2015) impact of climate and soil properties on yields of major agricultural crops across CONUS," *Frontiers Sustain. Food Syst.*, vol. 6, Apr. 2022, Art. no. 847892.

[8] M. Meroni, F. Waldner, L. Seguini, H. Kerdiles, and F. Rembold, "Yield forecasting with machine learning and small data: What gains for grains?" *Agricult. Forest Meteorol.*, vols. 308–309, Oct. 2021, Art. no. 108555.

[9] P. Kamath, P. Patil, S. Shrilatha, Sushma, and S. Sowmya, "Crop yield forecasting using data mining," *Global Transitions Proc.*, vol. 2, no. 2, pp. 402–407, Nov. 2021.

[10] M. Shahhosseini, G. Hu, and S. V. Archontoulis, "Forecasting corn yield with machine learning ensembles," *Frontiers Plant Sci.*, vol. 11, pp. 1–16, Jul. 2020.

[11] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine learning applications for precision agriculture: A comprehensive review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021.

[12] N. Bali and A. Singla, "Emerging trends in machine learning to predict crop yield and study its influential factors: A survey," *Arch. Comput. Methods Eng.*, vol. 29, no. 1, pp. 95–112, Jan. 2022.

[13] A. Cravero, S. Pardo, S. Sepúlveda, and L. Muñoz, "Challenges to use machine learning in agricultural big data: A systematic literature review," *Agronomy*, vol. 12, no. 3, p. 748, Mar. 2022.

[14] R. Alfred, J. H. Obit, C. P. Chin, H. Haviluddin, and Y. Lim, "Towards paddy rice smart farming: A review on big data, machine learning, and rice production tasks," *IEEE Access*, vol. 9, pp. 50358–50380, 2021.

[15] G. Pradeep, T. D. V. Rayen, A. Pushpalatha, and P. K. Rani, "Effective crop yield prediction using gradient boosting to improve agricultural outcomes," in *Proc. Int. Conf. Netw. Commun. (ICNWC)*, Apr. 2023, pp. 1–6.

[16] B. M. Sagar, N. K. Cauvery, T. Padmashree, and R. Rajkumar, "Rice and wheat yield prediction in India using decision tree and random forest," *Comput. Intell. Mach. Learn.*, vol. 3, no. 2, pp. 1–8, Oct. 2022.

[17] K. Choudhary, W. Shi, Y. Dong, and R. Paringer, "Random forest for rice yield mapping and prediction using Sentinel-2 data with Google Earth engine," *Adv. Space Res.*, vol. 70, no. 8, pp. 2443–2457, Oct. 2022.

[18] K. K. Paidipati, C. Chesneau, B. M. Nayana, K. R. Kumar, K. Polisetty, and C. Kurangi, "Prediction of rice cultivation in India—Support vector regression approach with various kernels for non-linear patterns," *AgriEngineering*, vol. 3, no. 2, pp. 182–198, Apr. 2021.

[19] V. Amaratunga, L. Wickramasinghe, A. Perera, J. Jayasinghe, and U. Rathnayake, "Artificial neural network to estimate the paddy yield prediction using climatic data," *Math. Problems Eng.*, vol. 2020, pp. 1–11, Jul. 2020.

[20] S. Rathod, S. Yerram, P. Arya, G. Katti, J. Rani, A. P. Padmakumari, N. Somasekhar, C. Padmavathi, G. Ondrasek, S. Amudan, S. Malathi, N. M. Rao, K. Karthikeyan, N. Mandawi, P. Muthuraman, and R. M. Sundaram, "Climate-based modeling and prediction of rice gall midge populations using count time series and machine learning approaches," *Agronomy*, vol. 12, no. 1, p. 22, Dec. 2021.

[21] S. Chairani, "The correlation between rainfall, temperature, relative humidity, and rice field productivity in Aceh Besar," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 1071, no. 1, pp. 1–17, 2022.

[22] A. Abdullah and M. H. Nahid, "Performance analysis rice yield model based on historical weather dataset in Bangladesh," in *Proc. 4th Int. Conf. Sustain. Technol. Ind. 4.0 (STI)*, Dec. 2022, pp. 1–6.

[23] K. T. Soberano, J. S. Pisueña, S. M. R. Tee, J. C. T. Arroyo, and A. J. P. Delima, "Predictive soil-crop suitability pattern extraction using machine learning algorithms," *Int. J. Adv. Appl. Sci.*, vol. 10, no. 6, pp. 8–16, Jun. 2023.

[24] Y. Iuchi, H. Uehara, Y. Fukazawa, and Y. Kaneta, "Stabilizing the predictive performance for ear emergence in rice crops across cropping regions," in *Proc. Pacific Rim Knowl. Acquisition Workshop*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, 2021, pp. 83–97.

[25] S. Ngandee and A. Taparugssanagorn, "Improved information dissemination services for the agricultural sector in Thailand: Development and evaluation of a machine learning based rice crop yield prediction system," *Inf. Develop.*, pp. 1–14, Nov. 2023. [Online]. Available: https://journals.sagepub.com/doi/10.1177/02666669231208017

[26] W. Mongkolnithithada, J. Nontapun, and S. Kaewplang, "Rice yield estimation based on machine learning approaches using MODIS 250 m data," *Eng. Access*, vol. 9, no. 1, pp. 75–79, 2023.

[27] P. Roy, B. Kumar, P. K. Bharti, V. K. Vishnoi, K. Kumar, S. Mohan, and K. P. Singh, "Paddy yield prediction based on 2D images of rice panicles using regression techniques," *Vis. Comput.*, vol. 40, no. 6, pp. 4457–4471, Jun. 2024.

[28] X. Zhang, R. Shen, X. Zhu, B. Pan, Y. Fu, Y. Zheng, X. Chen, Q. Peng, and W. Yuan, "Sample-free automated mapping of double-season rice in China using Sentinel-1 SAR imagery," *Frontiers Environ. Sci.*, vol. 11, pp. 1–11, Jul. 2023.

[29] L. Li, B. Wang, P. Feng, H. Wang, Q. He, Y. Wang, D. L. Liu, Y. Li, J. He, H. Feng, G. Yang, and Q. Yu, "Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China," *Agricult. Forest Meteorol.*, vols. 308–309, Oct. 2021, Art. no. 108558.

[30] Y. Tanaka et al., "Deep learning enables instant and versatile estimation of rice yield using ground-based RGB images," *Plant Phenomics*, vol. 5, pp. 1–16, Jan. 2023.

[31] P. Hoang-Phi, T. Nguyen-Kim, V. Nguyen-Van-Anh, N. Lam-Dao, T. Le-Van, and T. Pham-Duy, "Rice yield estimation in An Giang province, the Vietnamese Mekong Delta using Sentinel-1 radar remote sensing data," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 652, no. 1, 2021, Art. no. 012001.
[32] K. R. Thorp and D. Drajat, "Deep machine learning with sentinel satellite data to map paddy rice production stages across West Java, Indonesia," *Remote Sens. Environ.*, vol. 265, Nov. 2021, Art. no. 112679.
[33] A. Ansari, Y.-P. Lin, and H.-S. Lur, "Evaluating and adapting climate change impacts on rice production in Indonesia: A case study of the Keduang subwatershed, Central Java," *Environments*, vol. 8, no. 11, p. 117, Oct. 2021.
[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[35] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[36] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.
[37] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.
[38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
[39] *Environment Statistics of Indonesia 2023*, Central Bur. Statist. Indonesia, Jakarta, Indonesia, 2023.
[40] Meteorology, Climatology, and Geophysics Agency of Indonesia. (2023). *Data Iklim Indonesia*. Accessed: Nov. 2, 2023. [Online]. Available: https://dataonline.bmkg.go.id/home
[41] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, 3rd ed., Hoboken, NJ, USA: Wiley, 2019.
[42] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in *Classification, Clustering, and Data Mining Applications*, D. Banks, F. R. McMorris, P. Arabie, and W. Gaul, Eds. Berlin, Germany: Springer, 2004, pp. 639–648.
[43] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 15, no. 4, pp. 409–538, Aug. 2022.
[44] X. Han, F. Liu, X. He, and F. Ling, "Research on rice yield prediction model based on deep learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–9, Apr. 2022.
[45] Z. Longfei, M. Ran, Y. Xing, L. Yigui, H. Zehua, L. Zhengang, X. Binyuan, Y. Guodong, P. Shaobing, and X. Le, "Improved yield prediction of ratoon rice using unmanned aerial vehicle-based multi-temporal feature method," *Rice Sci.*, vol. 30, no. 3, pp. 247–256, May 2023.
[46] C. M. Manasa, B. Prince, G. R. Arpitha, and A. Verma, "Hybrid feature selection techniques to improve the accuracy of rice yield prediction: A machine learning approach," in *Coating Materials. Materials Horizons: From Nature to Nanomaterials*, A. Verma, S. K. Sethi, and S. Ogata, Eds., Singapore: Springer, 2023, pp. 409–421.
[47] M. S. Mia, R. Tanabe, L. N. Habibi, N. Hashimoto, K. Homma, M. Maki, T. Matsui, and T. S. T. Tanaka, "Multimodal deep learning for rice yield prediction using UAV-based multispectral imagery and weather data," *Remote Sens.*, vol. 15, no. 10, p. 2511, May 2023.
[48] Y. Liu, S. Wang, J. Chen, B. Chen, X. Wang, D. Hao, and L. Sun, "Rice yield prediction and model interpretation based on satellite and climatic indicators using a transformer method," *Remote Sens.*, vol. 14, no. 19, p. 5045, Oct. 2022.
[49] C.-H. Chang, J. Lin, J.-W. Chang, Y.-S. Huang, M.-H. Lai, and Y.-J. Chang, "Hybrid deep neural networks with multi-tasking for rice yield prediction using remote sensing data," *Agriculture*, vol. 14, no. 4, p. 513, Mar. 2024.

ERLIN received the Ph.D. degree in computer science from Universiti Teknologi Malaysia (UTM), in 2011. She is currently an Associate Professor with the Department of Informatics Engineering, Institut Bisnis dan Teknologi Pelita Indonesia. She has actively participated in several research projects funded by the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia. In addition, she has authored several reputable conferences and high-impact journal papers, with her research interests include machine learning, data science, data mining, and soft computing.

ARDA YUNIANTA received the Ph.D. degree in computer science from Universiti Teknologi Malaysia (UTM), in 2015. He is currently an Assistant Professor with the Department of Information Systems, Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Saudi Arabia. He has authored and co-authored papers published in several reputable conferences and high-impact journals. Additionally, he contributes as an editor and a reviewer in some international journals and reputable international conferences. His research interests include ontology, data integration, semantic technology, e-learning systems, big/intelligence data, deep learning, machine learning, data mining, and data science.

LILI AYU WULANDHARI received the Ph.D. degree in computer science from Universiti Teknologi Malaysia (UTM), in 2014. She is currently an Associate Professor with the Computer Science Department, School of Computer Science, Bina Nusantara University. She has actively participated in several research projects funded by the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia and Bina Nusantara University. She has authored several reputable conference and journal papers. In addition, she has expertise in the fields of artificial intelligence, artificial neural networks, and machine learning. She also has a good academic reputation and experience as a reviewer for several reputable journals.

YENNY DESNELITA received the Ph.D. degree in computer science from Universitas Negeri Padang (UNP), in 2020. She is currently an Associate Professor with the Department of Information Systems, Institut Bisnis dan Teknologi Pelita Indonesia. She has actively participated in several research projects funded by the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia. She has authored several conference and journal papers, with her research interests include artificial intelligence, decision support systems, and educational technology.

NURLIANA NASUTION received the Ph.D. degree in computer science from Universitas Negeri Padang (UNP), in 2019. She is currently an Associate Professor with the Department of Information Technology, Universitas Lancang Kuning. She has actively participated in several research projects funded by the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia and Universitas Lancang Kuning. She has authored several conference and journal papers, with her research interests include augmented reality, the IoT, blended learning, and machine learning.

JUNADHI received the master's degree in computer science from Universitas Putra Indonesia (UPI) "YPTK" Padang, in 2016, where he is currently pursuing the Ph.D. degree. He is currently a Senior Lecturer with the Informatics Engineering Department, Universitas Sains dan Teknologi Indonesia. He has authored several conference and journal papers, with his research interests include data science, UI UX design, and mobile programming.