



Comparison Of Machine Learning Algorithms For Rice Production Prediction

Abdul Karim¹, Yuwalidi Away², Syahrial³, Roslidar⁴, Jeperson Hutahaean⁵
William Ramdhan⁶, Yessica Siagian⁷

¹Faculty of Science and Technology, Information Technology Study Program, Universitas Labuhanbatu, Labuhanbatu, Indonesia

^{2,3,4}Graduate Program of Electrical and Computer Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia

^{5,6,7}Faculty of Computer Science, Information Systems, Universitas Royal, Asahan, Indonesia

*abdulkarim@ulb.ac.id

Abstract

Rice production forecasting plays an important role in supporting future agricultural planning, food supply management, and food security. Accurate yield prediction allows governments and farmers to estimate production outcomes and develop appropriate strategies to maintain stable food availability. This study addresses this gap by comparing four regression-based machine learning models: Random Forest, XGBoost, Support Vector Regression (SVR), and Artificial Neural Network (ANN). All models were trained and tested using the same dataset to ensure a fair evaluation. Model performance was measured using the coefficient of determination (R^2). The results show that Random Forest achieved the best performance ($R^2 = 0.963$), followed by XGBoost ($R^2 = 0.959$). In contrast, SVR ($R^2 = -0.064$) and ANN ($R^2 = -2.417$) performed poorly, indicating limited predictive capability. Overall, these findings suggest that ensemble-based methods, particularly Random Forest and XGBoost, are more reliable and effective for rice production forecasting compared to SVR and ANN.

Keywords: ANN; random forest; rice; SVR; XGBoost

How to Cite: A. Karim, "Comparison Of Machine Learning Algorithms For Rice Production Prediction", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 2, pp. 407 - 418, Apr. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i2.7453>

Received: January 7, 2026

Accepted: April 4, 2026

Available Online: April 25, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Rice production is very important in maintaining food security, especially in developing countries where rice is the main staple food. In Indonesia, rice remains the most important agricultural commodity, and its production is strongly influenced by several environmental and agronomic factors, including harvested area, rainfall, humidity, and temperature. Variations in these factors, especially under the increasing uncertainty caused by climate change, often lead to fluctuations in agricultural productivity that are difficult to predict using conventional statistical approaches. Accurate prediction of rice production is therefore essential to support agricultural planning, optimize resource allocation, and reduce the risk of food shortages. In recent years, advances in data-based technology have driven an increase in the use of machine learning methods to analyze agricultural data,

enabling more accurate and reliable predictions compared to conventional forecasting approaches

Machine learning approaches have demonstrated significant potential in agricultural yield prediction because of their ability to capture nonlinear relationships and complex interactions among environmental variables [1-3]. Various algorithms have been applied in crop production forecasting, including Random Forest [4], XGBoost [5], Support Vector Regression [6-8], and Artificial Neural Networks [9], [10]. Ensemble Learning method, particularly Random Forest, are well known for their ability to manage complex, high-dimensional datasets while limiting overfitting by combining the outputs numerous decision trees. In a similar vein, gradient boosting techniques like XGBoost are extensively applied in predictive analytics because they refine model performance step by step, continuously reducing errors

and enhancing overall accuracy. Meanwhile, SVR has been utilized in many regression-based prediction problems because it can effectively model nonlinear relationships using kernel functions. Artificial Neural Networks are widely used in agricultural forecasting because of their ability to capture complex relationships between input and output variables through layered network structures. Empirical findings from prior research indicate that such machine learning approaches often surpass conventional statistical models in estimating crop yields, particularly when handling large-scale and diverse agricultural data. [11].

Despite these advancements, several limitations remain in existing studies. Many previous works focus on implementing a single machine learning algorithm without conducting a systematic comparison among different models under the same experimental conditions. Furthermore, most studies are conducted using datasets from specific countries or regions, which limits the generalizability of their findings when applied to other agricultural environments. In the context of Indonesia, particularly in the Sumatra region, comprehensive comparative studies that evaluate multiple machine learning algorithms for rice production prediction using environmental and agronomic variables are still relatively limited. As a result, there is still uncertainty regarding which algorithm provides the most reliable predictive performance for rice production forecasting in this regional context.

To address this gap, this study aims to evaluate and compare the predictive capabilities of four regression-based machine learning models: Random Forest, XGBoost, SVR, and ANN in estimating rice production, using environmental variables and harvested land area as the primary inputs. The model's performance is assessed using several commonly used regression evaluation metrics, including Mean Absolute Error (MAE) [12-13], Root Mean Squared Error (RMSE) [14-15], and the coefficient of determination (R^2) [16]. This evaluation metric is widely applied in predictive modeling because it offers complementary perspectives on performance, including overall accuracy, the scale of prediction errors, and how effectively the model accounts for variations in the observed data.

This study makes three primary contributions. First, it delivers a structured comparative evaluation of several machine learning algorithms for predicting rice production, all assessed using the consistent dataset and a unified evaluation framework. Second, the study focuses on the agricultural context of the Sumatra region, which remains underexplored in previous machine learning-based agricultural prediction studies. Third, the research identifies the most influential environmental variables affecting rice production, providing valuable insights for agricultural stakeholders and policymakers. The results of this research are anticipated to support the advancement of more

dependable data-driven decision-support systems in agriculture and encourage the integration of machine learning approaches to strengthen food security efforts in Indonesia.

2. Methods

This research was carried out in a clear sequence of stages, starting from gathering the data, then cleaning and preparing it, followed by building the model, and finally assessing the results. The flow of these steps is shown in Figure 1.

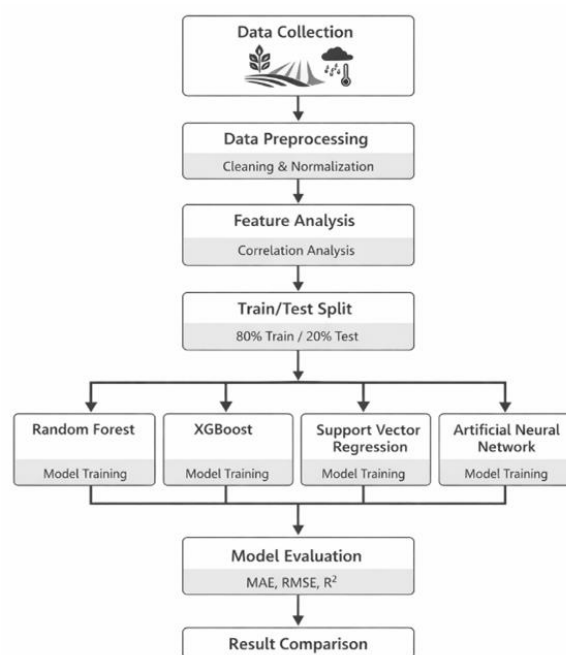


Figure 1. Research flow

The flow diagram shown in Figure 1 illustrates the methodological framework applied to build a machine learning model to predict rice production. This process begins with data collection, in which information regarding rice yields and key environmental factors such as harvested land area, rainfall, humidity, and average temperature is gathered from various available sources. Next, the dataset undergoes preprocessing procedures, including data cleaning, handling of missing value, and normalizations, to ensure its consistency and suitability for model development. Following this stage, feature analysis is conducted using correlation techniques to examine relationships between variables and identify those with the strongest influence on rice production.

The dataset was then divided into two groups, where 80% of the data was used to train the model, and the remaining 20% was set aside to test its performance. In the training process, four different machine learning methods were applied, namely Random Forest, XGBoost, SVR, and ANN.

All models were developed using the same dataset and the same experimental setup so that their prediction results could be compared fairly without any bias. After

training was complete, performance model was evaluation using several regression metric, including MAE, RMSE, and the coefficient of determination (R^2). These indicators provide insights into prediction accuracy and the extent to which each model can capture variability in the data. In the final stage, a comparative analysis is conducted to evaluate the performance of all algorithms and determine which model produces the most accurate rice production estimates. This structured approach allows for an objective assessment of various ML (machine learning) techniques and helps identify the most suitable method for rice production forecasting.

2.1 Dataset and Preprocessing

The dataset used includes rice production in Sumatra from various years with features such as harvested area, rainfall, humidity, and average temperature. Data preprocessing was carried out by addressing missing values, normalization, and correlation analysis between the obtained variables https://www.kaggle.com/datasets/ardikasatria/datasetta_namanpadisumatera.

In addition, the collected data must undergo a preprocessing step to remove irrelevant information, handle missing values, and convert it into a format suitable for analysis. The dataset was then divided into two parts: 179 observations (80%) were used as the training dataset, while 45 observations (20%) served as the test dataset. The models were built using the training dataset and later assessed using the test dataset to measure how well they perform in making predictions..

2.2 Machine Learning Algorithm

Four regression-based machine learning algorithms are implemented in this study to predict rice production. These algorithms were selected because they represent different methodological approaches within machine learning, including ensemble learning techniques, kernel-based models, and neural network architectures. Each method employs a unique strategy to recognize patterns and identify relationships between input features and desired outcomes. By utilizing various algorithms, this study can evaluate the performance of different modeling approaches when applied to agricultural production data.

By comparing these algorithms, the study can identify which model produces the most stable and accurate predictions of rice production. Evaluating different methods with varying learning mechanisms also helps reveal how well each model captures the complex relationships between environmental factors and agricultural yields. Through this research, the aim is to determine the most appropriate machine learning method for rice production forecasting in order to support better decision-making in data-driven agriculture..

The Random Forest method is an ensemble learning that builds decision trees and combines the results to

obtain more accurate predictions. This approach also helps reduce the likelihood of overfitting.[17]. In a regression context, the model generates its final output averaging the predictions produced by all individual trees. If there are N trees and each tree T_i yields a prediction $F_i(x)$ for a given input x , then the overall prediction is calculated as the mean of these outputs, as expressed in Equation 1

$$Y = \frac{1}{n} \sum_{i=1}^N F_i^x \quad (1)$$

In the framework of Random Forest regression, the final output is obtained by combining the predictions produced by many decision trees in the ensemble. In this context, y represents the overall predicted value generated by the model, obtained by averaging the outputs from all individual trees. The term N refers to the total number of decision trees included in this model, while the value $F_i(x)$ represents the prediction made by the i -th decision tree for input x . Thus, by aggregating the results from many decision trees, Random Forest increases predictive accuracy and reduces the likelihood of overfitting compared to relying on a single decision tree.

XGBoost Regression is a decision tree-based boosting algorithm designed to provide high performance in regression prediction and classification. XGBoost (Extreme Gradient Boosting) optimizes a model using a multilevel boosting technique to enhance prediction accuracy and computational efficiency. The prediction in XGBoost is calculated by adding the prediction results of each decision tree built. which is mathematically defined in Equation 2.

$$Y = \frac{1}{n} \sum_{m=1}^M F_m^{(x)} \quad (2)$$

In the XGBoost regression approach, the final output is produced by aggregating the contributions of multiple decision trees that are developed sequentially during training. Here, y denotes the predicted value generated by the model, while M represents the total number of trees involved in the boosting process. The expression $F_m(x)$ refers to the prediction provided by the m -th tree for a given input x . Unlike ensemble methods where trees are constructed independently, XGBoost builds each successive tree to address the errors made by the preceding ones. This step-by-step learning mechanism enables the model to continuously refine its predictions and enhance overall accuracy by reducing cumulative error

Support Vector Regression is a regression technique derived from Support Vector Machines with the aim of identifying the optimal function in predicting output values so that it can keep the prediction error within a predetermined tolerance limit. The goal is to strike a balance between model complexity and prediction accuracy by allowing for small deviations that remain within that limit. Mathematically, this formulation is expressed in Equation 3.

$$y = w^t x + b \quad (3)$$

In Support Vector Regression (SVR), the predicted value of (y) is obtained based on the relationship between two variables, namely the input and the learned model parameters. Here, x represents the input feature vector, while w is the weight vector that determines how much each feature influences the final prediction. The parameter b is the bias term, which shifts the regression function to better fit the data.

Additionally, SVR incorporates a tolerance parameter ϵ (epsilon), which defines a margin of acceptable error between predicted values and actual observations. This margin enables the model to disregard minor deviations that fall within the predefined boundary, while concentrating on reducing larger errors. By applying this tolerance mechanism, SVR achieves a balance between model complexity and predictive [17].

Artificial Neural Networks are a computational model inspired by the information processing mechanisms in the human brain. Composed of interconnected neurons arranged in several layers, this model is used to learn patterns and produce predictions from input data. Mathematically, this relationship can be expressed as Equation 4.

$$y = f(\sum_{i=1}^n w_i x_i + b) \quad (4)$$

In an Artificial Neural Network (ANN), the neuron's output, denoted as y , is obtained from a weighted combination of input signals that is subsequently processed through an activation function. In this formulation, x_i represents the input values supplied to the neuron, while w_i denotes the corresponding weights that regulate the influence of each input on the final output. The term bias is added to the sum of weights to adjust the activation threshold and increase the flexibility of the model and adapt to the data.

After calculating the total weight and applying the bias, the result is computed through the activation function f , which introduces non-linearity into the model. This non-linearity is the network's way of capturing complex relationships in data. The activation functions used include Rectified Linear Unit (ReLU), sigmoid, and hyperbolic tangent (Tanh). Through the combination of weighted inputs, bias adjustment, and non-linear transformation, the ANN model is able to learn complex patterns between input features and the predicted output. To evaluate the performance of a model, several regression metrics are used and applied to measure how close the predicted values are to the actual observations. One of them is the Mean Absolute Error (MAE), which calculates the average absolute difference between predicted values and actual values. This metric is easy to use because it is expressed in the same units as the target variable.

Thus, the Root Mean Squared Error (RMSE) is applied in measuring the performance of a model by taking the square root of the mean of the squared differences between the predicted values and the actual values. Since the errors are squared before averaging, this metric therefore gives more emphasis to larger errors,

making it useful for detecting models that produce significant prediction deviations.

The coefficient of determination (R^2) is used to evaluate how well a model performs and explains the variability in the observed data. A higher R^2 value means the model is better at identifying the underlying patterns in the dataset, resulting in more accurate and reliable predictions[18]

In an Artificial Neural Network (ANN), the output of a neuron is represented by Y , which is calculated based on the weighted combination of input values processed through an activation function. In this formulation, X_i denotes the input values provided to the neuron, while W_i represents the weights assigned to each corresponding input. These weights determine the relative importance of each input variable in influencing the final output of the neuron. The term B refers to the bias component, which is added to the weighted sum of inputs in order to adjust the activation threshold of the neuron.

After summing the weighted inputs and adding the bias term, the result is processed through an activation function represented as F . This function plays an important role in introducing non-linearity into the network, allowing the model to learn more complex patterns in the data. Common activation functions used in artificial neural networks include Rectified Linear Unit (ReLU), Sigmoid, and Hyperbolic Tangent (Tanh). By combining weighted inputs, bias adjustment, and non-linear activation functions, ANN is able to capture complex relationships within the data

3. Results and Discussions

The results of the experiment show a performance comparison among four machine learning algorithms in predicting rice production. The best models is determined based on the most optimal rmse, mae, and r^2 score. In addition, the most influential environmental factors in rice production prediction are analyzed using the shap method (shapley additive explanations) and this test uses python based on google colab

3.1 Data Exploration

Data exploration is the first stage of data analysis that focuses on understanding the structure and main characteristics of a dataset. This process involves carefully examining the data to find patterns, detect anomalies, and identify relationships between variables. Its main objective is to assess data quality by checking for issues such as missing values, duplicate records, and outliers. It also includes analyzing the distribution and range of each variable to ensure that the dataset is suitable for further analysis [19]. Common techniques in data exploration include descriptive statistics, which can summarize data using measures such as mean, median, mode, and standard deviation. In addition, data visualisation methods—such as histograms, box plots and scatter plots—are used to illustrate distributions and

reveal relationships between variables in a more intuitive way [20].

In addition correlation analysis is carried out to measure the strength and direction of the relationship between two or more variables, thereby helping to identify whether there is a significant relationship. The benefits of data exploration are significant, as it helps in making better decisions by providing initial insights into the data, making it easier to formulate hypotheses for further analysis, and reducing the risk of errors in more complex analysis models. Overall, Data exploration is a key step in the data analysis process as it helps researchers and analysts better understand the dataset before proceeding to more advanced analysis. This stage provides early findings, including those illustrated in Figure 2.

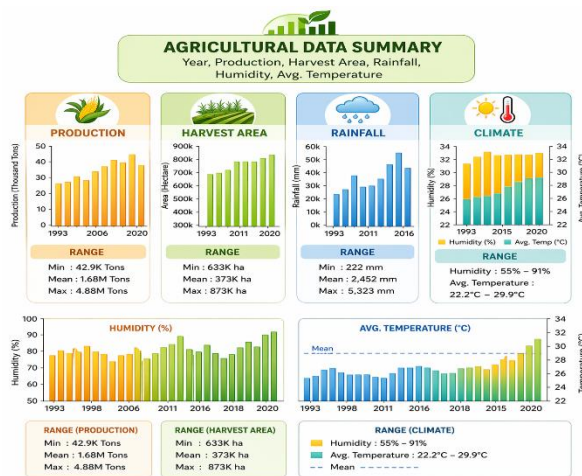


Figure 2. Data exploration

As shown in Figure 2, once the dataframe is properly loaded, we can analyze the correlation between the numerical features and the target variable "Production".

3.2. Data preparation

Data preparation is a key stage in the analytical process that involves cleaning and organizing data before it is used for further analysis. The purpose of this stage is to ensure that the dataset is accurate, consistent, and suitable for modeling. This process begins with collecting data from relevant internal and external sources that are aligned with the research objectives[21].

Subsequently, data cleaning is performed to detect and resolve issues such as missing values, duplicate entries, and outliers, all of which can significantly affect the validity and reliability of the results. In addition, data transformation is carried out to convert the dataset into an appropriate format for analysis. This may include normalization, standardization, or encoding categorical variables into numerical representations, enabling more effective application of analytical and machine learning techniques [22].

3.3 Data Splitting

Data splitting technique is a fundamental step in data analysis, especially in developing machine learning models. This method involves dividing a dataset into several separate parts, usually a training dataset and a testing dataset[23]. The training dataset is used to develop a model with an algorithm to learn patterns and relationships in data. In contrast, the testing dataset is provided for evaluation purposes, which can measure how well the trained model can generalize and make predictions on data that has not been seen before[24].

Data splitting is an important step to prevent overfitting, where a model performs well on training data but fails to generalize to new data. Generally, the dataset is divided into training and testing sets, such as 70% for training and 30% for testing, although the exact proportion can vary depending on the size and characteristics of the dataset. Proper data splitting helps ensure that the final model maintains good predictive performance and reliability in practical use. In this study, the dataset is divided into training and testing subsets to evaluate the performance of machine learning models. The training set contains 179 observations with 24 features, represented as $X_{train}(179,24)$, meaning that 179 samples are used to train the models using 24 predictor variables. The corresponding target values are represented as $y_{train}(179,)$, indicating that there are 179 output values linked to the training data.

Meanwhile, the testing set includes 45 observations with the same 24 features, represented as $X_{test}(45,24)$. This subset is used to assess model performance on unseen data. The corresponding target values are denoted as $y_{test}(45,)$, indicating that 45 actual observations of rice production are used for evaluation. This clear separation between training and testing data enables the models to be assessed objectively, ensuring their ability to generalize and produce reliable predictions when applied to new datasets.

3.4 Training model

Model training is an essential stage in developing machine learning models, where the algorithm learns to recognize patterns and relationships within the data. During this process, the model uses the training dataset to understand the connection between input features and the target variable, while continuously adjusting its internal parameters to generate predictions that closely match the actual observed values[25].

The training phase also encompasses several essential steps, including choosing a suitable algorithm, tuning model parameters, and applying optimization strategies to enhance performance. Appropriate parameter configuration enables the model to learn more efficiently from the data, while optimization techniques work to minimize prediction errors and improve generalization to unseen cases. Consequently, a well-trained model is capable of producing more accurate and dependable predictions [26].

Based on Table 1 in the data exploration stage, the Random Forest Regression model shows strong predictive performance in estimating rice production. The model achieved a Mean Absolute Error (MAE) value of 128,348, which indicates that the average difference between the predicted values and the actual values is relatively very small. This indicates that, overall, the model's predictions are quite close to the actual data, reflecting a very high level of consistency across the dataset. Practically, a low MAE means minimal deviation in each prediction, making this model reliable for general forecasting.

In addition, the model records a Root Mean Squared Error (RMSE) of 182,779, which is more sensitive to larger errors because it gives greater weight to bigger deviations. The fact that RMSE is slightly higher than MAE indicates that a small number of predictions have larger errors. However, these deviations are not frequent enough to significantly affect the overall results, so the error level remains acceptable. This also shows that while the model generally performs well, a few extreme conditions or unobserved factors may still cause larger prediction errors. The coefficient of determination (R^2) of 0.9622 indicates that the model can explain 96.22% of the variation in rice production, meaning that input variables such as harvested area and environmental factors are strongly influential and well captured by the Random Forest model. A high R^2 value also reflects a good model fit, showing that the main patterns in the data are well represented.

Overall, these findings confirm that the Random Forest Regression method provides very high prediction accuracy and stable performance on the dataset. Its ensemble structure, which combines multiple decision trees, enhances generalization and reduces the risk of overfitting. Therefore, this model is highly suitable for predicting rice crop yields and can serve as a reliable tool in data-driven decision making in agriculture. The comparison between the actual values and the values predicted by the Random Forest model is shown in Figure 3, further highlighting its effectiveness.

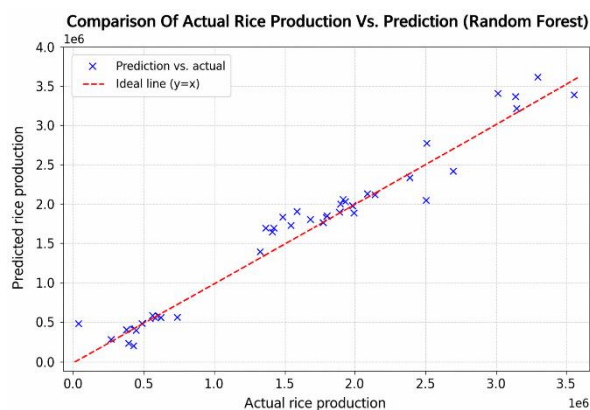


Figure 3. visualization of comparison between actual vs predicted rice production

Figure 3 shows that the blue points represent the relationship between the values predicted by the model

and the actual data, while the red dashed line ($y=x$) indicates the ideal condition where the predictions fully match the actual values. The proximity the data points to this reference line reflects the model's accuracy, with a tighter cluster of points around the line indicating better predictive performance. This visualisation demonstrates that the model is working effectively, as a large number of points align closely with the ideal line, indicating high accuracy in estimating rice production.

Figure 4 presents a correlation matrix illustrating the relationships between variables in the dataset.

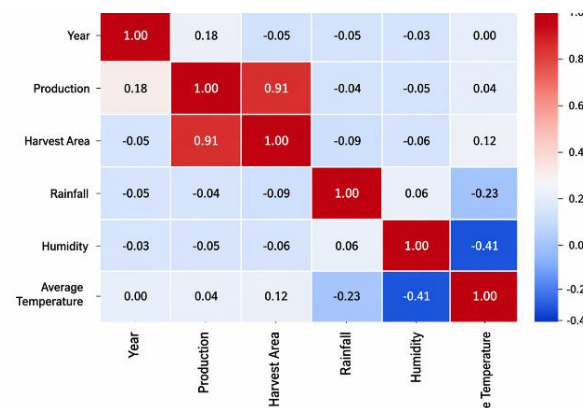


Figure 4. Correlation Matrix between variables

Figure 4 presents a heatmap of the correlation matrix showing the relationships between variables in the dataset, namely Year, Production, Harvested Area, Rainfall, Humidity, and Average Temperature. In this visualization, the red color represents a positive correlation, while the blue color indicates a negative correlation, with a darker color intensity signifying a stronger relationship between the variables.

The results reveal that Harvest Area has a very strong positive correlation with Production (0.91), indicating that an increase in harvested land is closely associated with higher rice production. In contrast, Humidity shows a moderately strong negative correlation with Average Temperature (-0.41), suggesting that humidity tends to increase when average temperature decreases.

Other variables such as Rainfall, Year, and several inter-variable relationships exhibit relatively weak correlations, implying that their impact on production is less significant compared to the primary factor. Overall, correlation coefficients range from -1 to 1, representing both the direction and strength of relationships. This analysis highlights that cultivated land area is the most influential variable affecting rice production.

Furthermore, Figure 5 presents the structure of a decision tree from one of the estimators within the Random Forest Regression model.

Figure 5 explains that each node in the decision tree represents a point where data is split based on the most important feature at that stage. The branches illustrate the different decision paths that eventually lead to the final outcomes at the leaf nodes. The variations in color within the nodes reflect how the data samples are

distributed and how dense they are in each segment. As part of the Random Forest approach, this tree is just one of many estimators, where each tree can produce its own prediction and the final result is calculated based on the average of all tree outputs to improve accuracy

and reduce the risk of overfitting. Meanwhile, XGBRegressor is a regression method based on the Extreme Gradient Boosting (XGBoost) algorithm, which is designed to deliver efficient and accurate predictions for regression problems.

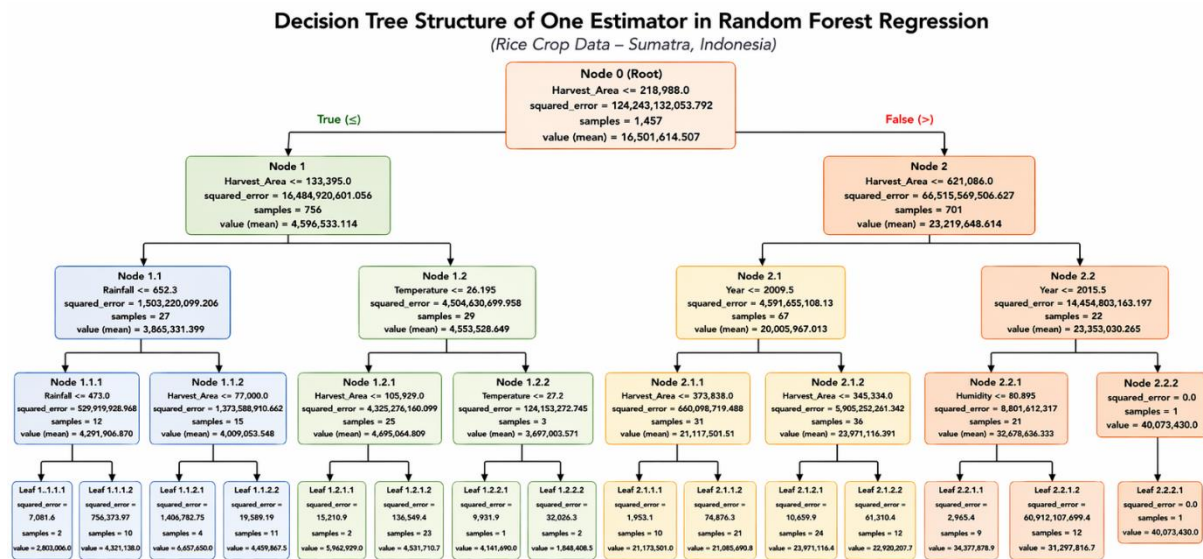


Figure 5. Tree Structure

This model provides several tunable parameters that can be adjusted to improve performance. Key parameters include `n_estimators`, which defines the number of trees used in the model; `learning_rate`, which controls how much the model updates itself in each iteration to minimize errors; and `max_depth`, which limits how deep each tree can grow in order to control model complexity and prevent overfitting.

Additionally, parameters such as `colsample_bytree` and `subsample` determine the proportion of features and data samples used during each training iteration, helping to reduce variance and improve generalization. The `gamma` parameter serves as a threshold for node splitting, ensuring that only meaningful splits are made, while `random_state` guarantees reproducibility of results across different runs. Owing to its flexibility, XGBRegressor is well-suited for handling complex datasets, including those with high dimensionality or missing values. To further improve the performance of a model, hyperparameter optimization techniques such as Grid Search or Random Search can be applied in identifying the most effective configuration of a parameter, as illustrated in Figure 6.

Figure 6 shows that the XGBoost regression model performs well in predicting rice production. It achieves an R^2 value of 0.958, meaning that about 95.8% of the variation in rice production can be explained by the model. This high value indicates that the model is effective in learning the relationship between the input variables and the target output.

From the error evaluation perspective, the model obtained a Mean Squared Error (MSE) value of 36,994,924,266.53, which represents the average

squared difference between predicted values and actual values. In addition, the Root Mean Squared Error (RMSE) was recorded at 192,340.65, indicating the magnitude of prediction errors in the same units as the target variable. Although there is still a gap between the predicted results and the actual values, the relatively low error value along with a high R^2 indicates that the model has a good level of accuracy and predictive capability.

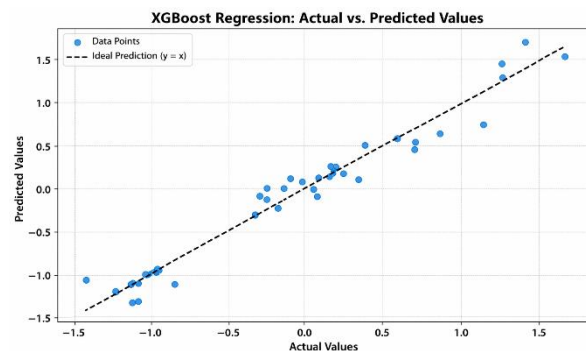


Figure 6. XGBRegressor Actual vs Predicted Values.

Further analysis shows that 'Production' is the target variable, while 'Harvested Area' has the strongest positive correlation (0.905622), indicating that this variable has a significant influence on rice production results. In addition, the data is divided into training and testing sets with an 80:20 ratio, and using a `random_state` value of 42 so that the results obtained remain consistent and reproducible. With this setup, the XGBoost model consistently delivers strong performance on the test data.

Despite these strong results, several challenges remain, particularly in feature engineering, where difficulties

were encountered when attempting to create and incorporate additional variables. Enhancing this stage could further improve model performance by introducing more informative features. Additionally, optimizing hyperparameters through techniques such as grid search or random search is recommended to refine the model and achieve even better predictive accuracy. A scatter plot illustrating the relationship between harvested area and rice production is presented in Figure 7.

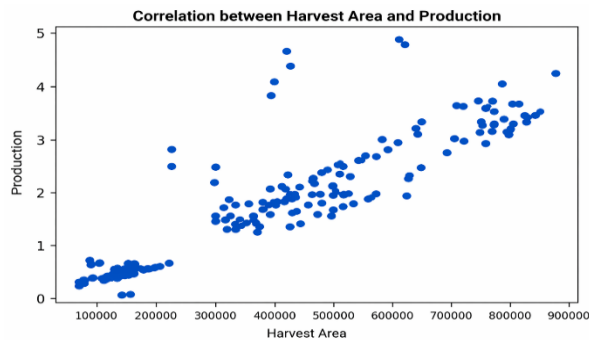


Figure 7. Visualization with Scatter Plot

The scatter plot presented in Figure 7 demonstrates a clear positive association between harvested area and rice production. The distribution of data points forms an upward trend, indicating that larger harvested areas are generally accompanied by higher levels of rice output. Although a few observations deviate from this pattern, the overall relationship remains consistently positive and approximately linear. This pattern suggests that harvested area plays a significant role in influencing rice production levels. In other words, expansion of cultivated land tends to contribute directly to increased agricultural yield. Despite the presence of outliers, the dominant trend reinforces the importance of harvested area as a key explanatory variable in rice production analysis.

A histogram helps to show how a numerical variable is distributed across its range. In this study, Figure 8 presents the distribution pattern of rice production values.

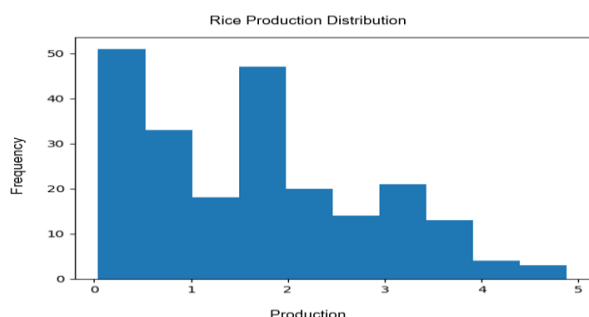


Figure 8. shows a histogram that illustrates the distribution of production values:

The histogram in Figure 8 shows the distribution of Rice Paddy Production based on the number of frequencies. The distribution of the data appears skewed to the right, where most regions have smaller

amounts of rice production, while only a few regions have very high production. This suggests that the majority of regions have lower production capacity, while only a few regions have large amounts of rice production. This pattern could be an indicator of factors that affect production, such as land area, agricultural technology, or climatic conditions.

A box plot showing the distribution of production by province is shown in Figure 9. Boxplots in Figure 9 show the distribution of Rice Production across provinces. Each boxplot depicts the distribution of the data, including the median (green line in the middle), lower and upper quartiles (box borders), as well as outliers (dots outside the whiskers).

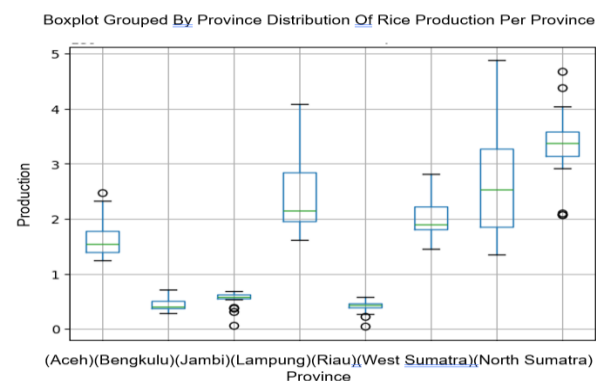


Figure 9. Box Plot Visualization

From Figure 9, it can be seen that some provinces have high production variability, such as North Sumatra and South Sumatra, which show a wider distribution of data and the presence of outliers. Meanwhile, other provinces such as Bengkulu and Jambi have a lower and more uniform distribution of production.

This pattern indicates that there are significant differences in rice production capacity between provinces, which may be due to factors such as farm size, cultivation techniques, or local climatic conditions.

The evaluation results show that the Support Vector Regression (SVR) model has weak performance in predicting rice production. This is evidenced by a relatively high Mean Absolute Error (MAE) of 801,088.64, indicating that, on average, there is a large gap between the predicted and actual values. Such a high level of error reflects low accuracy, making the model less reliable for practical prediction purposes.

In addition, the Root Mean Squared Error reaches 969,013.58, highlighting the presence of large prediction deviations that are heavily penalized in this metric. This confirms that the model not only makes frequent errors but also occasionally produces significantly large misestimations.

More importantly, the coefficient of determination (R^2) is -0.062 , which indicates that the model is unable to explain the variation in the rice production data. A negative R^2 value means that the performance of the

SVR model is even worse compared to a simple baseline method that only predicts the average value of the data. This indicates that the model does not fit the data and fails to effectively capture the patterns within it.

This poor performance is further evident in the scatter plot comparing actual and predicted values Figure 10. In an ideal scenario, the data points would closely follow the diagonal line, which represents perfect prediction accuracy. However, the substantial spread of points away from this line indicates a weak alignment between the predicted and actual values. This pattern reinforces the finding that the SVR model is not reliable for this dataset in its current form. Improvements such as hyperparameter tuning, better feature scaling, or more effective feature selection may be required to enhance its performance

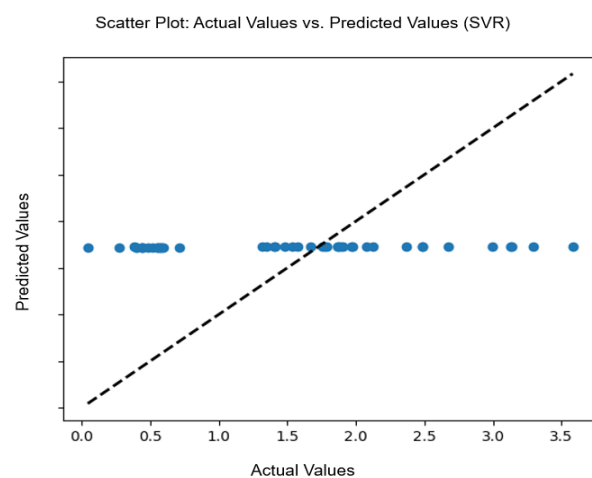


Figure 10. Scatter Plot: Actual Value vs Value

Figure 10 shows that the residual plot is useful for checking whether the assumptions of the SVR model are met. Ideally, the residuals should be randomly distributed around the horizontal zero line, as shown in Figure 11.

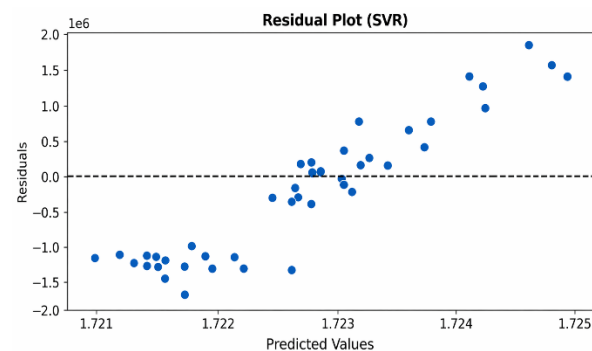


Figure 11. Residual plot

The importance of this visualization, which shows which features have the greatest influence on the SVR model's predictions, can be seen in Figure 12.

Artificial Neural Network (ANN) models were evaluated by dividing the dataset into training and testing sets to see how well they perform on new,

unseen data. Their performance was measured using MAE, RMSE, and R^2 , which indicate the size of prediction errors and how well the model explains variations in the data. To support the analysis, scatter plots and residual plots were also used to compare predicted and actual values and to examine error patterns more clearly.

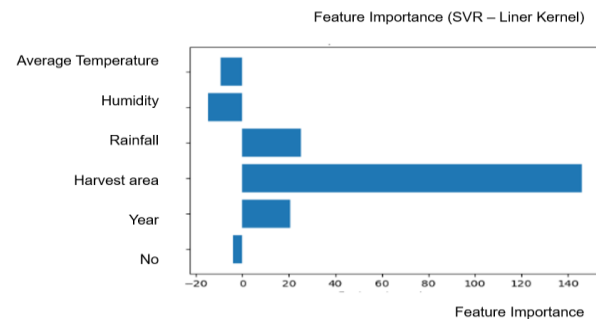


Figure 12. Feature Importance

The results indicate that the ANN model has poor performance in predicting rice production, with an MAE of 1,459,396.38 and an RMSE of 1,735,320.59. The R^2 value of -2.406 shows that the model was unable to learn meaningful patterns from the dataset and even performed worse than a simple mean-based prediction. This weak outcome is likely due to the limited amount of data, which was not enough to properly train the neural network. Consequently, the model exhibits low generalization ability when compared to other machine learning approaches

3.5 Comparison of Actual vs Predicted Rice Production (ANN)

Figure 13 presents a visualization that compares the real rice production values with the outputs estimated by the ANN model

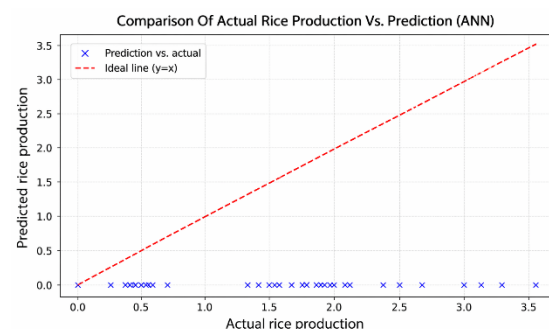


Figure 13. Comparison of Actual vs Predicted Rice Production

Figure 13 presents a comparison between the actual rice production values and those predicted by the Artificial Neural Network (ANN) model. The x-axis shows the observed data, while the y-axis displays the model's predicted results. Each blue point represents the relationship between actual and predicted values. In a well-performing model, these points would be closely aligned with the red dashed diagonal line ($y = x$), which reflects perfect prediction accuracy where both values are equal. However, the figure shows a clear deviation from this ideal pattern, as most data points are scattered

away from the diagonal line. In many cases, the predicted values are lower than the actual production figures. This indicates that the ANN model does not capture the data patterns effectively and may suffer from underfitting. Overall, this suggests that the model has limited capability in learning the underlying structure of the dataset, as also supported by the results shown in Figure 14.

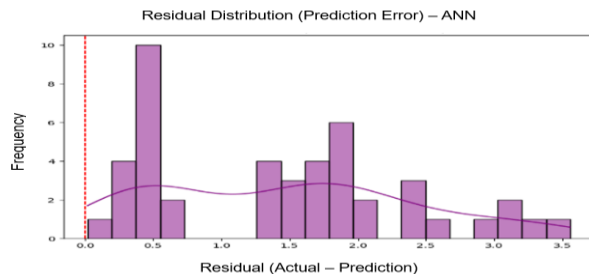


Figure 14. Residual Distribution (Prediction Error) - ANN

Figure 14 illustrates the distribution of prediction errors (residuals) produced by the Artificial Neural Network (ANN) model in estimating rice production. These residuals, defined as the difference between actual and predicted values (Actual – Predicted), are shown on the horizontal axis, while the vertical axis indicates the frequency of each error range. The purple histogram bars represent how often different residual values occur, and the smooth purple curve shows the overall trend of the error distribution. The red dashed line at zero marks the ideal point where predictions perfectly match the actual values, meaning no error occurs.

The graph indicates that most residuals fall on the positive side, implying that the model generally produces lower predictions than the actual rice production values (systematic underestimation). Moreover, the wide spread of residuals reflects a high level of prediction error variability. This pattern suggests that the ANN model has limited accuracy and does not consistently capture the underlying structure of the data, indicating suboptimal predictive performance as further confirmed in Figure 15.

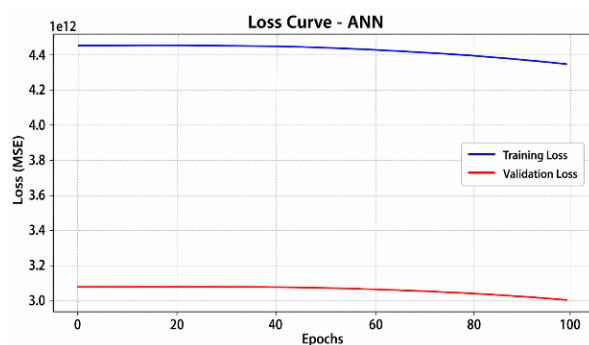


Figure 15. Loss Curve - ANN

Figure 15 shows the loss graph of the Artificial Neural Network (ANN) model, depicting changes in the error rate during the training process. The horizontal axis represents the number of epochs, while the vertical axis represents the loss values calculated using Mean

Squared Error (MSE). The blue line illustrates the loss on the training data, while the red line shows the loss on the validation data. The results indicate that both training and validation loss experienced very slight decreases during the training process, suggesting that the model is less capable of learning the data patterns optimally. This issue may be caused by an unsuitable model setup, a network structure that is too simple, or sub-optimal tuning of parameter. Furthermore, the persistently high loss values indicate that the model has not yet achieved satisfactory prediction accuracy. Therefore, further improvements such as hyperparameter optimisation or the use of a more suitable model architecture are required, as illustrated further in Figure 16.

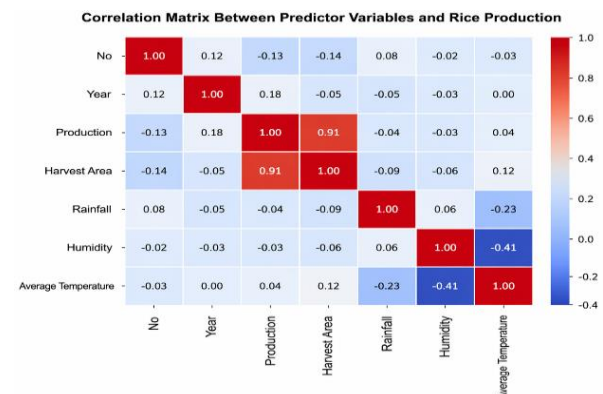


Figure 16. Correlation Matrix between Predictor Variables and Rice Production

Figure 16 presents a correlation matrix illustrating the relationships between the input variables and rice production. The correlation coefficients range from -1 to 1 , where values approaching 1 indicate a strong positive association, values approaching -1 indicate a strong negative association, and values close to 0 reflect a weak or negligible relationship. From the matrix, Harvested Area has the strongest positive correlation with rice production, with a value of 0.91 . This indicates that larger harvested areas are strongly associated with higher levels of rice production. Conversely, variables such as rainfall, humidity, and average temperature show weak or negative correlations, indicating that climatic factors may have a limited direct influence or that their relationship with production is more complex and indirect. The red color intensity represents stronger positive correlations, while blue tones indicate weaker or negative relationships. This analysis provides important insights into the key determinants of rice production and serves as a foundation for developing more accurate predictive models. Furthermore, the performance comparison of four regression algorithms used for rice production prediction is summarized, and the overall results of these model evaluations are presented in Table 1.

Table 1 shows a comparison of the performance of four regression models in predicting rice production. From the results, Random Forest is the model with the best performance, indicated by an R^2 value of 0.963 , which

means this model is able to explain approximately 96.3% of the data variation. In addition, this model also has a relatively low error rate, with an MAE of 125,524 and an RMSE of 180,136, making its predictions closer to the actual values compared to the other models. Meanwhile, XGBoost also delivers very good results with an R^2 of 0.959. However, its RMSE value is slightly higher than that of Random Forest, indicating a slight increase in the prediction error rate. Overall, both models demonstrate strong performance, although Random Forest still holds a slight edge in terms of accuracy and error control.

Table 1. Evaluation results

index	MAE	MSE	RMSE	R2
Random Forest	125.524	324.490	180136	0.963
XGBoost	115.389	360.613	189898	0.959
SVR	801.955	941.181	970145	-0.064
ANN	146.2767	302.153	1738256	-2.417

In contrast, both Support Vector Regression (SVR) and Artificial Neural Networks (ANNs) show unsatisfactory predictive results. The SVR model records a negative R^2 value of -0.064 , which means its predictions are less reliable than a simple approach that uses the average value of the dataset. The relatively high MAE and RMSE values also highlight a large gap between predicted and actual outcomes. This may be linked to unsuitable parameter settings or a poor match between the model assumptions and the structure of the data. Similarly, the ANN model performs even worse, with an R^2 of -2.417 . This suggests that the model is unable to capture meaningful relationships within the dataset. The very large error values further indicate potential problems such as limited training data, an unsuitable network structure, or insufficient training processes. As a result, the model shows weak generalization ability and fails to produce reliable predictions.

Overall, these findings suggest that ensemble-based methods such as Random Forest and XGBoost are more suitable for predicting rice production in this study. Conversely, SVR and ANN require further optimisation, including hyperparameter tuning, feature engineering, or architectural adjustments such as adding hidden layers, adjusting the learning rate, and expanding the dataset to improve learning performance, as illustrated in Figure 17.

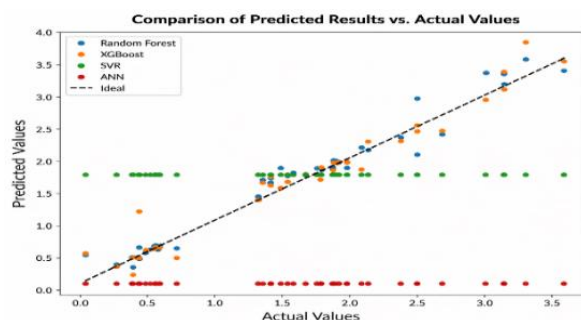


Figure 17. Comparison results of all algorithms

4. Conclusions

In conclusion, the comparative assessment of the four regression models demonstrates that Random Forest and XGBoost provide the most reliable and accurate performance for rice paddy production forecasting, achieving R^2 values of 0.963 and 0.959, respectively. These findings highlight the strong capability of both ensemble-based models to capture the inherent variability and complex relationships within the dataset, thereby establishing them as robust approaches for rice yield prediction. By contrast, SVR and ANN exhibited substantially weaker predictive performance, as indicated by their negative R^2 values, suggesting that these models were unable to adequately learn the underlying relationship between the input variables and the target output under the present experimental setting. Therefore, Random Forest and XGBoost are recommended as the preferred models for rice production forecasting, whereas further refinement of SVR and ANN through comprehensive hyperparameter optimization, improved model architecture, and more effective feature engineering is necessary to enhance their predictive reliability in future studies.

References

- [1] S. Singh, M. Singh, and S. Attri, "Mapping stock market dynamics: A tripartite neural network approach using modified grid search for stock market prediction," *Expert Syst. Appl.*, vol. 278, p. 127243, Jun. 2025, doi: [10.1016/j.eswa.2025.127243](https://doi.org/10.1016/j.eswa.2025.127243).
- [2] A. Alkan, "Analysis of Factors Affecting Academic Success with Machine Learning: Data-Driven Inferences in Education," *Turkish Journal of Engineering-2026*, vol. 10, no. 1, pp. 48–62, 2026, doi: [10.31127/10.31127/tuje.1779491](https://doi.org/10.31127/10.31127/tuje.1779491).
- [3] A. Chotirattanachote, N. Thiensuwan, C. Ngamcharussrivichai, T. Yokoi, and S. Sukpancharoen, "Machine learning-guided catalyst selection for bio-oil hydrodeoxygenation to maximize hydrocarbon yields toward renewable naphtha and sustainable aviation fuel," *Fuel*, vol. 426, Dec. 2026, doi: [10.1016/j.fuel.2026.139511](https://doi.org/10.1016/j.fuel.2026.139511).
- [4] Y. Xia, "Statistical properties of a class of randomized binary search algorithms," *Performance Evaluation*, vol. 168, p. 102478, Jun. 2025, doi: [10.1016/j.peva.2025.102478](https://doi.org/10.1016/j.peva.2025.102478).
- [5] A. I. Simsek, E. Bulut, Y. E. Gur, and E. Gültekin Tarla, "A novel approach to Predict WTI crude spot oil price: LSTM-based feature extraction with Xgboost Regressor," *Energy*, vol. 309, p. 133102, Nov. 2024, doi: [10.1016/j.energy.2024.133102](https://doi.org/10.1016/j.energy.2024.133102).
- [6] J. Zhang, F. Xie, and X.-J. Liu, "Dynamic modeling of hybrid robots through stiffness directionally releasing and multiple-node dynamics assembling," *Mech. Mach. Theory*, vol. 209, p. 106002, Jul. 2025, doi: [10.1016/j.mechmachtheory.2025.106002](https://doi.org/10.1016/j.mechmachtheory.2025.106002).
- [7] Z. Liu, Z. Zhang, Q. Zhang, and L. Zhao, "Prediction of landfill gases concentration based on Grey Wolf Optimization – Support Vector Regression during landfill excavation process," *Waste Management*, vol. 198, pp. 128–136, May 2025, doi: [10.1016/j.wasman.2025.02.040](https://doi.org/10.1016/j.wasman.2025.02.040).
- [8] N. Deng, R. Xu, Y. Zhang, H. Wang, C. Chen, and H. Wang, "Forest biomass carbon stock estimates via a novel approach: K-nearest neighbor-based weighted least squares multiple birth support vector regression coupled with whale optimization algorithm," *Comput. Electron. Agric.*, vol. 232, p. 110020, May 2025, doi: [10.1016/j.compag.2025.110020](https://doi.org/10.1016/j.compag.2025.110020).
- [9] Y. Liu, X. Liu, Y. Zhou, and P. Cao, "S-ANN: Synchronous TCAD device simulation of FinFET using Artificial Neural

- Network,” *Microelectronics J.*, vol. 159, p. 106630, May 2025, doi: [10.1016/J.MEJO.2025.106630](https://doi.org/10.1016/J.MEJO.2025.106630).
- [10] X. Chen et al., “Modelling and prediction of atmospheric drag coefficients in LEO satellite orbit determination and prediction with Bi-LSTM approach,” *Advances in Space Research*, vol. 75, no. 3, pp. 2874–2888, Feb. 2025, doi: [10.1016/J.ASR.2024.10.063](https://doi.org/10.1016/J.ASR.2024.10.063).
- [11] S. R. Andani, Poningsih, and A. Karim, “Enhancing Lung Cancer Detection: Optimizing CNN Architectures through Hyperparameter Tuning,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 4, pp. 944–954, Aug. 2025, doi: [10.29207/RESTI.V9I4.6357](https://doi.org/10.29207/RESTI.V9I4.6357).
- [12] J. Michańkó, P. Sakowski, and R. Ślepaczuk, “Mean Absolute Directional Loss as a new loss function for machine learning problems in algorithmic investment strategies,” *J. Comput. Sci.*, vol. 81, p. 102375, Sep. 2024, doi: [10.1016/J.JOCS.2024.102375](https://doi.org/10.1016/J.JOCS.2024.102375).
- [13] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, “Mean Absolute Percentage Error for regression models,” *Neurocomputing*, vol. 192, pp. 38–48, Jun. 2016, doi: [10.1016/J.NEUCOM.2015.12.114](https://doi.org/10.1016/J.NEUCOM.2015.12.114).
- [14] A. Langenbucher, N. Szentmáry, J. Wendelstein, A. Cayless, P. Hoffmann, and D. Gatineau, “Performance Evaluation of a Simple Strategy to Optimize Formula Constants for Zero Mean or Minimal Standard Deviation or Root-Mean-Squared Prediction Error in Intraocular Lens Power Calculation,” *Am. J. Ophthalmol.*, vol. 269, pp. 282–292, Jan. 2025, doi: [10.1016/J.AJO.2024.08.043](https://doi.org/10.1016/J.AJO.2024.08.043).
- [15] A. Langenbucher, N. Szentmáry, J. Wendelstein, A. Cayless, P. Hoffmann, and D. Gatineau, “Performance Evaluation of a Simple Strategy to Optimize Formula Constants for Zero Mean or Minimal Standard Deviation or Root-Mean-Squared Prediction Error in Intraocular Lens Power Calculation,” *Am. J. Ophthalmol.*, vol. 269, pp. 282–292, Jan. 2025, doi: [10.1016/J.AJO.2024.08.043](https://doi.org/10.1016/J.AJO.2024.08.043).
- [16] T. Kenarangi, F. Rahmani, A. Yazdani, G. D. Ahmadi, M. Lotfi, and T. A. Khalaj, “Comparison of GAP, R-GAP, and new trauma score (NTS) systems in predicting mortality of traffic accidents that injure hospitals at Mashhad University of medical sciences,” *Heliyon*, vol. 10, no. 16, p. e36004, Aug. 2024, doi: [10.1016/J.HELIYON.2024.E36004](https://doi.org/10.1016/J.HELIYON.2024.E36004).
- [17] J. Huang et al., “Support Vector Regression for Adaptive Endpoint Detection in Automated Potentiometric Titration: A Methodological Integration of Machine Learning and Process Control,” *Results in Control and Optimization*, p. 100698, Mar. 2026, doi: [10.1016/J.RICO.2026.100698](https://doi.org/10.1016/J.RICO.2026.100698).
- [18] T. M. Samo, C. Ominde, J. Maithya, and J. Muniyitha, “Artificial Neural Network (ANN) modeling for CO2 concentration prediction in geothermal fields,” *J. Atmos. Sol. Terr. Phys.*, vol. 279, p. 106735, Feb. 2026, doi: [10.1016/J.JASTP.2026.106735](https://doi.org/10.1016/J.JASTP.2026.106735).
- [19] N. Tylosky, A. Knutas, and A. Wolff, “Design practices in visualization driven data exploration for non-expert audiences,” *Comput. Sci. Rev.*, vol. 56, p. 100731, May 2025, doi: [10.1016/J.COSREV.2025.100731](https://doi.org/10.1016/J.COSREV.2025.100731).
- [20] N. T. N. Tien, N. Q. Thu, D. H. Kim, S. Park, and N. P. Long, “EasyPubPlot: A Shiny Web Application for Rapid Omics Data Exploration and Visualization,” *J. Proteome Res.*, Mar. 2025, doi: [10.1021/ACS.JPROTEOME.4C01068](https://doi.org/10.1021/ACS.JPROTEOME.4C01068).
- [21] Z. J. Salmi, M. I. Khodakarami, and F. Behnamfar, “Reducing data preparation calculations for estimating machine learning-based seismic fragility curves of structures,” *Journal of Building Engineering*, vol. 102, p. 111893, May 2025, doi: [10.1016/J.JOBE.2025.111893](https://doi.org/10.1016/J.JOBE.2025.111893).
- [22] S. Boonpan and W. Sarakorn, “Deep neural network model enhanced with data preparation for the directional predictability of multi-stock returns,” *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 11, no. 1, p. 100438, Mar. 2025, doi: [10.1016/J.JOITMC.2024.100438](https://doi.org/10.1016/J.JOITMC.2024.100438).
- [23] P. C. Li, S. Dey, and V. Merwade, “Analyzing the effects of data splitting and covariate shift on machine learning based streamflow prediction in ungauged basins,” *J. Hydrol. (Amst.)*, vol. 653, p. 132731, Jun. 2025, doi: [10.1016/J.JHYDROL.2025.132731](https://doi.org/10.1016/J.JHYDROL.2025.132731).
- [24] S. A. Memon, I. Ullah, G. E. M. Abro, I. Ullah, and A. Noor, “Advanced data association technique using integrated track splitting filter for multi-target tracking in clutter and occlusion,” *Expert Syst. Appl.*, vol. 277, p. 127193, Jun. 2025, doi: [10.1016/J.ESWA.2025.127193](https://doi.org/10.1016/J.ESWA.2025.127193).
- [25] A. Mukherjee, D. Gupta, and D. Bhattacharyya, “Mass-Constrained hybrid Gaussian radial basis neural networks: Development, training, and applications to modeling nonlinear dynamic noisy chemical processes,” *Comput. Chem. Eng.*, vol. 197, p. 109080, Jun. 2025, doi: [10.1016/J.COMPCHEMENG.2025.109080](https://doi.org/10.1016/J.COMPCHEMENG.2025.109080).
- [26] Y. Wang, “Personalized lightweight federated learning for efficient and private model training in heterogeneous data environments,” *Systems and Soft Computing*, vol. 7, p. 200212, Dec. 2025, doi: [10.1016/J.SASC.2025.200212](https://doi.org/10.1016/J.SASC.2025.200212).