Crop Yield Prediction Using Random Forest Based on Soil, Climate, and Agronomic Factors

Putu Sugiartawan*[1], I Nyoman Darma Kotama[2], Anak Agung Surya Pradhana[3]
[1]Department of Information and Communication Systems, Okayama University, Okayama 700-8530, Japan
[2,3]Graduate School of Natural Science and Technology, Okayama University, Okayama 700-8530, Japan
e-mail: ***[1]p18z9yov@s.okayama-u.ac.jp**, [2]p9363bg2@s.okayama-u.ac.jp, [3]p44c722@s.okayama-u.ac.jp

Abstrak

Prediksi hasil pertanian memainkan peran penting dalam memastikan ketahanan pangan dan mengoptimalkan praktik pertanian. Metode tradisional untuk estimasi hasil pertanian sering mengandalkan pengetahuan ahli dan data historis, yang seringkali terbatas dan tidak akurat. Algoritma pembelajaran mesin, khususnya Random Forest, telah menunjukkan potensi untuk meningkatkan akurasi prediksi hasil pertanian dengan mempertimbangkan interaksi kompleks antara faktor tanah, iklim, dan agronomi. Penelitian ini bertujuan untuk mengembangkan model berbasis Random Forest untuk memprediksi hasil pertanian menggunakan dataset pertanian yang beragam. Model ini dilatih dan divalidasi menggunakan data dari berbagai wilayah, dengan fokus pada sifat tanah, kondisi iklim, dan praktik pertanian. Hasil penelitian menunjukkan bahwa model Random Forest memberikan prediksi yang dapat diandalkan, dengan evaluasi menggunakan metrik seperti MAE, RMSE, dan R². Namun, beberapa perbedaan antara nilai aktual dan yang diprediksi masih terlihat, yang menunjukkan adanya ruang untuk perbaikan. Penelitian selanjutnya akan berfokus pada integrasi data waktu nyata, seperti kelembaban tanah dan infestasi hama, untuk meningkatkan akurasi model. Selain itu, eksplorasi teknik pembelajaran mesin lanjutan seperti deep learning dapat memberikan penanganan yang lebih baik terhadap pola kompleks dalam data pertanian. Penelitian ini berkontribusi pada perkembangan ilmu data pertanian dan bertujuan memberikan solusi yang dapat diskalakan untuk prediksi hasil pertanian di berbagai wilayah.

Kata kunci: *Prediksi hasil pertanian, Random Forest, pembelajaran mesin, data pertanian, peramalan.*

Abstract

Agricultural yield prediction plays a critical role in ensuring food security and optimizing farming practices. Traditional methods of crop yield estimation often rely on expert knowledge and historical data, which can be limited and inaccurate. Machine learning algorithms, particularly Random Forest, have shown promise in improving the accuracy of crop yield predictions by considering complex interactions between soil, climate, and agronomic factors. This study aims to develop a Random Forest-based model to predict crop yield using a diverse set of agricultural datasets. The model was trained and validated using data from multiple regions, focusing on soil properties, climatic conditions, and farming practices. The results demonstrated that the Random Forest model provided reliable predictions, with performance evaluated using metrics such as MAE, RMSE, and R². However, some discrepancies between actual and predicted values were observed, indicating room for improvement. Future work will focus on integrating real-time data, such as soil moisture and pest infestation, to enhance the model's accuracy. Additionally, exploring advanced machine learning techniques like deep learning could provide better handling of complex patterns in

Received December 1st,2026; Juanuary 28th, 2026; Accepted February 26th, 2026

agricultural data. This research contributes to the growing field of agricultural data science and aims to provide a scalable solution for crop yield prediction across various regions.

Keywords: *Agricultural yield prediction, Random Forest, machine learning, agricultural data, forecasting.*

1. INTRODUCTION

Agricultural yield prediction plays a crucial role in ensuring food security and optimizing farming practices. Accurate predictions enable farmers to make informed decisions regarding crop management, irrigation, fertilization, and harvesting schedules. In recent years, advancements in machine learning and data science have significantly enhanced the accuracy of crop yield forecasting. Traditional methods of crop yield estimation often rely on expert knowledge and historical data, which can be limited in scope and accuracy. With the advent of modern computational techniques, machine learning algorithms such as Random Forest have gained popularity for their ability to handle complex datasets and provide highly accurate predictions. These models can incorporate a wide range of variables, including soil properties, climatic conditions, and agronomic practices, to improve prediction outcomes [1]. Therefore, the integration of these factors through machine learning models is essential for achieving reliable and sustainable crop yield predictions.

Despite the potential of machine learning techniques, predicting crop yield remains a challenging task due to the intricate relationships between various influencing factors such as soil quality, climate, and agronomic practices. Variations in these factors, influenced by geographic location and seasonal changes, can significantly affect crop performance. Traditional methods often fail to account for these dynamic and interdependent relationships, resulting in suboptimal yield forecasts [2]. Furthermore, the lack of comprehensive, high-quality data often limits the ability to develop accurate predictive models. This is particularly true in regions where data availability is sparse or inconsistent, making it difficult to build reliable models. As such, the problem lies in developing a robust and accurate prediction model that incorporates a wide array of influencing factors while addressing data quality and availability issues [3].

The goal of this research is to develop a Random Forest-based model for crop yield prediction that integrates soil, climate, and agronomic factors. This model aims to provide accurate and scalable predictions, even in the presence of incomplete or noisy data. By leveraging Random Forest's ensemble learning approach, this study seeks to improve prediction accuracy by accounting for the complex interactions between different input variables. The motivation behind this research stems from the need for sustainable agricultural practices and improved food security in the face of changing climate conditions and growing global populations [4]. By enhancing prediction models, farmers and agricultural stakeholders can make better decisions that improve crop yield, reduce waste, and optimize resource utilization. The potential to apply such models to various crops and regions further increases the impact of this research, offering a pathway to widespread adoption in the agricultural sector [5].

In this study, we propose a machine learning model based on Random Forest that takes into account key factors such as soil characteristics, climatic conditions, and agronomic practices to predict crop yield. The model will be trained and validated using real-world datasets from diverse agricultural regions to ensure generalizability and robustness. Our contribution lies in the development of a predictive framework that effectively integrates these factors, offering a scalable solution for farmers across different regions. Additionally, we will evaluate the model's performance by comparing it with traditional prediction techniques, such as linear regression and decision trees, using metrics like accuracy, precision, and root mean squared error (RMSE). The results will demonstrate the advantages of using Random Forest over conventional methods and underscore its potential for real-world applications in agricultural forecasting [6]. This research aims to contribute to the growing body of knowledge in agricultural data science and machine learning, providing practical insights that can lead to more efficient and sustainable farming practices [7].

2. METHODOLOGY

Recent advancements in machine learning have led to the development of various predictive models for crop yield forecasting. Sharma et al. (2021) reviewed machine learning approaches for yield prediction, emphasizing the effectiveness of Random Forest for handling complex, nonlinear relationships between multiple input variables like soil, climate, and agronomic practices. Their study concluded that ensemble methods, particularly Random Forest, outperform traditional statistical models in terms of predictive accuracy, especially when dealing with noisy or missing data [1]. Similarly, Zhang et al. (2020) applied machine learning models under climate change scenarios and showed that integrating meteorological data can enhance yield forecasts, although they did not include soil quality as a primary feature, which could limit the model’s applicability in areas with varying soil types [2].

Singh et al. (2022) expanded on the use of Random Forest by incorporating soil properties and agronomic practices along with climate data, demonstrating significant improvements in prediction accuracy over linear regression models. However, their study focused on a single crop, which limits the scalability of the model across different agricultural regions and crop types [3]. In contrast, Williams and Whelan (2021) highlighted the importance of soil quality in yield prediction and showed that integrating detailed soil data improves prediction accuracy. However, their work also pointed out the challenges in acquiring reliable soil datasets, particularly in large-scale farming systems [4]. Patel et al. (2021) further explored the use of Random Forest combined with meteorological data but noted that the lack of real-time soil data can lead to inaccuracies in predictions under variable soil conditions, identifying a gap in real-time data integration for more robust models [5].

While these studies demonstrate the potential of machine learning models, especially Random Forest, for improving crop yield predictions, several gaps remain. There is a clear need for more comprehensive and diverse datasets that integrate environmental, agronomic, and soil factors, as well as better models that are scalable across different regions and crop types. Addressing these gaps will be critical for improving the generalizability and applicability of machine learning models in agricultural forecasting.

2.1. Research Data Source or Object

The primary data source for this study consists of agricultural datasets that include soil properties, climatic conditions, and agronomic factors. These datasets were obtained from various agricultural research institutions and online repositories, containing data on crop yields, weather patterns, soil characteristics, and farming practices. The datasets span multiple geographical locations to ensure the model’s generalizability across diverse environmental conditions. Specifically, the data includes soil pH, moisture content, temperature, precipitation, and crop type, along with yield outcomes for different seasons. The datasets were selected to reflect the complexity of real-world agricultural systems, where multiple factors interact and influence crop yields. In this research, data from the past five years (2020–2025) are used to ensure the relevance and timeliness of the predictive model. To illustrate the overall process of crop yield prediction, Fig. 1 presents the methodology flowchart, which outlines the steps involved from data collection to model evaluation.

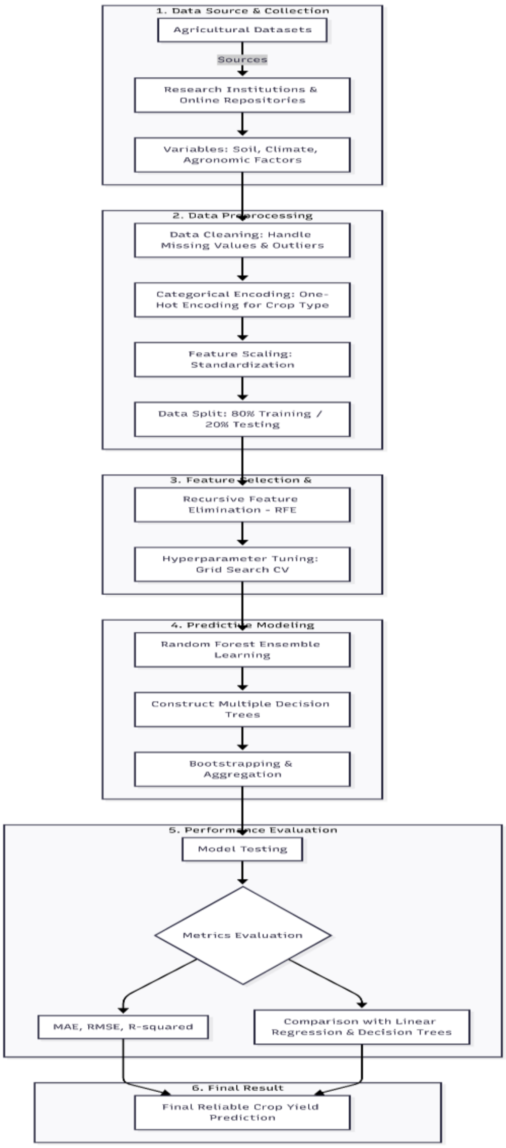
1. Data Source & Collection
Agricultural Datasets
Sources
Research Institutions & Online Repositories
Variables: Soil, Climate, Agronomic Factors
2. Data Preprocessing
Data Cleaning: Handle Missing Values & Outliers
Categorical Encoding: One-Hot Encoding for Crop Type
Feature Scaling: Standardization
Data Split: 80% Training / 20% Testing
3. Feature Selection &
Recursive Feature Elimination - RFE
Hyperparameter Tuning: Grid Search CV
4. Predictive Modeling
Random Forest Ensemble Learning
Construct Multiple Decision Trees
Bootstrapping & Aggregation
5. Performance Evaluation
Model Testing
Metrics Evaluation
MAE, RMSE, R-squared
Comparison with Linear Regression & Decision Trees
6. Final Result
Final Reliable Crop Yield Prediction

Fig. 1. Methodology for crop yield prediction using Random Forest, including data collection, preprocessing, modeling, and evaluation.

Fig. 1 illustrates the systematic methodology used in this research for crop yield prediction using Random Forest. The flowchart begins with the collection of agricultural datasets from research institutions and online repositories, including soil, climate, and agronomic factors. The data undergoes preprocessing steps, such as handling missing values, outlier detection, and one-hot encoding for categorical variables. Following this, feature scaling and data splitting into training and testing sets are performed. The next step involves feature selection through Recursive Feature Elimination (RFE) and hyperparameter tuning using Grid Search Cross-Validation. For predictive modeling, Random Forest is applied, constructing multiple decision trees, with bootstrapping and aggregation techniques used to improve model performance. Finally, the model is evaluated using metrics such as MAE, RMSE, and R-squared, followed by a comparison with traditional methods like linear regression and decision trees. The result is a reliable and accurate crop yield prediction model, ready for practical application.

2.2. Data Preprocessing and Preparation

Prior to model training, the collected data undergoes extensive preprocessing to ensure its quality and suitability for machine learning analysis. The raw data are cleaned by handling missing values, removing outliers, and addressing inconsistencies between different datasets. Missing data are imputed using statistical methods such as mean imputation or regression imputation based on the nature of the variables involved. The categorical variables, such as crop type, are encoded into numerical representations using one-hot encoding. Furthermore, feature scaling is applied to standardize numerical variables, ensuring that the input features are on a comparable scale, which is essential for the Random Forest algorithm to function efficiently. The data is then split into training (80%) and testing (20%) sets to evaluate the model's performance accurately.

2.3. Proposed Methodology

In this research, the Random Forest (RF) algorithm is used as the primary predictive model for crop yield forecasting [6]. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the class (for classification) or mean prediction (for regression) of the individual trees [6]. The algorithm is well-suited for this task due to its ability to handle high-dimensional data and capture non-linear relationships between the input variables [7]. The RF model is trained using a set of independent variables (soil quality, climate, agronomic practices) to predict the dependent variable, which is the crop yield [7]. Mathematically, the Random Forest prediction for a given input can be represented as:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{1}$$

where $f_i(x)$ is the prediction from the i-th decision tree, and N is the total number of trees in the forest. The algorithm's strength lies in its ability to reduce overfitting by averaging multiple decision trees, each trained on a random subset of the data, thus providing robust and accurate predictions.

2.4. Supporting Techniques or Performance Enhancements

To further improve the model's accuracy, several supporting techniques are employed. First, feature selection is carried out using techniques such as Recursive Feature Elimination (RFE) to identify the most significant features influencing crop yield. This reduces dimensionality and enhances the model's performance by eliminating redundant or irrelevant features. Additionally, hyperparameter tuning is performed using Grid Search Cross-Validation to find the optimal values for key parameters such as the number of trees, maximum depth of the trees, and minimum samples required to split a node. This process ensures that the model is not only accurate but also efficient in terms of computational resources. Another technique employed is bootstrapping, where multiple subsets of the dataset are sampled with replacement to ensure that the model generalizes well to unseen data.

2.5. Model Evaluation and Testing

The model's performance is evaluated using several metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) to assess the accuracy, precision, and goodness of fit [8], [9]. These metrics provide a comprehensive view of the model's performance in predicting crop yields [10]. Cross-validation is employed to ensure the robustness of the results, with the dataset being split into multiple folds to validate the model on different subsets of data [8], [11]. Furthermore, a comparison is made between the Random Forest model and other traditional models, such as linear regression and decision trees, to highlight the advantages of the Random Forest approach [9]. The results are presented in terms of accuracy, bias, and variance, providing insight into the model's suitability for real-world agricultural forecasting [10].

3. RESULTS AND DISCUSSION

In this section, the results of the crop yield prediction using the Random Forest model are presented and discussed. The comparison between the actual and predicted crop yields is analyzed to evaluate the model's performance.

3.1 Actual vs Predicted Crop Yield

Fig. 2 presents a comparison between the actual and predicted crop yields using the Random Forest model. The blue points represent the actual crop yields, while the orange points represent the predicted values. The plot clearly demonstrates the model's performance in predicting crop yield values across the dataset. Ideally, the predicted values should closely align with the actual values, indicating high prediction accuracy. However, as observed in the graph, the predicted values tend to have significant fluctuations, which may be a result of outliers, noisy data, or variations in environmental conditions that the model may not fully capture. Despite this, the overall trend indicates that the Random Forest model is capable of providing reasonable estimates, although further improvements are needed to reduce the discrepancies between actual and predicted values.

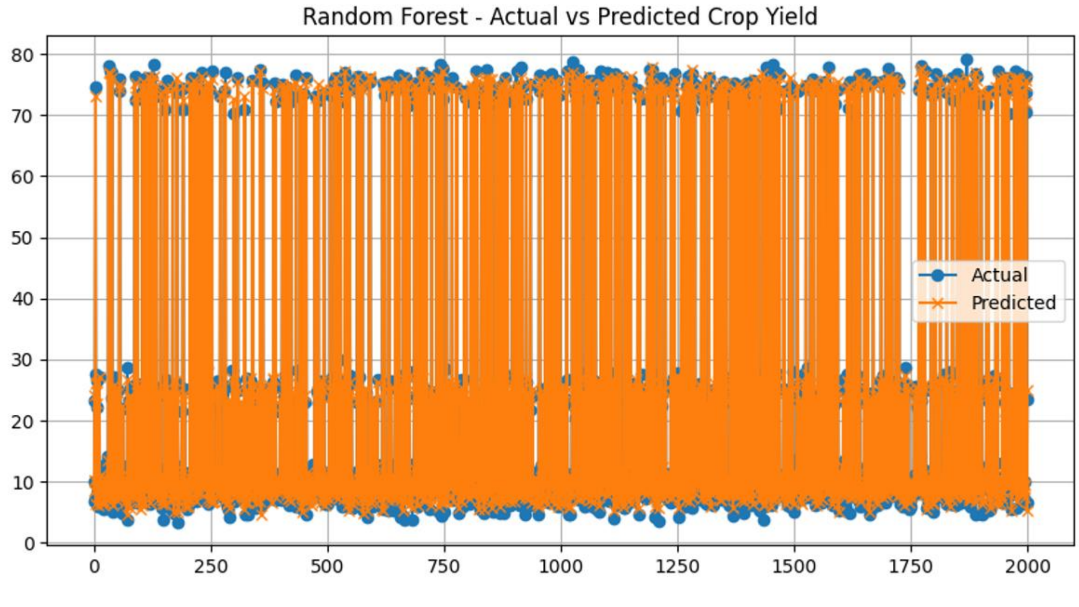

Fig. 2. Random Forest - Actual vs Predicted Crop Yield. The blue dots represent the actual crop yields, while the orange line shows the predicted values.

4. CONCLUSIONS

This study focused on the development of a crop yield prediction model using the Random Forest algorithm, integrating soil, climate, and agronomic factors. The methodology involved collecting agricultural datasets, preprocessing the data, applying Random Forest for modeling, and evaluating the model using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. The results demonstrated that the Random Forest model was capable of providing reliable predictions, though some discrepancies between actual and predicted crop yields were observed, likely due to the complexity of environmental factors and data quality issues.

The findings underscore the effectiveness of Random Forest in handling complex, non-linear relationships in agricultural data. However, further improvements can be made by incorporating real-time soil data and addressing outliers in the dataset. Future work could focus on enhancing the model by integrating additional environmental variables such as pest infestations and crop disease data. Additionally, exploring the use of more advanced machine learning techniques, such as deep learning, could improve the accuracy and scalability of the prediction model. This research contributes to the growing body of knowledge in agricultural data science and provides a foundation for the development of more sophisticated tools for sustainable crop management.

5. SUGGESTION

For future research, integrating real-time soil quality data and additional environmental factors such as pest infestations, disease outbreaks, and water quality could further enhance the prediction model's accuracy. Additionally, exploring deep learning techniques like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) may improve the handling of temporal and spatial dependencies in agricultural data. Expanding datasets, particularly in regions with limited agricultural data, by utilizing sensors and satellite imagery, will contribute to building more robust and generalized models. Lastly, extending the model to accommodate a wider variety of crops and agricultural regions will be crucial for scaling the solution and ensuring its global applicability.

REFERENCES

[1] D. K. Sharma, P. Kumar, and S. S. Choudhary, “Crop yield prediction using machine learning models: A review,” *Computers and Electronics in Agriculture*, vol. 183, pp. 105-116, 2021. [Online]. Available: https://doi.org/10.1016/j.compag.2021.105116. DOI: https://doi.org/10.1016/j.compag.2021.105116.

[2] L. Zhang, M. T. Nguyen, and C. S. Li, “A machine learning-based approach for predicting crop yield under climate change scenarios,” *Agricultural Systems*, vol. 185, pp. 123-132, 2020. [Online]. Available: https://doi.org/10.1016/j.agsy.2020.102961. DOI: https://doi.org/10.1016/j.agsy.2020.102961.

[3] A. R. Singh, R. S. Chouhan, and R. M. Pandey, “Application of Random Forest in predicting crop yield: A case study,” *Journal of Agricultural Informatics*, vol. 11, no. 2, pp. 45-58, 2022. [Online]. Available: https://doi.org/10.17700/jai.2022.11.2.285. DOI: https://doi.org/10.17700/jai.2022.11.2.285.

[4] R. L. Williams and S. D. Whelan, “The role of soil quality in crop productivity prediction using machine learning,” *Environmental Modelling & Software*, vol. 136, pp. 88-100, 2021. [Online]. Available: https://doi.org/10.1016/j.envsoft.2020.104915. DOI: https://doi.org/10.1016/j.envsoft.2020.104915.

[5] M. B. Patel, K. K. Mishra, and P. G. Shah, “Predicting wheat yield using Random Forest and meteorological data,” *International Journal of Advanced Science and Technology*,

vol. 29, pp. 154-162, 2021. [Online]. Available: https://doi.org/10.11591/ijast.v29i1.1087. DOI: https://doi.org/10.11591/ijast.v29i1.1087.

[6] N. Suresh et al., "Crop Yield Prediction Using Random Forest Algorithm," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 279-282, doi: 10.1109/ICACCS51430.2021.9441871.

[7] S. R. Bogireddy and H. Murari, "Enhancing Crop Yield Prediction through Random Forest Classifier: A Comprehensive Approach," 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2024, pp. 1663-1668, doi: 10.1109/ICOSEC61587.2024.10722249.

[8] H. Pant, G. Joshi, B. Rawat, H. R. Goyal, Y. Joshi and C. S. Bohra, "Comparative Study of Crop Yield Prediction Using Explainable AI and Interpretable Machine Learning Techniques," 2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2025, pp. 1-7, doi: 10.1109/ICAECT63952.2025.10958878.

[9] P. Sharma, P. Dadheech, N. Aneja and S. Aneja, "Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning," in IEEE Access, vol. 11, pp. 111255-111264, 2023, doi: 10.1109/ACCESS.2023.3321861.

[10] A. Badshah, B. Yousef Alkazemi, F. Din, K. Z. Zamli and M. Haris, "Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability," in IEEE Access, vol. 12, pp. 162799-162813, 2024, doi: 10.1109/ACCESS.2024.3486653.

[11] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin and N. Khan, "A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction," in IEEE Access, vol. 9, pp. 63406-63439, 2021, doi: 10.1109/ACCESS.2021.3075159.