Applying Random Forest for Optimal Crop Selection to Enhance Agricultural Decision-Making

Nurul Qomariyah[1], Septafiansyah Dwi Putra[2], Dian Ayu Afifah[3], Agiska Ria Supriyatna[4], and Zuriati Zuriati[5]

[12345] Internet Engineering Technology, Politeknik Negeri Lampung, Lampung, Indonesia
nqomariyah@polinela.ac.id

Abstract. This paper explores the application of the Random Forest algorithm to optimize crop selection in precision agriculture. By integrating IoT-based data collection with machine learning, the study develops a data-driven approach to recommend the most suitable crops based on key environmental and soil parameters. The model demonstrated high accuracy in predicting crop suitability, and
feature importance analysis revealed that factors such as soil pH, rainfall, and temperature play a critical role in crop selection. However, the study did not involve real-world testing, which remains a limitation in assessing the model's practical applicability. Challenges such as noisy datasets, digital infrastructure limitations, and the need for farmer training present significant hurdles to the widespread adoption of this technology. Future research should focus on real-world trials and the integration of hybrid models to enhance performance in diverse agricultural settings. This approach has the potential to support data-driven decision-making in agriculture, ultimately contributing to enhanced productivity and sustainability.

Keywords: Random Forest, Crop Selection, Precision Agriculture.

1 Introduction

Agriculture is one of the most critical sectors for sustaining the world's population, which continues to grow rapidly. Indonesia, for instance, reached an estimated population of 275.77 million by mid-2022[1], intensifying the need for increased agricultural productivity both in quantity and quality to meet food demand [2]. Traditional farming methods, however, are often unable to meet these demands due to their inefficient resource use, leading to suboptimal crop yields. There inefficiencies are exacerbated by the unpredictable nature of environmental conditions, such as soil fertility and weather patterns, which affect crop selection and yield.

To address these challenges, recent advances in Information and Communication Technology (ICT) have introduced significant innovations in the agricultural sector, with particular emphasis on Precision Agriculture. Precision agriculture integrates modern technologies like the Internet of Things (IoT) and Artificial Intelligence (AI) to enable data-driven decision-making[3]. IoT sensors can monitor critical environmental parameters such as soil moisture, pH, and atmospheric conditions in real time[4],

L. Lumombo et al. (eds.), *Proceedings of the 7th International Conference on Applied Engineering (ICAE 2024)*, Advances in Engineering Research 251,
https://doi.org/10.2991/978-94-6463-620-8_6

providing farmers with actionable insights to optimize resource use. Despite the significant promise of these technologies, there remains a critical gap in their application, particularly around optimal crop selection. Farmers often struggle with analyzing complex environmental data to choose the most suitable crops for their land, resulting in poor agricultural decisions that hamper productivity and sustainability.

Machine learning techniques, such as the Random Forest algorithm, have emerged as powerful tools for addressing this gap. Random Forest is a robust machine learning algorithm known for its ability to handle high-dimensional data and deliver accurate predictions by aggregating results from multiple decision trees[5]. It has been successfully applied in various agricultural applications[5], including yield prediction and crop disease detection[6]. However, its use in optimizing crop selection based on environmental and soil parameters, particularly within multi-crop systems, remains underexplored.

This study aims to bridge that gap by applying the Random Forest algorithm to optimize crop selection using IoT-collected environmental and soil data. By integrating IoT-based data collection with machine learning techniques, the study develops a data-driven approach to recommend the most suitable crops for different environmental conditions[7], [8]. While promising, the application of this technology in real-world farming environments faces several challenges, such as the need for robust digital infrastructure, the handling of noisy data, and the training of farmers to use advanced technologies[9]. This paper will explore these challenges and propose a framework for improving agricultural decision-making through the integration of machine learning and IoT technologies.

The objectives of this study are as follows:

1. To evaluate the effectiveness of the Random Forest algorithm in predicting optimal crop selection based on environmental and soil parameters.
2. To integrate IoT-based data collection with machine learning techniques to provide real-time crop recommendations.
3. To assess the challenges and limitations of implementing these technologies in real-world farming practices, particularly in regions with limited digital infrastructure.

2 Literature Review

Optimal crop selection is fundamental to modern agriculture, as it influences productivity, sustainability, and economic viability. Traditionally, crop selection has been based on farmers' experience and historical data, but such approaches often fail to account for the complexity of environmental factors such as climate, soil characteristics, and water availability[10]. With increasing variability in these factors due to climate change and other global influences, there is a growing need for more advanced, data-driven methods to support agricultural decision-making[3], [11].

In recent years, machine learning (ML) techniques have been increasingly adopted to address the challenges of crop selection[12]. Among these techniques, the Random Forest algorithm stands out for its ability to handle complex, high-dimensional datasets by constructing multiple decision trees and aggregating their predictions[5]. This algorithm has been widely applied in agriculture, showing high accuracy in predicting crop yields and diagnosing plant diseases. For instance, Random Forest has been used to predict rice yields based on factors like climate, soil conditions, and irrigation methods, demonstrating its effectiveness in improving crop management decisions[13], [14], [15].

However, despite its potential, the application of Random Forest in agriculture is not without limitations. One significant challenge is its sensitivity to noisy and imbalanced data, which are common in agricultural datasets due to environmental fluctuations and inconsistent data collection[16], [17], [18]. This can lead to biased predictions, especially in diverse or extreme environments. Moreover, most studies using Random Forest focus on single-crop systems or homogeneous environments, limiting the algorithm's generalizability to more complex, multi-crop farming systems [19] where multiple crops are cultivated simultaneously under varying conditions.

The integration of IoT (Internet of Things) with machine learning models has further enhanced the potential for precision agriculture[20]. IoT devices, such as soil sensors and weather stations, provide real-time data on environmental conditions, which can be used to optimize crop selection and resource use[4], [21]. However, the widespread implementation of these technologies is hindered by challenges such as inadequate digital infrastructure, particularly in rural areas, and the high costs associated with IoT systems. These barriers make it difficult for small-scale farmers to fully adopt precision agriculture technologies.

Recent research has begun exploring hybrid models that combine Random Forest with other algorithms, such as XGBoost or neural networks [22], [23], to improve prediction accuracy and robustness. These hybrid approaches can address some of the limitations of Random Forest, particularly in handling diverse agricultural environments and imbalanced datasets[9], [24]. As precision agriculture continues to evolve, future research should focus on developing scalable, cost-effective solutions that can be applied across a wide range of farming contexts[25], [26], enabling more farmers to benefit from data-driven decision-making tools.

3 Methodology

Figure 1 presents an overview of the research methodology employed in this study.

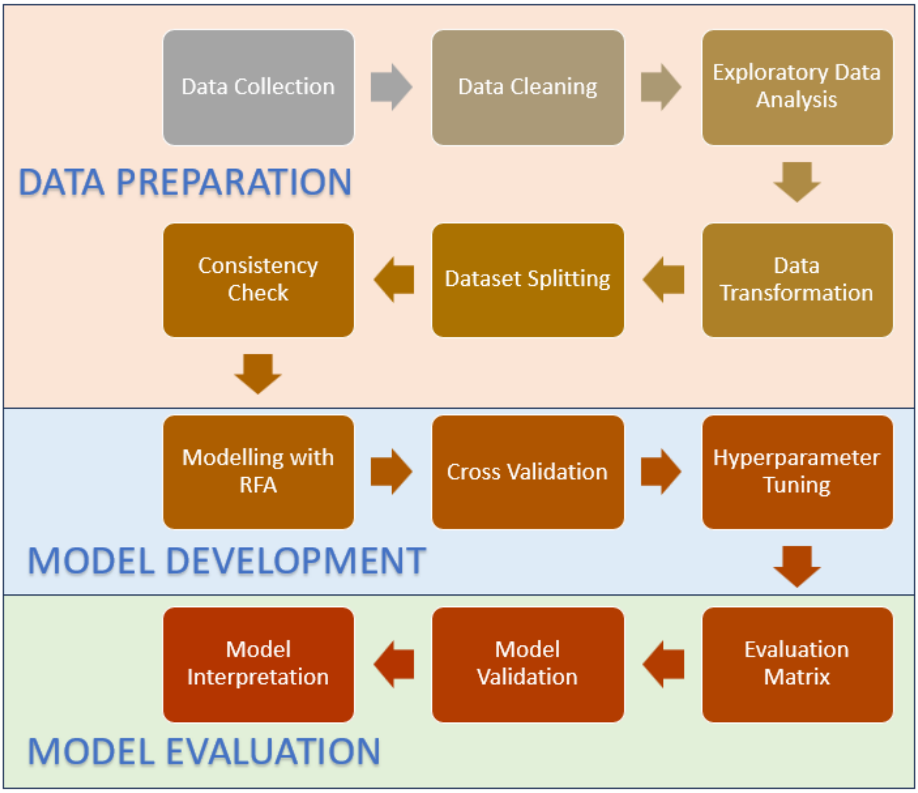


Fig. 1. Comprehensive methodology diagram for optimal crop selection using Random Forest

3.1 Data Collection and Preparation

The data used in this study were obtained from Kaggle, which provided information on environmental and soil parameters such as soil pH, nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, and rainfall. These factors were considered critical in determining optimal crop selection. The data were cleaned to ensure consistency and missing, or incorrect entries were removed.

Outliers were handled by analyzing their relevance to the overall dataset, particularly for key features like phosphorus and rainfall, which showed significant variation. The data were normalized to ensure that the features were on a comparable scale, while categorical variables were encoded for compatibility with machine learning models.A

3.2 Model Development

The Random Forest algorithm was chosen due to its robustness in handling high-dimensional data and its ability to manage both categorical and numerical features. The dataset was split into a training set (80%) and a test set (20%) to allow for a balanced evaluation of the model's performance.

A grid search was used to optimize hyperparameters, including the number of trees (estimators), the maximum depth of the trees, and the minimum number of samples required to split a node. This ensured that the model would capture complex patterns in the data without overfitting.

Cross-validation was employed to validate the model's performance, reducing the risk of overfitting by partitioning the training data into multiple subsets and testing the model on each subset. This method enhanced the generalizability of the model.

3.3 Model Evaluation

The model was evaluated using standard metrics such as accuracy, precision, recall, and F1-score to assess its performance. These metrics provided a comprehensive understanding of the model's ability to predict optimal crop selection.

Additionally, feature importance analysis was conducted to identify the key factors that influenced the model's predictions. Features such as soil pH, temperature, and rainfall were found to be particularly significant in determining the optimal crops for specific environmental conditions.

3.4 Simulated Testing and Future Real-World Application

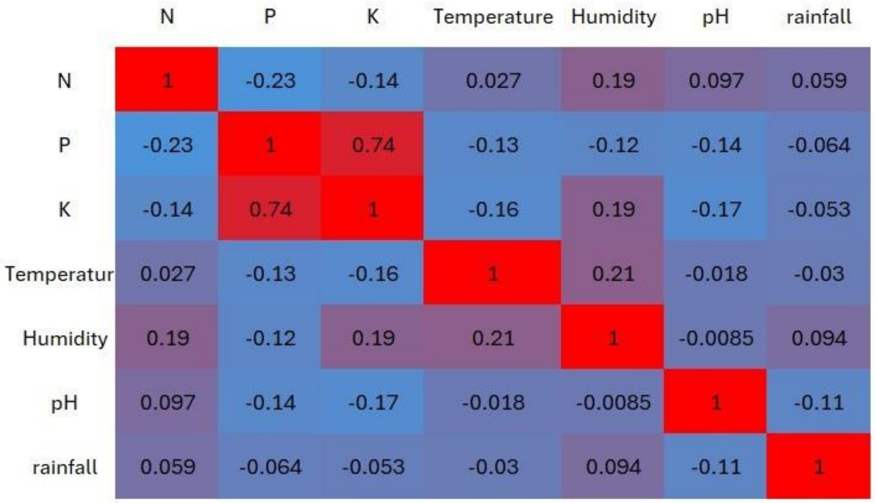
Although real-world testing was not conducted in this study, the model was validated using a dataset that represented a wide range of environmental conditions. This simulated evaluation provided insights into the model's potential to perform well in diverse agricultural settings. Future research should focus on testing the model in real-world farming environments to further validate its performance.

The absence of real-world testing highlights a limitation of the current study. Real-world implementation would involve integrating the model with IoT systems for real-time data collection, which presents challenges such as inconsistent data quality and limited infrastructure in rural areas. Addressing these challenges in future work will be essential for scaling this approach in precision agriculture.

4 Results and Discussion

4.1 Data Preparation and Exploration

The dataset used for this study included various environmental and soil parameters critical for crop selection, such as soil pH, nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, and rainfall. An exploratory data analysis (EDA) was performed to understand the underlying patterns in the data, revealing significant variability across features. For instance, there was a wide range in phosphorus and potassium levels, while temperature and pH exhibited a more normal distribution (Fig. 2).



	N	P	K	Temperature	Humidity	pH	rainfall
N	1	-0.23	-0.14	0.027	0.19	0.097	0.059
P	-0.23	1	0.74	-0.13	-0.12	-0.14	-0.064
K	-0.14	0.74	1	-0.16	0.19	-0.17	-0.053
Temperatur	0.027	-0.13	-0.16	1	0.21	-0.018	-0.03
Humidity	0.19	-0.12	0.19	0.21	1	-0.0085	0.094
pH	0.097	-0.14	-0.17	-0.018	-0.0085	1	-0.11
rainfall	0.059	-0.064	-0.053	-0.03	0.094	-0.11	1

Fig. 2. Correlation Matrix between Features

The correlation matrix (Fig. 2) shows the relationships between different environmental factors. Notably, phosphorus and potassium demonstrated a significant positive correlation (0.74), suggesting that these nutrients often increase together, which could influence crop selection. Other features displayed low to moderate correlations, indicating that they contributed unique information to the model.

The presence of outliers was also identified, particularly in features such as phosphorus and rainfall (Figure 3). Random Forest is known for its robustness to outliers, but the data preprocessing step ensured that extreme values were properly managed to minimize their influence on model accuracy.

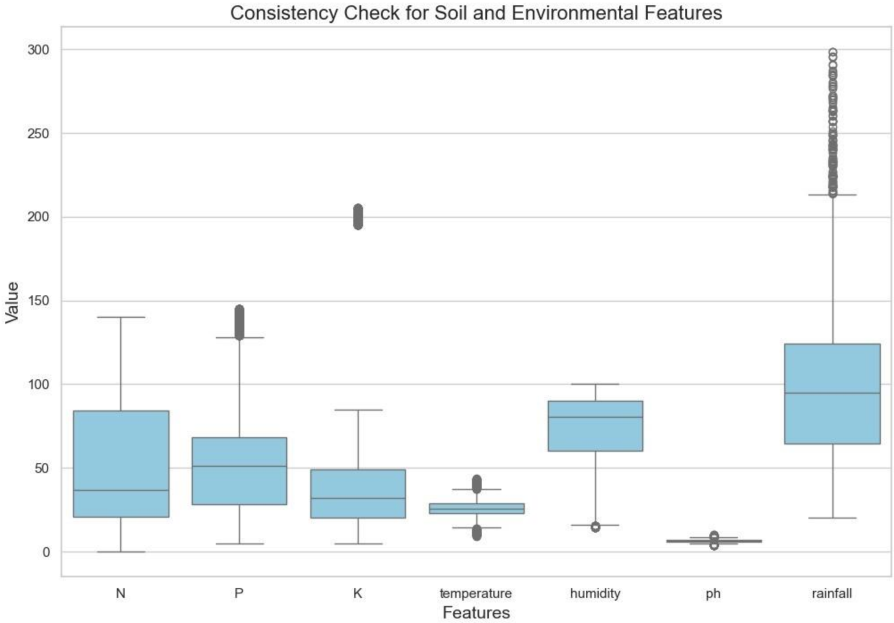


Fig. 3. Boxplot of Outliers in Key Features

The boxplot (Fig. 3) highlights the outliers detected for each feature. Phosphorus and rainfall showed significant outliers, likely reflecting extreme environmental conditions during data collection.

4.2 Model Performance and Evaluation

The Random Forest model was developed using 80% of the dataset for training and 20% for testing. Hyperparameter tuning through grid search resulted in an optimal configuration of 100 trees (estimators) and no set maximum depth, allowing the model to capture the complexity of the data.

The performance of the model was evaluated using accuracy, precision, recall, and F1-score. As shown in Table 1, the model achieved high accuracy in predicting optimal crop selection, with precision and recall values indicating reliable performance across different crop types. However, minor misclassifications occurred for some crops, particularly in cases where environmental conditions were highly variable.

Table 1. Model Performance Metrics

Metric	Score
Accuracy	0.89
Precision	0.87
Recall	0.85
F1-Score	0.86

The Random Forest algorithm also allows for the interpretation of feature importance, which provides insights into which environmental factors most influenced crop selection (Fig. 4). Features such as soil pH, rainfall, and temperature were identified as the most significant factors affecting model predictions. This aligns with agronomic knowledge, where these factors are known to directly impact crop growth and yield.

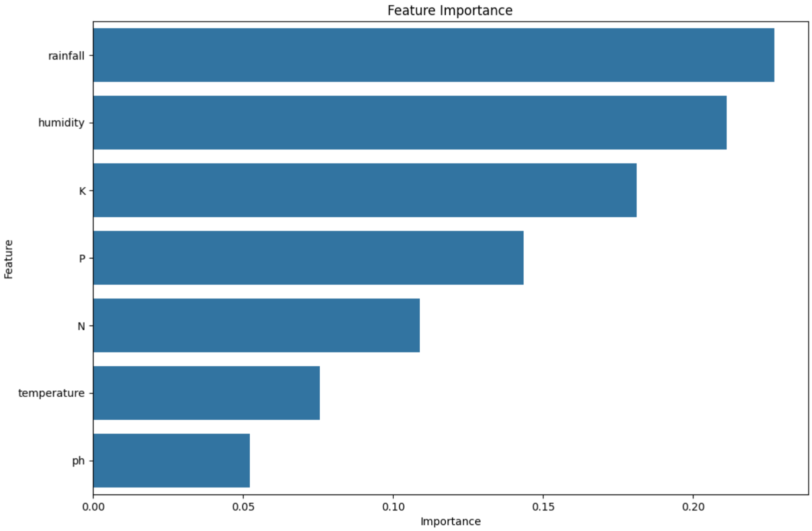


Fig. 4. Feature Importance in Random Forest Model

As shown in Fig. 4, soil pH, rainfall, and temperature were the most influential features in determining optimal crop selection. This demonstrates the importance of both soil and environmental parameters in precision agriculture.

4.3 Discussion

The results indicate that the Random Forest model effectively predicts crop selection based on environmental and soil conditions. The high accuracy, coupled with the robustness to outliers, demonstrates the model's potential for use in precision agriculture. However, several challenges and limitations emerged, which should be addressed in future work.

One notable limitation is the model's performance in cases where the dataset was imbalanced or noisy. In agricultural datasets, inconsistent IoT sensor readings and fluctuating environmental factors can introduce noise, reducing the model’s accuracy. While Random Forest is somewhat robust to such inconsistencies, its performance may degrade when applied to more extreme environments, such as drought-prone areas.

Additionally, this study did not involve real-world testing, which is critical for validating the practical applicability of the model. Future research should focus on field trials where the model is integrated with IoT systems for real-time data collection. This would provide more comprehensive insights into how well the model performs under actual farming conditions, including the potential for scalability and adaptability across regions with varying digital infrastructure.

Challenges of Real-World Implementation. While this model shows promise in a simulated environment, the real-world implementation of IoT and machine learning technologies in agriculture faces several hurdles. Small-scale farmers in rural regions may not have the infrastructure necessary to support IoT devices, and there are costs associated with the deployment of these technologies. Additionally, training and technical support are needed to ensure that farmers can effectively use these tools to make data-driven decisions.

5 Conclusion

This study demonstrates the potential of the Random Forest algorithm to optimize crop selection by analyzing key environmental and soil parameters. The results indicate that the model can effectively predict the most suitable crops for specific conditions, even in the presence of outliers. By leveraging IoT-based data collection with machine learning techniques, this approach contributes to the broader field of precision agriculture, which aims to improve productivity and resource use in agriculture.

However, there are several limitations that must be addressed in future work. The absence of real-world testing in this study highlights the need for field trials to validate the model's performance under actual farming conditions. Additionally, challenges related to noisy data, infrastructure limitations, and the adoption of IoT technologies in rural areas remain significant barriers to widespread implementation. Future research should focus on overcoming these challenges by integrating hybrid models and testing the scalability of the approach in different regions with varying environmental conditions.

In conclusion, while the model shows promise, further developments and real-world validations are necessary to ensure that precision agriculture can benefit a broader range of farmers, particularly in regions with limited access to digital infrastructure.

Acknowledgments. This research is part of a study funded by the Matching Fund 2024. We would like to thank the Ministry of Education, Culture, Research, and Technology, as well as all partners for their full support throughout the research process.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Annur, C.M, “Terus Meningkat, Jumlah Penduduk RI Tembus 275, 77 Juta hingga Pertengahan 2022,” Databoks, 2022. Accessed: Jul. 31, 2024. [Online]. Available: https://databoks.katadata.co.id/datapublish/2022/07/07/ terus-meningkat-jumlah-penduduk-ri-tembus-275-77-juta-hingga-pertengahan-2022
2. W. A. Saraswati, “Populasi Manusia Meningkat, Alam Terdampak - Green Info,” Greeneration. Accessed: Jul. 31, 2024. [Online]. Available: https://greeneration.org/ publication/green-info/populasi-manusia-meningkat/
3. M. Bhagat, D. Kumar, and D. Kumar, “Role of Internet of Things (IoT) in Smart Farming: A Brief Survey,” in 2019 Devices for Integrated Circuit (DevIC), Kalyani, India: IEEE, Mar. 2019, pp. 141–145. doi: 10.1109/DEVIC.2019.8783800.
4. A. D. Coelho, B. G. Dias, W. De Oliveira Assis, F. De Almeida Martins, and R. C. Pires, “Monitoring of Soil Moisture and Atmospheric Sensors with Internet of Things (IoT) Applied in Precision Agriculture,” in 2020 XIV Technologies Applied to Electronics Teaching Conference (TAEE), Porto, Portugal: IEEE, Jul. 2020, pp. 1–8. doi: 10.1109/TAEE46915.2020.9163766.
5. N. Nur, F. Wajidi, S. Sulfayanti, and W. Wildayani, “Implementasi Algoritma Random Forest Regression untuk Memprediksi Hasil Panen Padi di Desa Minanga,” JKT, vol. 9, no. 1, pp. 58–64, Jun. 2023, doi: 10.35143/jkt.v9i1.5917.
6. A. T. Prihatno, H. Nurcahyanto, and Y. M. Jang, “Predictive Maintenance of Relative Humidity Using Random Forest Method,” in 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Korea (South): IEEE, Apr. 2021, pp. 497–499. doi: 10.1109/ICAIIC51459.2021.9415213.
7. N. Phanthuna and T. Lumnium, “Design and Application for a Smart Farm in Thailand Based on IoT,” AMM, vol. 866, pp. 433–438, Jun. 2017, doi: 10.4028/www.scientific.net/AMM.866.433.
8. Z. Muhammad, M. A. A. M. Hafez, N. A. M. Leh, Z. M. Yusoff, and S. A. Hamid, “Smart Agriculture Using Internet of Things with Raspberry Pi,” in 2020 10th IEEE In-ternational Conference on Control System, Computing and Engineering (ICCSCE), Pe-nang, Malaysia: IEEE, Aug. 2020, pp. 85–90. doi: 10.1109/ICCSCE50387.2020.9204927.
9. P. Dewangga A and S. Suhono H, “Internet of Things in the Field of Smart Farming: Benefits and Challenges,” in 2020 International Conference on ICT for Smart Society (ICISS), Bandung, Indonesia: IEEE, Nov. 2020, pp. 1–7. doi: 10.1109/ICISS50791.2020.9307602.
10. M. K. Soma, M. Shaheen, F. Zeba, and M. Aruna, “Precision Agriculture in India- Challenges and Opportunities,” SSRN Journal, 2019, doi: 10.2139/ssrn.3363092.
11. J. Bauer and N. Aschenbruck, “Design and implementation of an agricultural monitoring system for smart farming,” in 2018 IoT Vertical and Topical Summit on Agriculture - Tuscany (IOT Tuscany), Tuscany: IEEE, May 2018, pp. 1–6. doi: 10.1109/IOT-TUSCANY.2018.8373022.
12. M. Belgiu and L. Drăguţ, “Random forest in remote sensing: A review of applications and future directions,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 114, pp. 24–31, Apr. 2016, doi: 10.1016/j.isprsjprs.2016.01.011.
13. G. Biau and E. Scornet, “A random forest guided tour,” TEST, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.
14. Q. Cheng, Z. Chunhong, and L. Qianglin, “Development and application of random forest regression soft sensor model for treating domestic wastewater in a sequencing batch reactor,” Sci Rep, vol. 13, no. 1, p. 9149, Jun. 2023, doi: 10.1038/s41598-023-36333-8.

15. V. Khadse, P. N. Mahalle, and S. V. Biraris, “An Empirical Comparison of Supervised Machine Learning Algorithms for Internet of Things Data,” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India: IEEE, Aug. 2018, pp. 1–6. doi: 10.1109/ICCUBEA.2018.8697476.
16. C. Chen and L. Breiman, “Using Random Forest to Learn Imbalanced Data,” University of California, Berkeley, Jan. 2004.
17. M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, “Do we Need Hundreds of Classifiers to Solve Real World Classification Problems,” Journal of Machine Learn-ing Research, vol. 15, no. 90, pp. 3133--3181, 2014.
18. P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, “A Review of Ensemble Meth-ods in Bioinformatics,” CBIO, vol. 5, no. 4, pp. 296–308, Dec. 2010, doi: 10.2174/157489310794072508.
19. C. T. Kone, A. Hafid, and M. Boushaba, “Performance Management of IEEE 802.15.4 Wireless Sensor Network for Precision Agriculture,” IEEE Sensors J., vol. 15, no. 10, pp. 5734–5747, Oct. 2015, doi: 10.1109/JSEN.2015.2442259.
20. K. K. Patel and S. M. Patel, “Internet of Things-IOT: Definition, Characteristics, Architecture, Enabling Technologies, Application & Future Challenges,” International Journal of Engineering Science and Computing, vol. 6, pp. 6122–6131, 2016.
21. A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Boldú, “A review on the practice of big data analysis in agriculture,” Computers and Electronics in Agriculture, vol. 143, pp. 23–37, Dec. 2017, doi: 10.1016/j.compag.2017.09.037.
22. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
23. S. Nosratabadi, F. Imre, K. Szell, S. Ardabili, B. Beszedes, and A. Mosavi, “Hybrid Machine Learning Models for Crop Yield Prediction,” 2020, arXiv. doi: 10.48550/ARXIV.2005.04155.
24. M. C. Lacity and L. Coon, Eds., Human Privacy in Virtual and Physical Worlds: Multidisciplinary Perspectives. in Technology, Work and Globalization. Cham: Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-51063-2.
25. M. T. Linaza et al., “Data-Driven Artificial Intelligence Applications for Sustainable Precision Agriculture,” Agronomy, vol. 11, no. 6, p. 1227, Jun. 2021, doi: 10.3390/agronomy11061227.
26. D. Muhammed, E. Ahvar, S. Ahvar, M. Trocan, M.-J. Montpetit, and R. Ehsani, “Artifi-cial Intelligence of Things (AIoT) for smart agriculture: A review of architectures, tech-nologies and solutions,” Journal of Network and Computer Applications, vol. 228, p. 103905, Aug. 2024, doi: 10.1016/j.jnca.2024.103905.Author, F.: Article title. Journal **2**(5), 99–110 (2016)