**Computer Science (CO-SCIENCE)**
Volume 6 Issue 1 January 2026
Accreditation **Sinta 4** No. SK : 230/E/KPT/2022

# Optimization of Crop Recommendation Model Using Ensemble Learning Techniques for Multiclass Classification

**Siti Marlina[1], Titik Misriati[2*], Riska Aryanti[3]**

[1,2,3]Bina Sarana Informatika University
Jl. Kramat Raya No. 98, RT.2/RW.9, Kwitang, Senen District, Jakarta, Indonesia

e-mail: [1]siti.smr@bsi.ac.id, [2*]titik.tmi@bsi.ac.id, [3]riska.rts@bsi.ac.id

(*) Corresponding Author

Article Info: Received: 24-09-2025 | Revised : 10-11-2025 | Accepted : 20-11-2025

**Abstracts** - Crop recommendation systems play a crucial role in modern agriculture by helping farmers make data-driven decisions to maximize yield, optimize resource use, and ensure sustainable farming practices. By analyzing environmental and soil parameters, these systems can suggest the most suitable crops for specific conditions, reducing the risks of crop failure and improving overall productivity. This study evaluates the performance of five ensemble learning algorithms—Random Forest, Extra Trees, CatBoost, XGBoost, and LightGBM—for multiclass classification in a crop recommendation system. All models achieved high accuracy above 98%, with Random Forest demonstrating the best and most stable performance. The feature importance analysis revealed that climatic factors, particularly rainfall and humidity, contributed the most to prediction outcomes, followed by macronutrients such as potassium, phosphorus, and nitrogen. In contrast, temperature and soil pH showed relatively lower influence. These findings highlight the dominance of climatic factors over soil chemical properties and demonstrate the capability of ensemble learning methods to capture complex data patterns. Random Forest is recommended as the primary model to support more effective land management and crop cultivation strategies.

Keywords : Ensemble Learning, Classification, Crop Recommendation, Random Forest Algorithm, Multiclass

## INTRODUCTION

The global demand for sustainable agriculture continues to grow as climate change, population increase, and food security concerns place pressure on agricultural productivity (Food and Agriculture Organization (FAO), 2021).Optimizing crop selection plays a crucial role in improving yield, resource utilization, and resilience to environmental variability. Traditional crop recommendation systems, which rely heavily on expert knowledge and rule-based approaches, often lack adaptability to diverse environmental and soil conditions (Prity et al., 2024) Consequently, data-driven approaches have gained significant attention in recent years (Patel & B. Patel, 2023).

Machine learning (ML) methods have been widely applied in precision agriculture, enabling the integration of soil characteristics, climate data, and historical yield information to recommend suitable crops. Among them, ensemble learning techniques (Sudianto & Cahyadi, 2025), such as Random Forest (Meiriyama & Sudiadi, 2022), Extremely Randomized Trees (Hussein & R. M. Zeebaree, 2024), Gradient Boosting frameworks like XGBoost (Arumugam S. S. L. et al., 2024) and LightGBM (Nguyen et al., 2024), as well as CatBoost (Srinivasu et al., 2024) have demonstrated strong predictive performance across various agricultural tasks (Hasan et al., 2023). These models leverage multiple weak learners to reduce variance and bias, thereby improving classification accuracy in complex, multiclass scenarios.

Building on this foundation, recent studies highlight the effectiveness of ensemble models in agricultural recommendation systems. For instance, an ensemble-based framework was developed to recommend suitable crops, achieving superior accuracy compared to single models (Hasan et al., 2023). Similarly, previous research employed Random Forest and XGBoost regressors for crop selection, demonstrating robust predictive capability (Rahman et al., 2024). Ensemble approaches also extend beyond agriculture, with successful applications in structural engineering (Daniel, 2024), healthcare (Elshewey et al., 2025), and financial risk assessment (Ying et al., 2025), underscoring their versatility and adaptability.

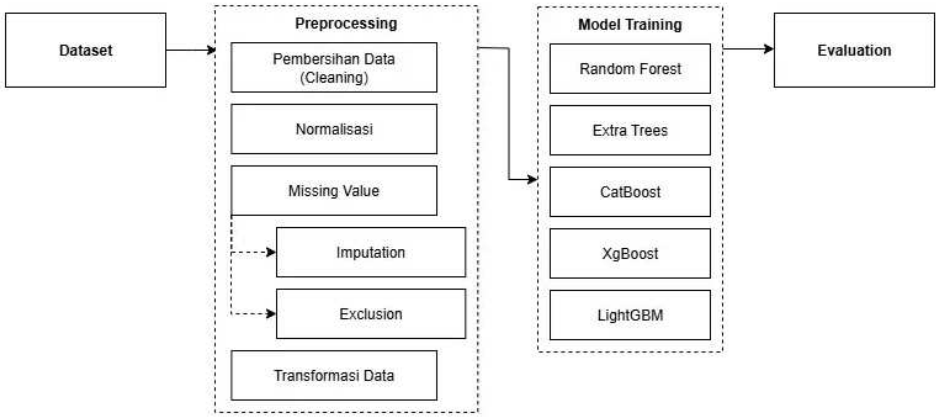
Despite these advancements, several challenges remain in optimizing ensemble models for multiclass classification in agriculture. Hyperparameter sensitivity, data imbalance, and model interpretability continue to hinder practical adoption (Haddouchi & Berrado, 2024); (Fajar et al., 2024). Furthermore, the heterogeneity of soil and climatic conditions across regions demands models that generalize well while maintaining high predictive

accuracy.

Therefore, this study proposes the optimization and performance evaluation of crop recommendation models using ensemble learning techniques to tackle the challenges of multiclass classification. By systematically evaluating and fine-tuning algorithms such as Random Forest, XGBoost, LightGBM, and CatBoost, this research aims to enhance classification accuracy, interpretability, and robustness. The novelty of this study lies not only in its comprehensive assessment of ensemble learning algorithms for multiclass agricultural data but also in its introduction of a comparative framework that integrates feature‑response interactions unique to crop‑specific conditions. This framework enables a deeper understanding of how climatic and soil variables collectively influence model performance, offering a more transparent and interpretable basis for precision-agriculture decision-making.

## RESEARCH METHOD

The research method is designed to develop and optimize a multiclass classification-based crop recommendation model using an ensemble learning approach, as illustrated in Figure 1. The stages of the study include data collection, preprocessing, model development, and model evaluation.




Source: Research Results (2025)
Figure 1. Research Methodology

1. Dataset

   Agricultural data were collected from the publicly available Crop Recommendation Dataset on Kaggle, originally compiled to support machine learning applications in agriculture. The dataset consists of 2,200 records, each representing a unique set of environmental and soil parameters associated with a specific crop. The features include nitrogen (N), phosphorus (P), potassium (K), temperature (°C), humidity (%), pH, and rainfall (mm). The target variable, crop type, contains 22 distinct crop classes, covering a wide variety of grains, pulses, fruits, and vegetables such as rice, maize, chickpea, banana, and mango. This stage is essential to ensure the availability of representative data (Hasan et al., 2023).

   To evaluate model performance effectively, the dataset was randomly divided into training and testing subsets. A 80:20 split ratio was applied, where 80% of the data 1,760 samples were used for model training and 20% 440 samples were reserved for testing. This division ensured that each crop class was proportionally represented in both subsets, maintaining class balance for fair model evaluation.

2. Preprocessing

   The preprocessing stage was carried out to ensure that the dataset was clean, consistent, and suitable for model training. The process included data cleaning to eliminate duplicates, inconsistencies, and noise; normalization to standardize feature scales and improve model stability; and handling of missing values through imputation techniques (mean, median, mode, or predictive methods) or exclusion of records with excessive missing entries. In addition, data transformation such as one-hot encoding was applied to categorical features to ensure compatibility with machine learning algorithms. These steps were essential to improve dataset quality and ensure higher accuracy and reliability of the resulting model (Shahid et al., 2024).

3. Model Training

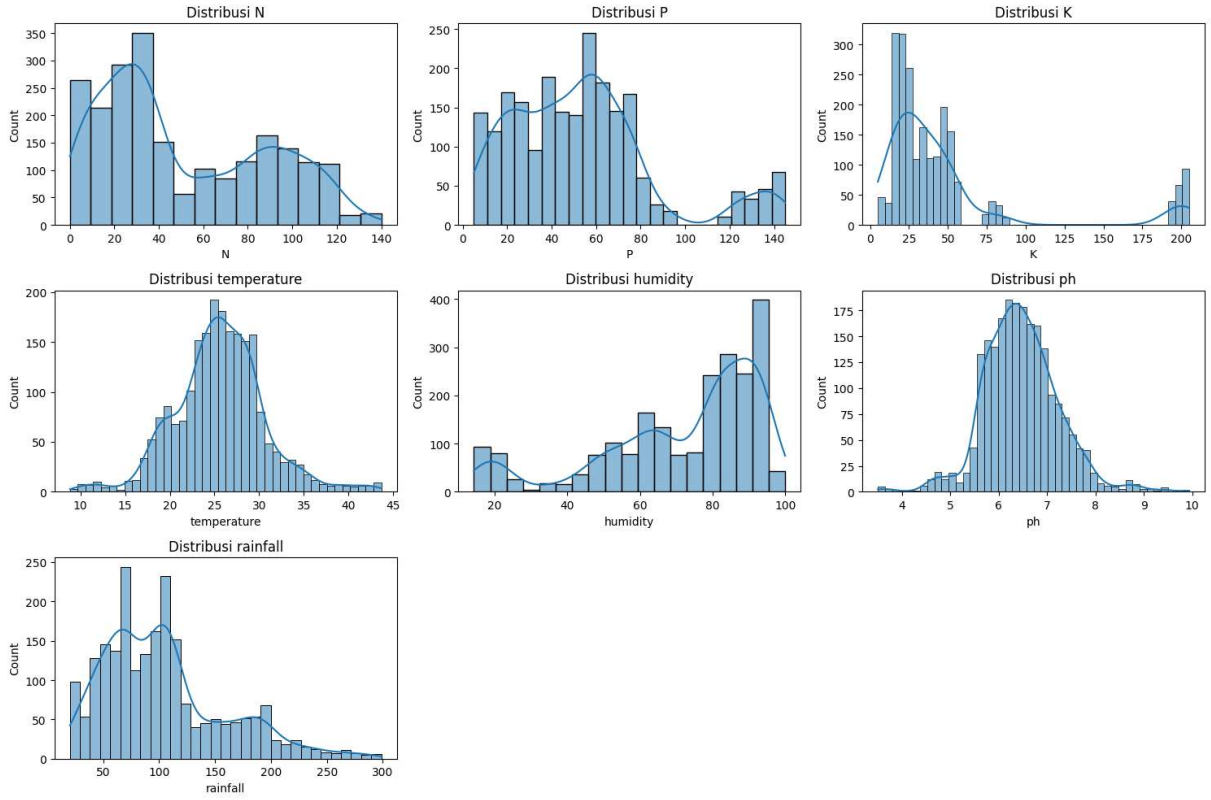
   The crop recommendation model was developed using several popular ensemble learning algorithms, namely Random Forest, which applies bagging based on decision trees; Extra Trees, which is similar to Random Forest but introduces more randomness in node splitting to increase variance, CatBoost, an optimized boosting method for handling categorical data, XGBoost, a widely used and efficient gradient boosting algorithm for

tabular data, and LightGBM, a high-efficiency boosting technique that employs histogram-based learning. Employing multiple algorithms allows for performance comparison, thereby facilitating the selection of the most optimal model for multiclass classification tasks.

4. Model Evaluation
   After training, the models were evaluated using multiclass classification metrics such as accuracy, precision, recall, F1-score, confusion matrix, and Macro-F1 for handling imbalanced data. This evaluation ensured that the models were not only accurate for majority classes but also fair in representing minority classes (Ramzan et al., 2023).
   With this method, the study is expected to produce a crop recommendation system that is accurate, adaptive, and capable of assisting farmers in determining the most suitable crop types for their land conditions.

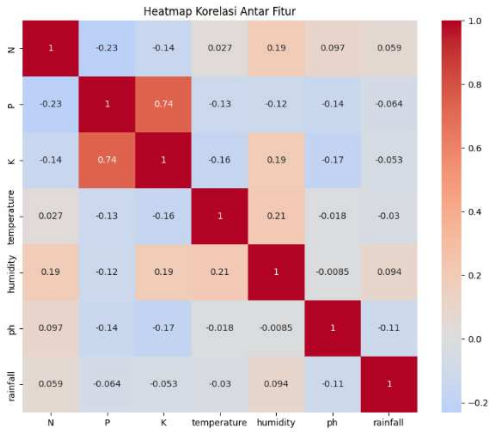
## RESULTS AND DISCUSSION

Variable distribution analysis was carried out to understand the characteristics of the data in developing the crop recommendation model, as shown in Figure 2. The results indicate that soil nutrients exhibit diverse distribution patterns: Nitrogen (N) is concentrated in the range of 20–40, Phosphorus (P) is relatively evenly distributed up to 80 with some anomalies above 100, and Potassium (K) is dominated by low values with high outliers reaching up to 200. Environmental variables show a more regular distribution: temperature approaches normal with an average of 25–27°C, humidity is high within the range of 60–100%, pH is nearly normal at 6–7, and rainfall is widely spread between 50–150 mm with some extreme values exceeding 250 mm. These findings emphasize that understanding the initial distribution is crucial for addressing non-normality and outliers, thereby improving the performance of ensemble learning models in providing optimal crop recommendations.



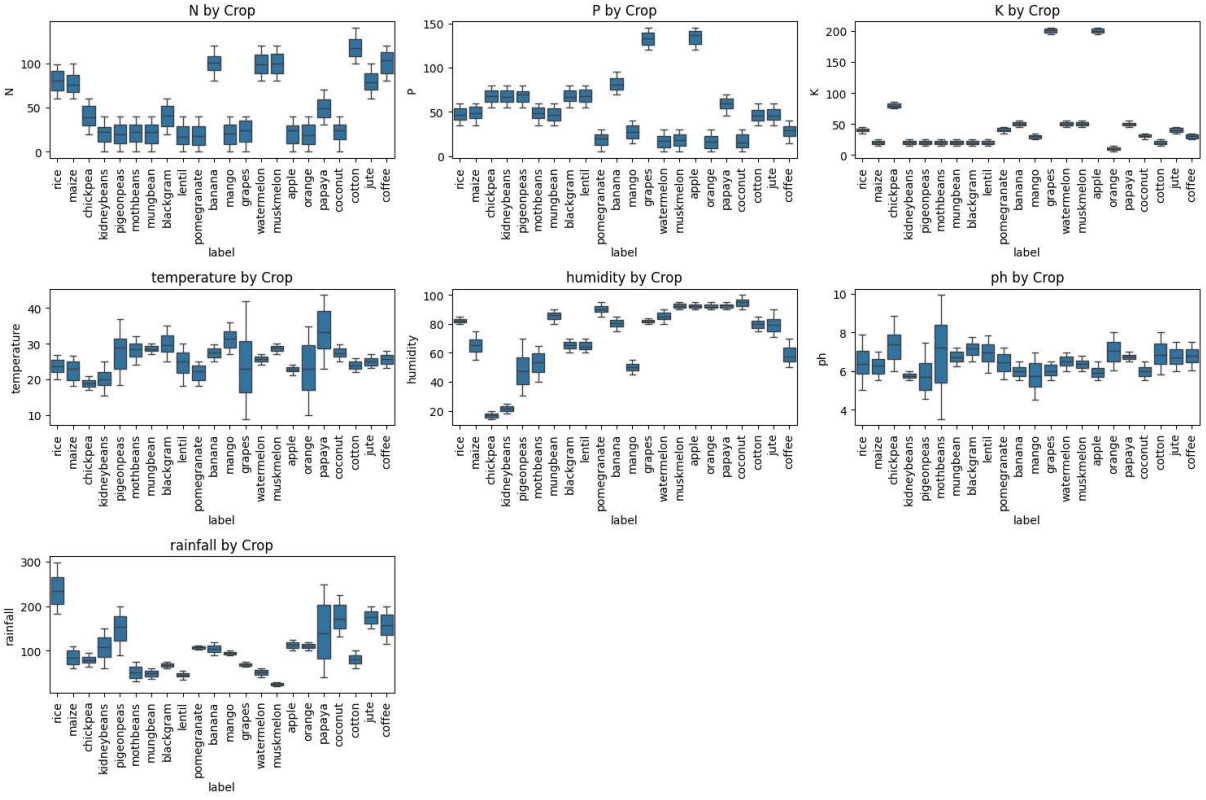

Source: Research Results (2025)
Figure 2. Variable Distribution

Correlation analysis among variables using Pearson's coefficient showed that most features have weak correlations, indicating that each variable may provide unique information for the crop recommendation model, as depicted in Figure 3. The strongest correlation was observed between Phosphorus (P) and Potassium (K) with a value of 0.74, suggesting a tendency to increase together. In contrast, Nitrogen (N) exhibited low and negative correlations with P and K (–0.23 and –0.14, respectively). Environmental variables such as temperature, humidity, pH, and rainfall did not show significant relationships, with the highest correlation being only 0.21 between temperature and humidity. These findings highlight the low multicollinearity among features, which is advantageous in machine learning as it reduces information redundancy and enhances the model's ability to capture complex patterns.




Source: Research Results (2025)
Figure 3. Feature Correlation Heatmap

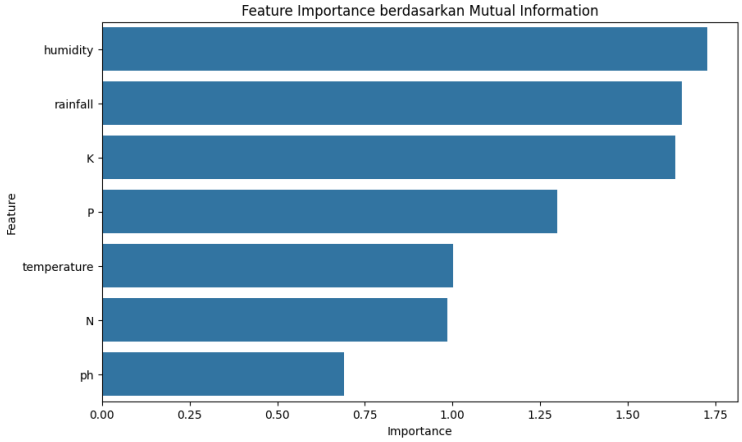
Variable distribution analysis by crop type using boxplots revealed differences in nutrient requirements and environmental conditions across commodities, as presented in Figure 4. Nitrogen (N) content was higher in paddy, maize, cotton, and coffee, while Phosphorus (P) and Potassium (K) were highest in grape and apple. Environmental factors also showed variation: tropical crops such as papaya, coconut, and mango thrive at higher temperatures, whereas apple grows better in cooler climates. High humidity was required by rice, coconut, and coffee, while pomegranate and lentil were more tolerant of low humidity. Soil pH was generally stable within 5.5–7, except for chickpea and lentil, which can grow in near-alkaline conditions. Rainfall requirements varied greatly, with rice, coconut, and coffee needing more than 200 mm, whereas mungbean, mothbeans, and pomegranate grow well with less than 100 mm. These findings confirm that each crop has distinct ecological and nutritional needs, underscoring the importance of developing machine learning–based recommendation models capable of recognizing the unique characteristics of each plant.




Source: Research Results (2025)
Figure 4. Distribution of Soil Nutrients and Environmental Factors by Crop Type

Feature importance analysis using the mutual information method revealed that environmental variables were more dominant than soil nutrients in crop classification, as illustrated in Figure 5. Humidity (1.726) and rainfall (1.654) emerged as the most influential factors, followed by Potassium (K) (1.636) and Phosphorus (P) (1.299). Temperature (1.002) also played an important role in distinguishing tropical and subtropical crops,

whereas Nitrogen (N) (0.986) and pH (0.690) had relatively lower influence. These findings confirm that agroclimatic factors play a greater role than soil chemical properties in determining crop suitability, consistent with previous studies highlighting the importance of climatic conditions in land mapping using machine learning.




Source: Research Results (2025)
Figure 5. Feature importance by Mutual Information

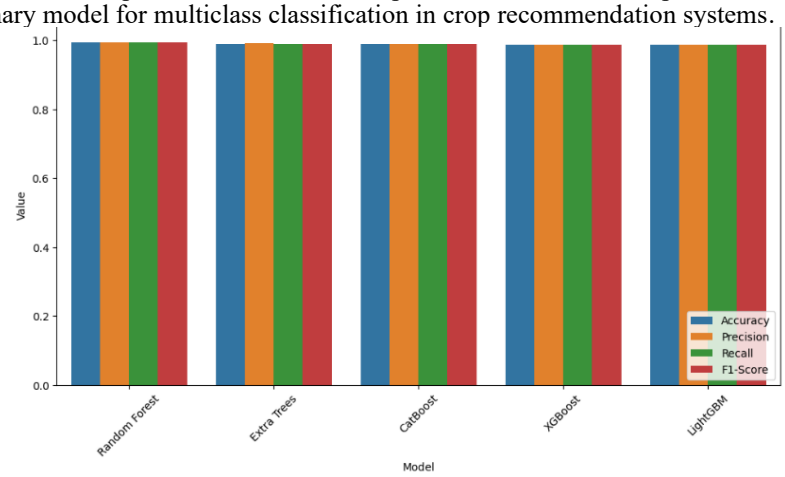
The feature importance analysis using Mutual Information showed that agroclimatic variables, particularly humidity and rainfall, had the greatest influence on crop classification, followed by Potassium (K) as the main soil nutrient factor. Phosphorus (P) ranked at a medium level, while temperature and Nitrogen (N) contributed less. Soil pH demonstrated the smallest influence. Overall, climatic factors proved to be more dominant than soil chemical properties, consistent with previous studies highlighting the importance of agroclimatic conditions in land suitability mapping and crop recommendation using machine learning.

Table 1. Evaluation Results of Ensemble Learning Models

| Model | Accuracy | Precision | Recall | F1Score |
|---|---|---|---|---|
| Random Forest | 0.993182 | 0.993735 | 0.993182 | 0.993175 |
| Extra Trees | 0.988636 | 0.990461 | 0.988636 | 0.988787 |
| CatBoost | 0.988636 | 0.989808 | 0.988636 | 0.988698 |
| XGBoost | 0.986364 | 0.986901 | 0.986364 | 0.986347 |
| LightGBM | 0.986364 | 0.987666 | 0.986364 | 0.986428 |

Source: Research Results (2025)

The evaluation results show that all five ensemble learning algorithms (Random Forest, Extra Trees, CatBoost, XGBoost, and LightGBM) achieved excellent performance with accuracy above 98%, as shown in Table 1. Random Forest emerged as the best model, with an Accuracy of 0.993, Precision of 0.9937, Recall of 0.9931, and F1-Score of 0.9931, demonstrating very low classification error. Its superiority is supported by the bagging mechanism and random feature selection, which enhance generalization. Extra Trees and CatBoost followed with accuracy close to 0.989, while XGBoost and LightGBM achieved around 0.986. The performance differences were mainly influenced by the algorithms' sensitivity to parameters and data structure. Overall, although all models were competitive, Random Forest proved to be the most superior and stable, making it the recommended primary model for multiclass classification in crop recommendation systems.




Source: Research Results (2025)
Figure 6. Comparison of Machine Learning Model Performance

The performance comparison of five ensemble learning models demonstrates that all algorithms (Random Forest, Extra Trees, CatBoost, XGBoost, and LightGBM) achieved high accuracy with evaluation metrics close to 1.0, as depicted in Figure 6. Random Forest consistently showed a slight advantage in Accuracy, Precision, and Recall, highlighting the effectiveness of the bagging mechanism in enhancing generalization. Extra Trees and CatBoost exhibited performance nearly equivalent to Random Forest, while XGBoost and LightGBM remained competitive though slightly lower. Overall, all models proved reliable for multiclass classification, but Random Forest emerged as the most optimal choice for the crop recommendation system.

Table 2. Feature Importance Analysis using Random Forest Model

| Feature | Importance |
|---|---|
| rainfall | 0.227036 |
| humidity | 0.211279 |
| K | 0.181222 |
| P | 0.143622 |
| N | 0.108859 |
| temperature | 0.075682 |
| ph | 0.052301 |

Source: Research Results (2025)

The feature importance analysis using Random Forest, as presented in Table 2, revealed that climatic factors, particularly rainfall (0.227) and humidity (0.211), exert the greatest influence on crop classification, followed by the soil nutrient potassium (K) (0.181). Phosphorus (P) and nitrogen (N) contributed at a moderate level, while temperature (0.076) and soil pH (0.052) played the least significant roles. These findings confirm that climatic factors are more dominant than soil properties in determining crop suitability and further demonstrate the capability of Random Forest to capture complex patterns, thereby enabling more accurate recommendations.

## CONCLUSION

The findings of this study demonstrate that climatic factors, particularly humidity and rainfall, exert the greatest influence on crop classification, followed by soil nutrients such as potassium (K) and phosphorus (P). In contrast, nitrogen (N), temperature, and pH show relatively minor effects. Among the ensemble learning algorithms tested, Random Forest achieved the best and most stable performance, with accuracy, precision, recall, and F1-score values reaching approximately 0.993, outperforming other models in the range of 0.986–0.988. These results emphasize that climatic conditions should be prioritized in data collection and modeling for crop recommendation systems. Furthermore, the optimized Random Forest model demonstrates high reliability and interpretability, providing a practical foundation for data-driven decision-making in precision agriculture, including land-use planning, resource optimization, and sustainable crop cultivation.

Future research should focus on improving the generalization and adaptability of ensemble models across diverse climatic and geographical regions by incorporating larger, multi-regional datasets. Integrating real-time climatic data, remote sensing imagery, and Internet of Things (IoT)-based agricultural monitoring can enhance model responsiveness and scalability. Additionally, exploring hybrid ensemble frameworks that combine Random Forest with deep learning or metaheuristic optimization techniques could further improve predictive accuracy and interpretability. Continued investigation into explainable AI methods is also essential to ensure that future crop recommendation systems remain transparent, user-friendly, and applicable for smart, sustainable agricultural decision support.

## ACKNOWLEDGEMENT

This research was funded by the Bina Sarana Informatika University Foundation under the 2025 Foundation Research Grant Scheme. The authors would like to express their sincere gratitude to the foundation for its support and trust in the implementation of this study. This support has been instrumental in advancing research efforts that contribute to the development of knowledge and the application of data-driven technologies in the field of agriculture.

## REFERENCES

Arumugam S. S. L., D., R., P. K., B., M., & V., A. (2024). Crop Recommendation using XG Boost Algorithm for Sustainable Agrarian Application. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3), 306–311. https://ijisae.org/index.php/IJISAE/article/view/5253

Daniel, C. (2024). A robust LightGBM model for concrete tensile strength forecast to aid in resilience-based structure strategies. *Heliyon*, *10*(20), e39679. https://doi.org/10.1016/j.heliyon.2024.e39679

Elshewey, A. M., Selem, E., & Abed, A. H. (2025). Improved CKD classification based on explainable artificial intelligence with extra trees and BBFS. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-02355-7

Food and Agriculture Organization (FAO). (2021). *Climate-Smart Agriculture Sourcebook*. https://openknowledge.fao.org/server/api/core/bitstreams/b21f2087-f398-4718-8461-b92afc82e617/content

Haddouchi, M., & Berrado, A. (2024). *A survey and taxonomy of methods interpreting random forest models*. http://arxiv.org/abs/2407.12759

Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M. I., Kardy, S., Ma, S., & Nam, Y. (2023). Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. *Frontiers in Plant Science*, *14*. https://doi.org/10.3389/fpls.2023.1234555

Hussein, N., & R. M. Zeebaree, S. (2024). Performance Evaluation of Extra Trees Classifier by using CPU Parallel and Non-Parallel Processing. *Indonesian Journal of Computer Science*, *13*(2). https://doi.org/10.33022/ijcs.v13i2.3802

Meiriyama, M., & Sudiadi, S. (2022). Penerapan Algoritma Random Forest Untuk Klasifikasi Jenis Daun Herbal. *Jurnal Teknologi Sistem Informasi*, *3*(1), 131–138. https://doi.org/10.35957/jtsi.v3i1.3176

Nguyen, K.-T., Tran, T.-N., & Nguyen, H.-T. (2024). Research on the Influence of Hyperparameters on the LightGBM Model in Load Forecasting. *Engineering, Technology & Applied Science Research*, *14*(5), 17005–17010. https://doi.org/10.48084/etasr.8266

Patel, K., & B. Patel, H. (2023). Multi-criteria Agriculture Recommendation System using Machine Learning for Crop and Fertilizesrs Prediction. *Current Agriculture Research Journal*, *11*(1), 137–149. https://doi.org/10.12944/carj.11.1.12

Prity, F. S., Hasan, MD. M., Saif, S. H., Hossain, Md. M., Bhuiyan, S. H., Islam, Md. A., & Lavlu, M. T. H. (2024). Enhancing Agricultural Productivity: A Machine Learning Approach to Crop Recommendations. *Human-Centric Intelligent Systems*, *4*(4), 497–510. https://doi.org/10.1007/s44230-024-00081-3

Rahman, A., Udjulawa, D., & Mulyati. (2024). Rekomendasi Pemilihan Jenis Tanaman Menggunakan Algoritma Random Forest dan XGBoost Regressor. *Computer Science (CO-SCIENCE)*, *4*(2). https://doi.org/10.31294/coscience.v5i1

Ramzan, Z., Asif, H. M. S., & Shahbaz, M. (2023). Multimodal crop cover identification using deep learning and remote sensing. *Multimedia Tools and Applications*, *83*(11), 33141–33159. https://doi.org/10.1007/s11042-023-17140-9

Shahid, M. F., Khanzada, T. J. S., Aslam, M. A., Hussain, S., Baowidan, S. A., & Ashari, R. B. (2024). An ensemble deep learning models approach using image analysis for cotton crop classification in AI-enabled smart agriculture. *Plant Methods*, *20*(1). https://doi.org/10.1186/s13007-024-01228-w

Srinivasu, P. N., Jaya Lakshmi, G., Gudipalli, A., Narahari, S. C., Shafi, J., Woźniak, M., & Ijaz, M. F. (2024). XAI-driven CatBoost multi-layer perceptron neural network for analyzing breast cancer. *Scientific Reports*, *14*(1), 28674. https://doi.org/10.1038/s41598-024-79620-8

Sudianto, S., & Cahyadi, E. F. (2025). A Soil Nutrient Assessment for Crop Recommendation Using Ensemble Learning and Remote Sensing. *International Journal of Intelligent Systems and Applications*, *17*(3), 34–47. https://doi.org/10.5815/ijisa.2025.03.03

Ying, C., Shi, A., & Li, X. (2025). Hybrid boosted attention-based LightGBM framework for enhanced credit risk assessment in digital finance. *Humanities and Social Sciences Communications*, *12*(1). https://doi.org/10.1057/s41599-025-05230-y