Journal of Engineering Science and Technology
Special Issue on INCITEST2025 Vol. 21, No. 2 (2026) 33 - 40

A LIGHTWEIGHT COMPRESSED HYBRID RF-XGBOOST ALGORITHM FOR IOT-BASED CROP RECOMMENDATION SYSTEMS

MIMI HARYANI TASANI[1], MUHAMMAD ARIA RAJASA POHAN[2], SENNY LUCKYARDI[3], MOHD KAMIR YUSOF[1]

[1]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Malaysia
[2]Department of Robotics and Artificial Intelligence Engineering, Universitas Komputer Indonesia, Indonesia
[3]Department of Management, Universitas Komputer Indonesia, Indonesia
*Corresponding Author: mimiharyanitasani95@gmail.com

Abstract

This study focused on compressed random forest with XGBoost (CRF+XGB) for internet of things (IoT)-based crop recommendation. The method combines the high performance of random forest (RF) with the high-speed computing capacity of extreme gradient boosting (XGB). RF identifies the most significant environmental and soil factors while XGB interpret subtle relationships. The hybridization yields a compact model with minimal accuracy loss. The RF+XGB hybrid achieved 99.55% accuracy and the lightweight CRF+XGB is 98.86% with significant reduction in inference time and memory usage. The purpose of study is to develop a highly accurate, resource-efficient model for real-time agriculture systems, recognising that IoT devices operate under strict computational constraints. SHAP analysis illustrates both models incorporate important features such as rainfall, soil moisture, and nutrients (N, P, K), and thus they are straightforward to interpret. The model provides an efficient, scalable, and practical solution that balances performance and resource use for real agricultural application.

Keywords: Compression model, Crop recommendation, Hybrid model, Precision agriculture, Internet of things.

1. Introduction

The agriculture sector faces increasing challenges to boost food production, keep operations sustainable, and achieve economic growth. Precision agriculture integrations have been introduced using internet of things (IoT) sensors, cloud computing and artificial intelligence (AI) for solutions [1, 2]. Machine learning (ML) models such as Decision Trees, RF, and gradient boosting methods like XGB are able to predict soil health, crop yield, and suitability of crops, but these models typically require heavy-duty computational capabilities, which limit them from being deployed on small IoT or edge devices [1, 3, 4].

Previous works were succeeded in simulation environments such as hydroponic systems but had high complexity and low portability in real-time field use [4-7]. To counter such limitations, this paper introduces CRF+XGB which provides precise, efficient, and explainable agronomic crop recommendations even when run on hardware-constrained devices. The approach combines the best of RF and XGB by integrating their respective benefits, optimized with the model compression [8]. Transparency and interpretability are maintained using shapely additive explanations (SHAP) analysis, even after compression.

The novelty of this work is firstly, the integration of RF and XGB using a stacking approach to take advantage of their complementary strengths. Secondly, knowledge distillation has been applied to compress the hybrid ensemble into a lightweight model. Thirdly, a thorough evaluation is performed across accuracy, memory usage, inference speed, and SHAP interpretability. This study offers an efficient and easily deployable solution for data-driven crop recommendation in precision agriculture.

2. Literature Review

Machine learning (ML) and AI have revolutionized agriculture into data-driven systems [9-11]. With IoT, cloud computing, and real-time computation of data, farmers are able to track environment and soil parameters to make sustainable decisions [12, 13]. Integrating IoT with ML models, improve agricultural harvests, utilize computing and other resources more efficiently, and improve environment well-being [14-16]. However, it is challenging to achieve the proper balance between knowledge, accuracy, and computation speed, especially in implementations of IoT from extremely low-power edge devices with very low processing capabilities [17-19]. Traditional models like decision Trees, SVMs, and RF have been applied widely but struggle with nonlinear soil-climate relationships [20]. Gradient boosting, like XGB, overcomes these drawbacks by identifying richer patterns, which increase higher computational complexity and long train duration [14].

Recent advances have brought IoT-based ML into real-time AIoT systems, enabling adaptive irrigation and crop management [16, 21]. However, many high-performance ML models remain too computationally demanding for embedded IoT devices [13, 19]. Although some studies explored RF–XGB hybrids for combined interpretability and accuracy, research on compressing such hybrids for resource-constrained IoT settings is scarce [4]. Key gaps include limited attention to efficiency and interpretability, few IoT-oriented compression methods, and minimal work on heterogeneous RF–XGB compression [9-13, 20]. The proposed CRF+XGB framework directly addresses these gaps by combining ensemble strengths with compression for an interpretable, scalable IoT solution. The three key studies that inform the current research are summarised in Table 1 where the methodologies, performance results, and limitations are noted. The current study seeks to address these limitations by creating an efficient, interpretable and IoT-ready system.

Table 1. Comparison of related studies.

Ref.	Model and Dataset	Limitations
[1]	RF (IoT soil & weather data) & IoT-Based Precision Farming	Data collected offline; not adaptable to real-time environmental changes.
[3]	XGB (AIoT hydroponic monitoring) & Hydroponic Crop System	IoT systems require continuous data connectivity; model scalability remains a challenge.
[4]	ML for nutrient & water uptake & Hydroponic Soybean Growth	Evaluation conducted in controlled lab environment, not real farms.

3. Research Methodology

The CRF+XGB framework follows five phases: preprocessing, base-learner training, hybrid construction, compression, and performance analysis. The dataset has 2,200 samples with seven features across 22 crop classes [1]. After LabelEncoder cleaning, 80/20 splitting, and Min-Max scaling, RF (n_estimators = 200) and XGB (n_estimators = 200, learning_rate = 0.05) were stacked using Logistic Regression meta-learning [22, 23].

Formally, the hybrid model is constructed as follows: Step 1 is train base models, RF ($RF\ model{:}\ \hat{y}_{RF} = RF.fit\ (X_{train}, y_{train})$) and XGB ($XGB\ model{:}\ \hat{y}_{XGB} = XGB.fit\ (X_{train}, y_{train})$) [1, 3]. For any dataset X, their outputs are ($P_{RF} = RF.predict\ (X),\ P_{XGB} = XGB.predict\ (X)$) where P_{RF} and P_{XGB} are vectors contains class predictions or probabilities. Step 2, prepare meta-model training data, meta features are formed as ($X_{meta} = [P_{RF}\ \ P_{XGB}]$, $y_{meta} = [y_{train}]$). Each row of X_{meta} has the base-model predictions for the same sample. Step 3, train the meta-learner, logistic regression classifier is trained ($f_{meta}.fit = (X_{meta}, y_{meta})$) for classifications like logistic regression, simple neural network, linear regression or another tree-based model. Step 4, hybrid predictions, entailed a new dataset X_{test} to first get base predictions ($P_{RF}^{test} = RF.predict(X_{test})$, $P_{XGB}^{test} = XGB.predict(X_{test})$), form the meta-input ($X_{meta}^{test} = [\ P_{RF}^{test}\ \ P_{XGB}^{test}\ ]$), and then attain the final hybrid prediction ($\hat{y}_{hybrid} = f_{meta}(\ X_{meta}^{test})$).

The relationship is summarized as ($\hat{y}_{hybrid} = f_{meta}(P_{RF}, P_{XGB})$), where ($P_{RF} = RF.predict(X)$ and $P_{XGB} = XGB.predict(X)$) the class probability distributions produced by RF and XGB are integrated. This configuration allows the hybrid model to combine the robustness of RF with the high predictive accuracy of XGB. After training, the hybrid ensemble was compressed through knowledge distillation, in which the hybrid model was induced as the teacher and a smaller XGB model was induced as the student [24, 25]. Knowledge distillation minimizes the Kullback-Leibler divergence between the teacher's and student's predicted probability distributions ($L_{KD} = \sum_i \sum_j T_i^{(j)} \, log \frac{T_i^{(j)}}{S_i^{(j)}}$). The teacher's predicted probability for class in an instance is directly related to the student prediction. By this optimization, the compressed model learns to represent the teacher's decision boundaries using fewer parameters and lower computational cost [26].

Model performance was assessed using accuracy, precision, recall, F1-score, log Loss, inference time, and memory usage. SHAP explainability confirmed that hybrid and compressed models preserved the same feature importance patterns with humidity, rainfall, and soil nutrients (N, P, K). The full process is outlined in Fig. 1 [15].

Input: Crop dataset D with features (N, P, K, temperature, humidity, pH, rainfall)	
Preprocess D	clean, encode, split (80/20), normalize.
Train base models	RF_model – train (RandomForest, D_train) XGB_model – train (XGBoost, D_train)
Generate hybrid	RF_probs, XGB_probs – predict_proba(D_train) meta_input – concatenate (RF_probs, XGB_probs) Hybrid_model – train (LogisticRegression, meta_input)
Compress hybrid	(Knowledge Distillation) soft_labels Hybrid_model.predict_proba(D_train) weights – max (soft_labels, axis=1) Compressed_model – train (XGBoost, D_train, labels=soft_labels, sample_weights=weights)
Evaluate	Compare metrics: Accuracy, F1, Log Loss, Time, Memory. Compute SHAP values for Hybrid_model and Compressed_model
Output: Predicted crop recommendation (IoT deployable model)	

Fig. 1. Pseudocode for the CRF+XGB framework.

In summary, the proposed CRF+XGB framework couples ensemble learning and model compression well to create an effective, accurate, and interpretable solution for IoT-based crop recommendation. The workflow is able to deliver high prediction accuracy, low computational cost, and understandable reasoning. It is thus practical for real-world agricultural decision support systems.

4. Results and Discussion

4.1. Quantitative evaluation

Figure 2 shows the hybrid (RF+XGB) model achieved 99.55% accuracy, while CRF+XGB model achieved 98.86% with 0.7% less while running ten times faster and consuming 90% less memory, in comparison to the previous study using the same dataset, making it suitable for IoT [27, 28]. These results are better than previous works [1, 3] with higher accuracy [4]. The approach offers an efficient, interpretable, and real-time functionality for low resource agriculture [29].

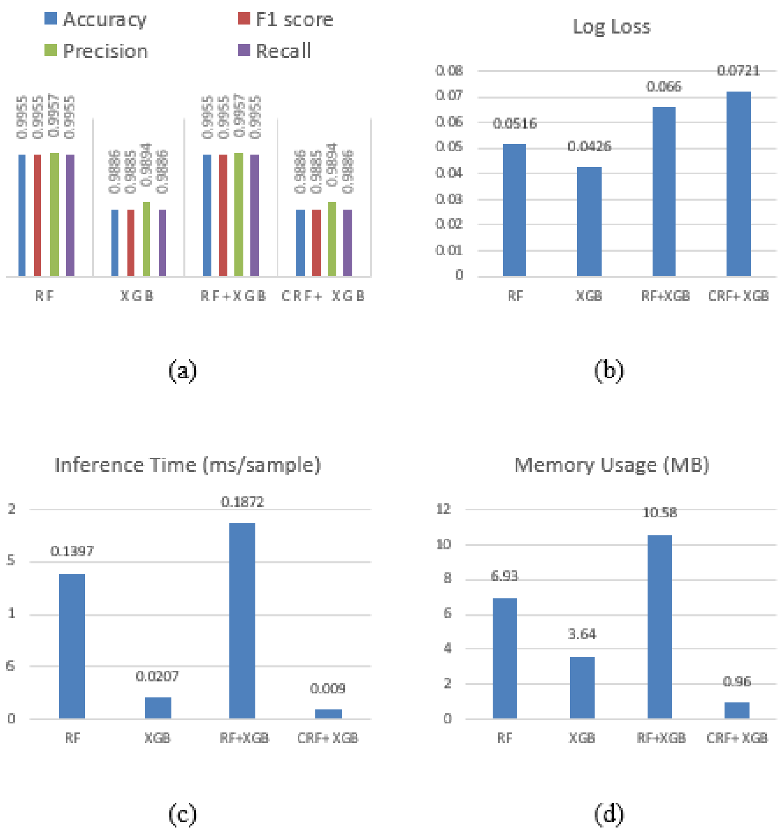

Fig. 2. Performance and efficiency comparison among RF, XGB, Hybrid and CRF+XGB models; (a) Performance, (b) Log loss, (c) Inference time (ms/sample), (d) Memory usage (MB).

4.2. Interpretability and model comparison

SHAP explainability was used to identify how each input contributed to predictions. The feature-importance patterns from the compressed model were very similar to the hybrid ensemble, which confirms that interpretability is well preserved. The result of SHAP has shown that the most influencing factors are soil nutrients, including N, P, and K, rainfall, and humidity. This agreed with the findings of previous studies. Spearman correlations with hybrid and XGB models revealed perfect ranking agreement, $\rho = 1.00$, in proof that the compressed model maintains accuracy and transparency for precision agriculture. Figure 3: A perfect SHAP correlation that CRF+XGB maintains both explainability and efficiency, hence bridging the gap between the experimental ML system and real-world agricultural practice.

While other interpretable techniques have been explored in agricultural application domains, such as LIME and permutation importance, SHAP was used in this study because it is founded on a stronger theoretical framework and can provide consistent additive feature attributions [15, 17, 25].

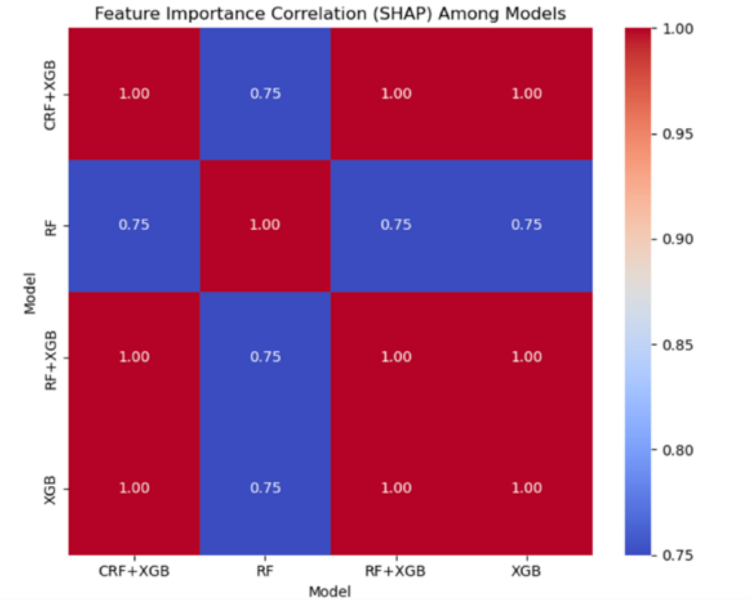

Fig. 3. SHAP feature importance correlation (SHAP) among models.

4.3. Discussion summary

Experiments showed the hybrid RF+XGB model achieved the highest accuracy, while the compressed CRF+XGB model matched its performance with faster processing and lower memory use. Both models produced identical feature-importance rankings, preserving interpretability. Overall, CRF+XGB offers an accurate, efficient, and practical solution for large-scale IoT-based intelligent crop recommendation.

5. Conclusion

This paper proposes a hybrid compression model, CRF with XGB, for IoT based precision agriculture. It combines the strengths of RF and XGB to perform with high accuracy, interpretability and lightweight deployment. Hybrid RF+XGB obtained 99.55% accurate and CRF+XGB was 98.86%, with faster inference and lower memory use. SHAP analysis indicated that both models utilized similar important features such as humidity, rain, and soil nutrient components such as N, P, and K for transparent and interpretable decision patterns. Proposed future research includes employing the CRF+XGB model on low power devices such as Raspberry Pi or ESP32 to measure real time performance to improve data privacy, flexibility, and scalability across a wide range of agricultural environments.

References

1. Akhter, R.; and Sofi, S.A. (2022). Precision agriculture using IoT data analytics and machine learning. *Journal of King Saud University-Computer and Information Sciences*, 34(10), 9319-9330.
2. Afrianto, I.; Mainusar, R.; and Atin, S. (2024). Integration of wireless sensor networks with wireless mesh networks in IoT-based patient monitoring systems. *Journal of Engineering Science and Technology*, 19(5), 1660-1671.

3. Rahman, M.A.; Chakraborty, N.R.; Sufiun, A.; Banshal, S.K.; and Tajnin, F.R. (2024). An IoT-based hydroponic system for crop recommendation and nutrient parameter monitorization. *Smart Agricultural Technology*, 8, 100472-100486.
4. Dhal, S.B.; Mahanta, S.; McClurkinMoore, J.; and Kalafatis, S. (2024). Machine learning-based analysis of nutrient and water uptake in hydroponically grown soybeans. *Scientific Reports*, 14, 24337-24351.
5. Saputra, H.; Sumitra, I.D.; Hirawan, D.; Lesmana, R.; and Soegoto, E.S. (2023). Smart urban farming application: UV light in hydroponic installations. *Journal of Engineering Science and Technology*, 18(2), 1007-1018.
6. Choi, J.W.; Hidayat, M.S.; Cho, S.B.; Hwang, W.H.; Lee, H.; Cho, B.K.; and Kim, G. (2025). Recent trends in machine learning, deep learning, ensemble learning, and explainable artificial intelligence techniques for evaluating crop yields under abnormal climate conditions. *Plants*, 14(18), 2841-2866.
7. Arısoy, A.; and Açıkgözoğlu, E. (2024). Optimizing soil fertility through machine learning: Enhancing agricultural productivity and sustainability. *Bilge International Journal of Science and Technology Research*, 8(2), 124-133.
8. Kumari, P.; Chandan, G.Y.; Agarwal, D.; Joshi, S.; and Vekariya, P. (2024). Combining predictive models: Hybrid approaches in crop recommendation systems. *International Journal of Research in Agronomy*, 7(11), 108-115.
9. Condran, S.; Bewong, M.; Islam, M.Z.; Maphosa, L.; and Zheng, L. (2022). Machine learning in precision agriculture: A survey on trends, applications and evaluations over two decades. *IEEE Access*, 10, 73786-73803.
10. Warlina, L.; Soegoto, E.S.; Supatmi, S.; Oktafiani, D.; and Jatnika, R. (2023). Regional competitive advantage of agriculture as the leading sector in Garut Regency, West Java Province, Indonesia. *Journal of Eastern European and Central Asian Research*, 10(1), 1-20.
11. Nursikuwagus, A.; Purwanto, H.; and Dewi, D.A. (2023). Feature classification based on heterogenous data using hybrid machine learning: A review. *Journal of Theoretical and Applied Information Technology*, 101(17), 6991-7010.
12. Choudhary, V.; Guha, P.; Pau, G.; and Mishra, S. (2025). An overview of smart agriculture using Internet of Things (IoT) and web services. *Environmental and Sustainability Indicators*, 26, 100607-100624.
13. Miller, T.; Mikiciuk, G.; Durlik, I.; Mikiciuk, M.; Łobodzińska, A.; and Śnieg, M. (2025). The IoT and AI in agriculture: The time is now—A systematic review of smart sensing technologies. *Sensors*, 25(12), 3583-3614.
14. Oikonomidis, A.; Catal, C.; and Kassahun, A. (2022). Hybrid deep learning-based models for crop yield prediction. *Applied Artificial Intelligence*, 36(1), 2031822-2031840.
15. Hussein, E.E.; Zerouali, B.; Bailek, N.; Derdour, A.; Ghoneim, S.S.; Santos, C.A.G.; and Hashim, M.A. (2024). Harnessing explainable AI for sustainable agriculture: SHAP-based feature selection in multi-model evaluation of irrigation water quality indices. *Water*, 17(1), 59-92.
16. Ahmed, R.M. (2024). Integration of wireless sensor networks, Internet of Things, artificial intelligence, and deep learning in smart agriculture: A

comprehensive survey. *Journal of Innovative Intelligent Computing and Emerging Technologies*, 1(01), 8-19.

17. Hassan, M.; Rahman, S.; and Karim, M. (2023). Comparative analysis of model interpretability techniques for agricultural AI systems. *Information Processing in Agriculture*, 10(3), 514-527.
18. Patil, V.; Kulkarni, N.; and Sharma, R. (2023). Data-driven approaches for precision agriculture: A comprehensive review. *Environmental Modelling and Software,* 162, 105601-105650.
19. Kalyani, Y.; and Collier, R. (2021). A systematic survey on the role of cloud, fog, and edge computing combination in smart agriculture. *Sensors,* 21(17), 5922-5948.
20. Bouni, M.; Hssina, B.; Douzi, K.; and Douzi, S. (2024). Integrated IoT approaches for crop recommendation and yield-prediction using machine learning. *Internet of Things*, 5(4), 1-16.
21. Ramli, M.I.; Ariffin, M.A.M.; Zainol, Z.; Amin, M.N.M.; Hirawan, D.; Sumitra, I.D.; and Jamil, N. (2023). Design of a smart portable farming kit for indoor cultivation using the Raspberry Pi platform. *Pertanika Journal of Science and Technology*, 31(4), 1-24.
22. Xu, C.; Gao, W.; and Li, T. (2023). Teacher-student collaborative knowledge distillation for image classification. *Applied Intelligence*, 53, 1997-2009.
23. Halder, K.; Srivastava, A.K.; Ghosh, A.; Nabik, R.; Pan, S.; Chatterjee, U.; and Islam, M.K. (2024). Application of bagging and boosting ensemble machine learning techniques for groundwater potential mapping in a drought-prone agriculture region of eastern India. *Environmental Sciences Europe*, 36(1), 155-182.
24. Safaei, N.; Safaei, B.; Seyedekrami, S.; Talafidaryani, M.; Masoud, A.; Wang, S.; and Moqri, M. (2022). E-CatBoost: An efficient machine learning framework for predicting ICU mortality using the eICU collaborative research database. *PLOS One*, 17(5), e0262895-e0262927.
25. Mishra, V.; Shen, H.; and Yadav, N. (2024). Knowledge distillation with ensemble calibration. *ACM Transactions on Intelligent Systems and Technology*, 15(2), 1-13.
26. Xu, Q.; Chen, Z.; Ragab, M.; Wang, C.; Wu, M.; and Li, X. (2022). Contrastive adversarial knowledge distillation for deep model compression in time-series regression tasks. *Neurocomputing*, 485, 242-251.
27. Parez, S.; Dilshad, N.; Alanazi, T.M.; and Lee, J.W. (2023). Towards sustainable agricultural systems: A lightweight deep learning model for plant disease detection. *Computer Systems Science and Engineering*, 47(1), 515-536.
28. Wang, T.; Guo, J.; Zhang, B.; Yang, G.; and Li, D. (2025). Deploying AI on edge: Advancement and challenges in edge intelligence. *Mathematics*, 13(11), 1878-1898.
29. Gelete, G. (2023). Hybrid extreme gradient boosting and nonlinear ensemble models for suspended sediment load prediction in an agricultural catchment. *Water Resources Management*, 37(14), 5759-5787.