Sensor-Driven Ensemble Learning for Crop Recommendation and Disease Prediction in Precision Agriculture

Gerasimos Vonitsanos*, Emmanouela-Electra Economopoulou*,Spyros Sioutas*, Andreas Kanavos† and Phivos Mylonas‡
*Computer Engineering and Informatics Department
University of Patras, Patras, Greece
mvonitsanos, std1057466, sioutas@ceid.upatras.gr
†Department of Informatics
Ionian University, Corfu, Greece
akanavos@ionio.gr
‡Department of Informatics and Computer Engineering
University of West Attica, Athens, Greece
mylonasf@uniwa.gr

***Abstract*—Agriculture plays a vital role in ensuring global food security, yet it faces growing challenges from climate change, resource limitations, and increasing demand. This paper presents an ensemble learning framework that leverages Internet of Things (IoT) sensor data for crop recommendation and plant disease prediction in precision agriculture. Environmental parameters including temperature, humidity, rainfall, soil pH, and nutrient levels are modeled using Random Forest, Neural Networks, and Logistic Regression classifiers. Experimental evaluation on a publicly available dataset shows that Random Forest achieves superior performance, reaching 89.20% accuracy and the highest F1-score, outperforming baseline models in robustness and interpretability. The integration of Apache Spark enables scalable and near real-time analysis, making the approach suitable for practical deployment. By combining ensemble learning with sensor-driven environmental monitoring, the proposed framework supports sustainable, interpretable, and data-driven agricultural decision-making for farmers, researchers, and policymakers.**

***Index Terms*—Precision Agriculture, Internet of Things (IoT), Ensemble Learning, Crop Recommendation, Plant Disease Prediction, Random Forest, Machine Learning**

I. Introduction

The integration of sensor technologies into agricultural practices has transformed farming into a data-driven discipline, enabling unprecedented insights into crop health and disease management [12], [24]. Plant diseases remain a major threat to global food security, causing significant yield losses and affecting the livelihoods of millions of farmers. As the global population grows, ensuring sustainable food production has become an urgent priority. In this context, the convergence of the Internet of Things (IoT) and machine learning (ML) offers powerful opportunities to enhance crop yield prediction and disease prevention. These technologies can optimize farming practices, support rural economies, and promote sustainability by reducing chemical dependence while addressing climate-related challenges such as irregular rainfall, rising temperatures, and pest outbreaks.

Affordable IoT sensors, high-resolution satellite imagery, and scalable computational frameworks now enable continuous monitoring of soil moisture, temperature, humidity, and solar radiation, generating real-time data at unprecedented granularity [24]. When combined with ML algorithms, these data streams can be transformed into actionable insights, revealing hidden patterns and supporting precise, timely interventions. Such capabilities provide farmers and stakeholders with context-aware decision support, extending beyond traditional observational methods and improving both efficiency and sustainability in modern agriculture.

Nevertheless, the adoption of sensor-based disease prediction systems still faces challenges. Data privacy, cybersecurity, and ethical considerations are critical concerns in the design of agricultural IoT networks. Addressing these issues requires interdisciplinary collaboration among researchers, technologists, policymakers, and practitioners to ensure trustworthy, equitable, and practical solutions.

This work makes four main contributions. First, we evaluate the predictive performance of ensemble and baseline ML models—Random Forest, Logistic Regression, and Neural Networks—on a sensor-derived crop dataset. Second, we examine the impact of environmental factors such as soil nutrients, temperature, humidity, and rainfall on predictive accuracy. Third, we demonstrate the scalability of the approach through integration with big data frameworks such as Apache Spark. Finally, we propose a semantic-aware and deployable framework that unifies crop recommendation and disease risk prediction, advancing the practical implementation of precision agriculture systems.

979-8-3315-0450-2/24/$31.00 © 2024 IEEE

The remainder of this paper is organized as follows. Section II reviews related work on IoT and ML in precision agriculture. Section III presents the ML models applied, highlighting their mathematical foundations and suitability for agricultural prediction tasks. Section IV details the experimental evaluation, including dataset description, preprocessing and setup, evaluation metrics, and hyperparameter tuning. Section V discusses the results supported by comparative analysis. Finally, Section VI concludes the paper, highlighting key findings, practical implications, and directions for future research.

II. Related Work

The selection of related literature was guided by four key factors: (i) crop performance and yield prediction, (ii) the use and integration of machine learning (ML) and Internet of Things (IoT) technologies, (iii) evaluation metrics employed, and (iv) the diversity of techniques and methodologies applied across regions. The review was conducted in two stages. First, abstracts, introductions, and conclusions of relevant papers were screened to classify works by topic and relevance. Second, shortlisted sources were examined in detail with an emphasis on agriculture-specific applications that address yield optimization and disease prediction. Databases such as Elsevier, ScienceDirect, and MDPI were used as primary sources.

The need to enhance plant health and productivity has long attracted attention from both researchers and practitioners, particularly as globalization and population growth intensify pressure on sustainable food supply [18], [23]. Demand for higher quality and productivity indicators has fostered widespread adoption of agricultural management tools [18]. A broad family of sensing and information technologies—including satellite navigation, sensor networks, and ubiquitous computing—has been introduced to support data-driven monitoring and decision-making in farming [20]. Empirical studies confirm that sensor deployment and networking positively influence crop monitoring and resource allocation [7], [16]. This evolution has given rise to precision agriculture paradigms such as Smart Farming, Variable Rate Technology, GPS-guided cultivation, and site-specific crop management [9].

Recent advancements at the intersection of IoT and ML further extend these capabilities. IoT-based systems enable real-time monitoring of soil moisture, temperature, humidity, and rainfall, while ML algorithms transform such data into actionable predictions. Studies show that the integration of IoT with ML improves farm-management accuracy, resource optimization, and yield outcomes [3], [14]. Beyond yield prediction, artificial intelligence, sensing technologies, and robotics are increasingly employed in plant phenotyping and sustainable farming, offering more effective ways to manage environmental impacts [22]. Remote sensing modalities, such as UAVs and satellite imagery, have also been recognized as critical for scalable monitoring and precision management across diverse and complex environments [6].

Despite these advances, several limitations remain. Many works rely on small-scale or controlled datasets, constraining their applicability to real-world agricultural conditions. Others focus exclusively on either yield prediction or disease detection, without integrating both into a unified predictive framework. Furthermore, computational efficiency and scalability are often neglected, despite their importance for deployment in resource-constrained settings. Few approaches explicitly incorporate semantic or context-aware mechanisms to enhance interpretability and adaptability of predictions—an aspect particularly relevant to intelligent agricultural decision support.

Addressing these gaps, this study evaluates ensemble and baseline ML algorithms on sensor-derived agricultural data, incorporates scalable processing through Apache Spark, and advances a unified framework for both crop recommendation and disease prediction. By emphasizing semantic-aware, context-driven insights, the proposed approach aims to deliver practical, reliable, and adaptive solutions for farmers, researchers, and policymakers seeking to strengthen agricultural productivity and sustainability.

III. Machine Learning Models

This study employs three supervised learning algorithms—Neural Networks, Random Forests, and Logistic Regression—to predict crop suitability and plant disease occurrence from IoT sensor data. These models were selected for their complementary strengths: Neural Networks capture complex, non-linear relationships; Random Forests provide robustness and feature interpretability; and Logistic Regression offers simplicity and transparency.

A. Neural Networks

Artificial Neural Networks (ANNs) are inspired by the structure of the human brain and consist of interconnected nodes organized into input, hidden, and output layers. Each neuron applies a weighted sum of its inputs followed by a non-linear activation function, which enables the network to model complex mappings [5], [13]. In agricultural contexts, where environmental variables and sensor signals may interact in highly complex ways, ANNs are particularly effective in handling noisy, high-dimensional data from IoT monitoring systems [12]. Their general operation can be expressed as:

$$h^{(l)} = f(W^{(l)}h^{(l-1)} + b^{(l)}) \tag{1}$$

where $W^{(l)}$ and $b^{(l)}$ are the parameters of layer l, $h^{(l-1)}$ is the previous layer's output, and $f(\cdot)$ is a non-linear activation. In this study, ANNs are leveraged to detect subtle patterns in temperature, humidity, and soil conditions that may indicate early signs of plant stress. Their adaptability allows the same architecture to be applied across different crop types and environmental settings.

B. Random Forest

Random Forest is an ensemble algorithm that constructs multiple decision trees using bootstrapped training samples and random feature subsets [2]. For classification, the model aggregates predictions by majority voting:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \tag{2}$$

where $h_t(x)$ is the prediction of tree t and T is the total number of trees. RFs are robust to overfitting, require little preprocessing, and are particularly valuable in identifying influential environmental variables, thus supporting agronomic decision-making [17], [19]. In this study, RFs are used to highlight which environmental features—such as soil pH, rainfall, or nutrient levels—are most critical for predicting disease onset. Their interpretability makes them suitable for providing farmers with actionable insights into crop management.

C. Logistic Regression

Logistic Regression is a widely used classification algorithm that models the probability of a binary outcome using the logistic function [8]. It provides interpretable coefficients that quantify the influence of each variable on the outcome, making it a transparent baseline model for agriculture [15]. The probability of disease occurrence is estimated as:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}} \tag{3}$$

where w and b are the model parameters and x is the feature vector. In this work, LR serves as a baseline for distinguishing between healthy and diseased crops using sensor data. Its computational efficiency makes it well-suited for resource-constrained agricultural environments, such as small farms.

IV. Experimental Evaluation

This section presents the experimental evaluation of the ML models applied to plant disease prediction and crop recommendation using sensor-derived environmental data. By comparing three algorithms—Neural Networks, Random Forests, and Logistic Regression—we assess their ability to classify crop suitability under varying conditions such as temperature, humidity, and soil nutrients. The evaluation aims to determine which model provides the most accurate predictions, offering insights into the effectiveness of data-driven approaches in precision agriculture. Each model was assessed using standard evaluation metrics, including Accuracy, Precision, Recall, F1-Score, and the Concordance Index (C-Index).

A. Dataset

The experiments use the publicly available *Crop Recommendation Dataset* from Kaggle [11], which contains 2,200 records covering a wide range of crops, including cereals, pulses, fruits, and industrial crops (e.g., rice, maize, chickpea, kidneybeans, pigeonpeas, lentils, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee). Each record is annotated with environmental and soil features:

- **N**: Nitrogen content in the soil
- **P**: Phosphorus content in the soil
- **K**: Potassium content in the soil
- **temperature**: Temperature in degrees Celsius
- **humidity**: Relative humidity in percentage
- **ph**: Soil pH value
- **rainfall**: Rainfall in millimeters
- **label**: Recommended crop type

The dataset is balanced across classes, making it well-suited for testing classification models under realistic agricultural conditions. Data preprocessing involved normalization of environmental features to ensure consistent scaling. As no missing values were present, imputation was not required. Label encoding was applied to the categorical crop labels for compatibility with supervised learning algorithms.

B. Experimental Setup and Preprocessing

All experiments were conducted using Python 3.9 with Scikit-learn and TensorFlow libraries for model implementation. Apache Spark was employed for scalable data handling and distributed training. Experiments were executed on a workstation equipped with an Intel i7 processor, 32 GB RAM, and NVIDIA GPU acceleration.

Preprocessing steps included feature normalization to ensure consistent scaling of environmental variables, while categorical crop labels were encoded numerically to support supervised learning. Since the dataset contained no missing values, imputation was not required.

C. Evaluation Metrics

To evaluate model performance, the confusion matrix was used as the basis for deriving key metrics. Accuracy measures the overall proportion of correct predictions, while Precision quantifies the share of true positives among predicted positives, reducing the risk of misclassifying healthy crops as diseased. Recall (sensitivity) reflects the proportion of actual positives correctly identified, critical in scenarios where missing a disease case could cause yield loss. The F1-score, defined as the harmonic mean of precision and recall, balances these two aspects, particularly in imbalanced data [21]. In addition, the Concordance Index (C-index) was used to measure discriminative ability, indicating how well models rank predictions across instance pairs [4].

D. Hyperparameter Tuning

Hyperparameters, unlike model parameters learned during training, are set prior to the learning process and strongly influence model generalization. Examples include the learning rate in neural networks, the number of trees in Random Forests, and the regularization strength in Logistic Regression. In this study, hyperparameters were optimized using a grid search strategy with cross-validation, following best practices for balancing bias and variance [1], [10]. This approach ensured that each model converged stably while minimizing overfitting, thereby improving predictive reliability.

V. Results and Discussion

This section presents the performance evaluation of the three machine learning models—Random Forest, Neural Networks, and Logistic Regression—applied to crop recommendation and plant disease prediction. Models were assessed using Accuracy, Precision, Recall, F1-Score, and the Concordance Index (C-Index). The results are presented through figures and tables, followed by detailed comparative analysis and discussion.

Figure 1 presents the accuracy scores achieved by the three models.

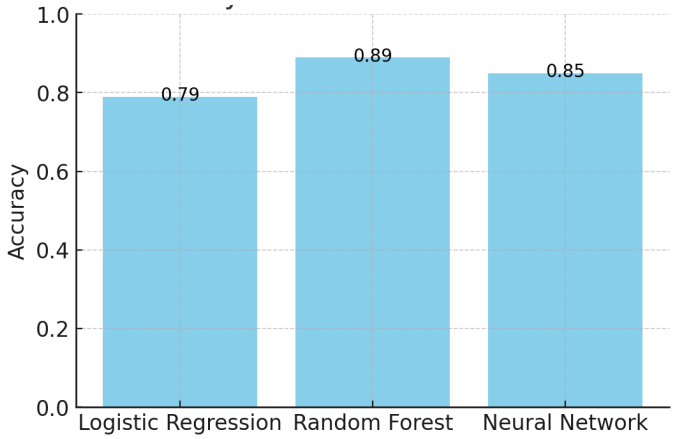

Fig. 1. Accuracy Scores of Machine Learning Models

As shown in Figure 1, Random Forest achieved the highest accuracy (89.2%), followed by Neural Networks (85.0%) and Logistic Regression (79.3%). The superior performance of Random Forest highlights the strength of ensemble methods in capturing complex environmental interactions. Neural Networks performed competitively but were more sensitive to hyperparameter settings, while Logistic Regression underperformed due to its linear assumptions.

Figure 2 illustrates the F1-scores of the three models, which provide a balanced view of Precision and Recall.

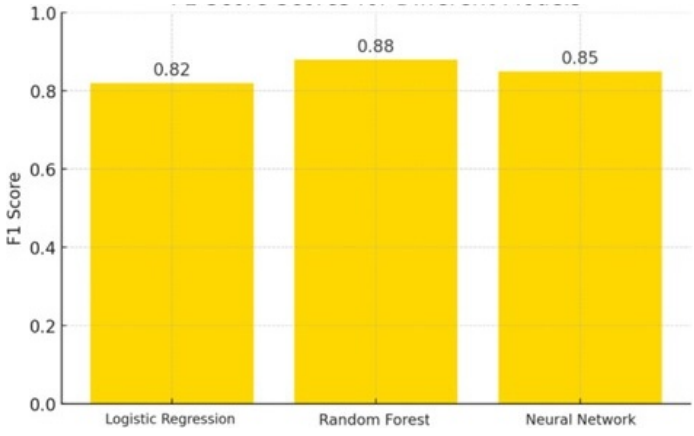

Fig. 2. F1-Score Comparison Across Models

Figure 2 shows that Random Forest achieved the highest F1-score (88.0%), followed by Neural Networks (85.0%) and Logistic Regression (82.0%). These results indicate that Random Forest not only improves overall accuracy but also provides balanced predictions, minimizing both false positives and false negatives. This robustness is particularly valuable in agriculture, where both types of misclassification can have economic consequences.

Figure 3 displays the confusion matrix of the Random Forest model.

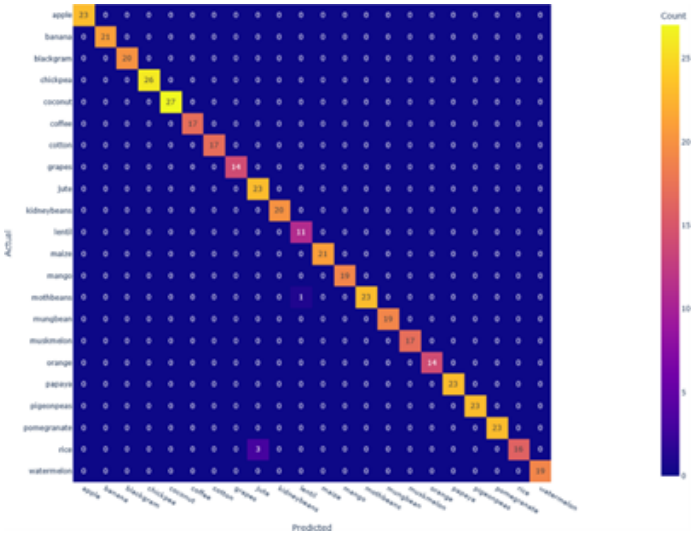

Fig. 3. Confusion Matrix of Random Forest Model

As illustrated in Figure 3, the Random Forest model shows strong diagonal dominance, with most predictions aligning with actual crop classes. Misclassifications were rare and mostly occurred between crops with similar agronomic requirements (e.g., rice and maize). This demonstrates the model's ability to generalize across diverse crop types while also indicating areas where domain-specific knowledge or hybrid approaches could further improve performance.

Table I summarizes the evaluation metrics across all three models.

TABLE I
Performance Comparison of Machine Learning Models

Metric	Random Forest	Neural Network	Logistic Regression
Accuracy	**89.20%**	85.00%	79.32%
Precision	**88.87%**	84.10%	77.56%
Recall	**84.02%**	83.10%	79.43%
F1-Score	**88.00%**	85.00%	82.00%
C-Index	**0.87**	0.84	0.792

Table I confirms that Random Forest consistently outperformed the other models across all metrics. Neural Networks achieved competitive but slightly lower results, while Logistic Regression produced the weakest performance, though still offering acceptable predictive capacity. Random Forest's ability to provide feature importance adds further value by identifying

the most influential variables—such as soil pH, nitrogen, and rainfall—for agricultural decision-making.

Numerically, Random Forest achieved the highest Accuracy (89.20%), Precision (88.87%), Recall (84.02%), F1-Score (88.00%), and C-Index (0.87). Neural Networks followed with Accuracy of 85.00%, Precision of 84.10%, Recall of 83.10%, F1-Score of 85.00%, and C-Index of 0.84. Logistic Regression lagged behind, reaching 79.32% Accuracy, 77.56% Precision, 79.43% Recall, 82.00% F1-Score, and a C-Index of 0.792. These values clearly demonstrate the performance gap between ensemble methods and simpler models, while also confirming that even the baseline Logistic Regression provides a reasonable starting point for interpretable crop prediction tasks.

A. Discussion

The comparative evaluation reveals three main findings. First, Random Forest is the most effective model, demonstrating both strong predictive performance and robustness, making it highly suitable for multi-class agricultural classification tasks. Second, Neural Networks remain a powerful tool for capturing non-linear dependencies, but their performance depends heavily on careful tuning and sufficient training data. Third, Logistic Regression, although less accurate, remains relevant for resource-constrained contexts due to its simplicity and interpretability.

From an agricultural perspective, these results emphasize the importance of models that balance predictive power with transparency and computational feasibility. Ensemble methods such as Random Forest provide reliable predictions while also delivering interpretable insights through feature importance. This dual capability is particularly valuable for farmers and policymakers, who require not only accurate recommendations but also clear explanations of underlying factors.

Overall, these findings highlight the value of integrating ensemble-based approaches within semantic-aware, context-driven frameworks for precision agriculture. By combining predictive robustness, interpretability, and scalability, such approaches can support sustainable farming practices, improve resource allocation, and guide data-driven policy decisions.

VI. Conclusions and Future Work

This study has demonstrated the effectiveness of supervised machine learning models in supporting agricultural decision-making through the prediction of crop suitability and plant disease risks. By leveraging structured environmental variables—such as temperature, rainfall, humidity, soil pH, and nutrient concentrations—the proposed framework provides actionable insights that can guide sustainable crop management and improve resilience to climate-driven challenges.

Among the evaluated models, Random Forest consistently achieved the best performance, with an accuracy of 89.20% and the highest F1-score. Its robustness and interpretability make it particularly suitable for real-world agricultural deployment. These findings underscore the value of ensemble learning methods for optimizing crop selection, fertilization strategies, and disease prevention, thereby reducing uncertainty and enhancing yield outcomes.

Future work will advance this framework in four directions. First, enhanced hyperparameter optimization will be pursued through automated methods such as Bayesian Optimization and Genetic Algorithms. Second, additional machine learning architectures—including Gradient Boosting Machines, Deep Neural Networks, and Support Vector Machines—will be evaluated across diverse crop categories. Third, heterogeneous datasets from multiple geographic regions will be incorporated to capture variability in soil profiles, climate zones, and farming practices. Finally, field-level validation in collaboration with local farmers will assess the framework's practicality under real-world conditions, enabling iterative refinement based on user feedback.

Looking forward, integrating this predictive framework into mobile applications and intelligent decision support systems can provide real-time, personalized recommendations for farmers. Such tools can also inform policymakers in shaping sustainable food production strategies. Embedding semantic- and context-aware mechanisms into these systems will further improve interpretability and adaptability, ensuring that predictive models not only achieve high accuracy but also deliver trustworthy and actionable insights. Promoting knowledge transfer through education and training initiatives will be essential to maximize the societal impact of smart farming technologies and strengthen resilience in agricultural ecosystems.

Overall, this work contributes to the growing field of precision agriculture by demonstrating how machine learning, IoT, and scalable data-driven frameworks can be combined to improve food security and sustainability. By bridging methodological advances with real-world agricultural challenges, it paves the way toward intelligent, context-aware farming systems that empower farmers, support policymakers, and promote resilient agricultural practices worldwide.

References

[1] Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**(2), 281–305 (2012)

[2] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)

[3] Chen, X., et al.: The role of precision technologies in sustainable farm management. Sustainability in Agriculture **12**(1), 142–156 (2021)

[4] Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics **21**(1), 1–13 (2020)

[5] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)

[6] Guo, D., Wang, H., Romanovsky, V.E., Haywood, A.M., Pepin, N., Salzmann, U., Sun, J., Yan, Q., Zhang, Z., Li, X., Otto-Bliesner, B.L., Feng, R., Lohmann, G., Stepanek, C., Abe-Ouchi, A., Chan, W., Peltier, W.R., Chandan, D., von der Heydt, A.S., Contoux, C., Chandler, M.A., Tan, N., Zhang, Q., Hunter, S.J., Kamae, Y.: Highly restricted near-surface permafrost extent during the mid-pliocene warm period. Proceedings of the National Academy of Sciences **120**(36), e2301954120 (2023). https://doi.org/10.1073/pnas.2301954120

[7] Haverkort, A.: Ancha srinivasan (ed): Handbook of precision agriculture: Principles and applications. Euphytica **156**, 269–270 (2007). https://doi.org/10.1007/s10681-007-9390-7, book review

[8] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. John Wiley & Sons, 3rd edn. (2013)

[9] Huang, H.F.: A novel access control protocol for secure sensor networks. Computer Standards & Interfaces **31**(2), 272–276 (2009). https://doi.org/10.1016/j.csi.2007.12.007
[10] Hutter, F., Kotthoff, L., Vanschoren, J.: Automated machine learning: Methods, systems, challenges. Springer (2019)
[11] Ingle, A.: Crop recommendation dataset. https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset (2020), accessed: August 2025
[12] Kamilaris, A., Prenafeta-Boldú, F.X.: Deep learning in agriculture: A survey. Computers and Electronics in Agriculture **147**, 70–90 (2018)
[13] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
[14] Lee, H., Kim, Y.: Advancements in iot and machine learning for precision agriculture. International Journal of Smart Farming **6**(4), 778–792 (2022)
[15] Menard, S.: Applied Logistic Regression Analysis. Quantitative Applications in the Social Sciences, SAGE Publications (2002)
[16] Ogunti, E.O., Adebayo, S., Akinwunmi, A.O.: A survey of medium access control protocols in wireless sensor network. International Journal of Computer Applications **116**(22), 1–8 (2015). https://doi.org/10.5120/20400-2542
[17] Pal, M.: Random forest classifier for remote sensing classification. International Journal of Remote Sensing **26**(1), 217–222 (2005)
[18] ur Rehman, A., Shaikh, Z.: Smart agriculture. In: Smart Agriculture, pp. 120–129. Elsevier (2009)
[19] Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P.: An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing **67**, 93–104 (2012)
[20] Ruiz-Garcia, L., Lunadei, L., Barreiro, P., Robla, J.I.: A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends. Sensors **9**(6), 4728–4750 (2009). https://doi.org/10.3390/s90604728
[21] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427–437 (2009)
[22] Su, D., Qiao, Y., Jiang, Y., Valente, J., Zhang, Z., He, D.: Ai, sensors and robotics in plant phenotyping and precision agriculture, volume ii. Frontiers in Plant Science **14**, 1215899 (2023). https://doi.org/10.3389/fpls.2023.1215899
[23] Wang, N., Zhang, N., Wang, M.: Wireless sensors in agriculture and food industry—recent development and future perspective. Computers and Electronics in Agriculture **50**(1), 1–14 (2006). https://doi.org/10.1016/j.compag.2005.09.003
[24] Zhang, Y., Ren, Y., Liu, Y., Wang, G.: Smart agriculture: Technologies and challenges. Computers and Electronics in Agriculture **197**, 106913 (2022). https://doi.org/10.1016/j.compag.2022.106913