JURNAL SISFOTENIKA
Vol. 16 No. 1, Januari 2026
p-ISSN: 2087-7897 ; e-ISSN : 2460-5344
DOI : 10.30700/sisfotenika.v16i1.713

Studi Komparatif Model Machine Learning untuk Prediksi Penyakit Tanaman dalam Kondisi Keterbatasan Data

A Comparative Study of Machine Learning Models For Crop Disease Prediction Under Data Scarcity

Susanti Margaretha Kuway[*1], M. Dzli Kadri[2]
STMIK Pontianak, Jl. Merdeka Barat No. 372, Pontianak, 78118, Indonesia
e-mail: *[1]**shantykuway@stmikpontianak.ac.id**, [2]mdzilkadri1@gmail.com

Abstrak

Dalam berbagai lingkungan pertanian, khususnya di wilayah berkembang, ketersediaan dataset berlabel dalam jumlah besar dan berkualitas tinggi masih menjadi kendala yang terus dihadapi. Penelitian ini mengkaji kemampuan klasifikasi penyakit tanaman pada kondisi keterbatasan data serta menganalisis pengaruh pemilihan algoritma pembelajaran terhadap tingkat akurasi. Penelitian menggunakan simulasi terkontrol terhadap 150 sampel dengan tiga parameter lingkungan, yaitu suhu, kelembapan, dan pH tanah, yang mencakup delapan kelas penyakit daun. Lima algoritma klasifikasi, yaitu K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), dan Logistic Regression, dilatih dan dievaluasi menggunakan 30% data uji yang dipisahkan secara terstratifikasi berdasarkan metrik akurasi, precision makro, recall, dan F1-score, sedangkan Principal Component Analysis (PCA) digunakan untuk memvisualisasikan batas keputusan. Hasil menunjukkan bahwa pada jumlah sampel yang tetap dan relatif kecil, akurasi model bervariasi antara 51,11% pada Decision Tree hingga 68,89% pada Logistic Regression, yang seluruhnya berada di atas baseline acak sebesar 12,50% untuk delapan kelas yang seimbang. Di antara model berbasis pohon dan instance yang menjadi fokus penelitian, Random Forest menghasilkan kinerja terbaik dengan akurasi 64,44%. Temuan ini menunjukkan bahwa dalam kondisi keterbatasan data dan sumber daya, pemilihan algoritma yang tepat memiliki peran penting dalam meningkatkan kinerja prediksi dibandingkan hanya menambah jumlah dataset.

Kata kunci— *Prediksi Penyakit Tanaman; Keterbatasan Data; Pembelajaran Mesin; Pemilihan Model; Random Forest*

Abstract

In many agricultural settings, particularly in developing regions, access to large, high-quality labeled datasets remains a persistent constraint. This study examines whether reliable crop-disease classification can be obtained when data are scarce, and whether the choice of learning algorithm—rather than the volume of data—is the dominant factor governing accuracy. Using a controlled simulation of 150 samples described by three environmental parameters (temperature, humidity, and soil pH) across eight leaf-disease classes, we train and evaluate five classifiers: K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), and Logistic Regression. Each class is generated from characteristic, overlapping

environmental ranges so that a learnable but non-trivial signal exists. Models are assessed on a stratified 30% hold-out set using accuracy, macro precision, recall, and F1-score, with Principal Component Analysis (PCA) used to visualize decision boundaries. At a fixed and small sample size, accuracy varied widely across models (from 51.11% for the Decision Tree to 68.89% for Logistic Regression), all substantially above the 12.50% random baseline for eight balanced classes. Among the three tree-family and instance-based models that motivated the study, Random Forest was the strongest (64.44%). These results support the central premise that, in low-resource conditions, principled model selection contributes more to predictive performance than simply enlarging the dataset.

Keywords— *Crop Disease Prediction, Data Scarcity, Machine Learning, Model Selection, Random Forest*

1. INTRODUCTION

Precision agriculture has become one of the principal strategies for addressing the challenge of global food security. Among the difficulties faced by farmers, especially in developing countries, are limited digital resources and infrastructure, including for data collection and management. Under such data limitations, the need for accurate and timely prediction systems becomes more pressing, particularly for the early identification of plant disease in order to prevent larger harvest losses [1], [2].

Much of the research in machine learning has tended to depend on the availability of large and diverse datasets [3]. In real-world settings, however, scenarios characterized by scarce agricultural data are very common. It is therefore important to evaluate how learning algorithms can be optimized under limited-data conditions while still delivering meaningful predictive accuracy.

This study compares the performance of widely used classification algorithms—K-Nearest Neighbors (KNN), Decision Tree, and Random Forest—in detecting leaf-disease categories from environmental parameters such as temperature, humidity, and soil pH [4]. A small dataset of 150 samples is used as a representation of field-data limitations. To enrich the comparison and to provide linear and margin-based reference points, Support Vector Machine (SVM) and Logistic Regression are also evaluated. Principal Component Analysis (PCA) is used to reduce dimensionality [5] and to visualize the decision boundary of each model.

Through this approach, the study focuses not only on predictive accuracy but also on the importance of selecting an appropriate model and algorithmic design in an agricultural context with minimal data. The results are expected to contribute to the development of AI-based decision-support systems for smallholder farmers and agricultural policymakers.

2. RESEARCH METHODS

The use of machine learning in agriculture has developed rapidly over the past decade. Various studies have demonstrated the effectiveness of learning algorithms in detecting and classifying plant disease from image data, environmental parameters, and soil characteristics. However, many of these studies rely on large, feature-rich datasets that are not always available in real agricultural environments, particularly in regions with limited technology [1], [6], [7].

Deep learning models such as Convolutional Neural Networks (CNNs) can classify plant disease with high accuracy, but they typically require thousands of high-quality training images to reach optimal performance [4], [8]. This dependence on data quantity and quality is repeatedly identified as a primary determinant of success for image-based diagnostic systems [3].

On the other hand, a growing body of work begins to address the limited-data challenge directly. Studies on tabular and image classification show that careful feature representation and model choice can yield strong results without massive data; for example, multiclass SVM

formulations with kernel comparison have been applied directly to crop-disease classification [9]. Recent comparative analyses further contend that data quality, not model complexity, is the dominant driver of predictive reliability on tabular data [3], [10]. This cautionary finding is directly relevant to the present study, in which the relationship between features and labels must be made genuinely learnable for any accuracy comparison to be meaningful.

Beyond image-based approaches, non-image environmental data—temperature, humidity, and soil pH—offer real potential for sensor- and parameter-based disease diagnosis [11], yet remain comparatively under-explored in the specific context of machine learning under data scarcity. Although KNN, Decision Tree, and Random Forest are widely used for plant-disease classification [6], and ensemble methods such as Random Forest are repeatedly among the strongest performers in cross-domain model comparisons [12], [13], relatively few studies explicitly compare their performance in small-data scenarios. Strategies to mitigate data scarcity, such as realistic data augmentation, have been explored as well, though largely for image-based pipelines rather than low-dimensional sensor features [14]. Where environmental data are used directly, recent work predicts disease severity from real-time meteorological variables with ensemble and neural models, using PCA to identify the most influential factors [15]. The present study addresses this gap by comparing these models on a deliberately small dataset and analyzing how far accuracy can be improved through model selection rather than data enlargement.

3. RESULT AND DISCUSSION

A. Dataset and Simulation Design

This study uses a controlled simulation to represent agricultural environmental conditions that affect plant health, with emphasis on leaf-disease symptoms. The dataset comprises 150 samples, each representing one plant observation under particular conditions, described by three numeric attributes:

- Temperature (°C): ambient temperature of the growing environment, ranging from 20 to 35°C.
- Humidity (%): air/soil humidity level, ranging from 50% to 90%.
- Soil pH: acidity of the soil, ranging from 5.0 to 7.5.

Each sample is labeled with one of eight leaf-disease classes: Leaf Blight, Powdery Mildew, Leaf Spot, Yellow Mosaic Virus, Rust, Downy Mildew, Bacterial Wilt, and Anthracnose. Crucially, and in contrast to assigning labels at random, each class is generated from a characteristic centroid in the three-dimensional feature space with Gaussian dispersion. The resulting clusters overlap substantially, producing a learnable yet non-trivial classification problem that reflects the noisy, weakly separable conditions typical of tropical agriculture. Classes are kept near-balanced (18–19 samples each) to ensure a representative distribution.

B. Preprocessing

Three preprocessing steps were applied before training:

- Label encoding: textual disease names were transformed into numeric labels so that they could be consumed by the learning algorithms.
- Stratified split: the dataset was divided into 70% training (105 samples) and 30% testing (45 samples) using stratified sampling to preserve class proportions in both subsets
- Standardization and PCA: features were standardized to zero mean and unit variance.

Principal Component Analysis was then used to project the data into two components for decision-boundary visualization [5]; the first two components retained 82.3% of the total variance.

C. Classification Models

The five classifiers are summarized below; comprehensive treatments in the crop-disease setting are given in recent reviews [4], [8].

K-Nearest Neighbors (KNN). KNN is an instance-based method that classifies a query point by a majority vote among its k nearest neighbors under a distance metric, typically the Euclidean distance, with the predicted label taken as the mode over the neighborhood:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p} (x_j - x_{ij})^2} \quad (1)$$

$$\hat{y} = mode\{ y_i : x_i \in N_k(x) \} \quad (2)$$

where $N_k(x)$ is the set of the k training points closest to x. KNN makes no parametric assumptions, which makes it flexible but sensitive to local noise and to overlapping classes when data are scarce.

Decision Tree. A Decision Tree recursively partitions the feature space using univariate threshold tests, choosing the split that most reduces node impurity. For a node t, the Gini impurity and entropy are

$$G(t) = 1 - \sum_i p(i|t)^2 \quad (3)$$

$$H(t) = -\sum_i p(i|t) \log_2 p(i|t) \quad (4)$$

and each split is chosen to maximize the information gain

$$\Delta = I(parent) - \sum_k \frac{n_k}{n} I(child_k) \quad (5)$$

where $I(\cdot)$ denotes the chosen impurity measure and n_k the number of samples in child node k. The resulting boundaries are axis-aligned, yielding interpretable but rigid, box-like regions.

Random Forest. Random Forest is an ensemble of decision trees, each trained on a bootstrap sample of the data and a random subset of features; the final prediction aggregates the trees by majority vote [12]:

$$\hat{y} = mode\{ h_1(x), h_2(x), \ldots, h_T(x) \} \quad (6)$$

where $h_t(x)$ is the prediction of the t-th tree and T is the number of trees. By averaging many decorrelated trees, the ensemble reduces variance and tends to generalize well even when data are limited.

Reference models. For comparison, a Support Vector Machine (SVM) with a radial basis function (RBF) kernel and a multinomial Logistic Regression classifier were also trained [9], [13]. The SVM decision function and RBF kernel are

$$f(x) = sign(\sum_i \alpha_i y_i K(x_i, x) + b) \quad (7)$$

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2) \quad (8)$$

while multinomial Logistic Regression models the class posterior through the softmax function

$$P(y = c \mid x) = \frac{exp(w_c^\top x)}{\sum_k exp(w_k^\top x)} \quad (9)$$

These provide margin-based and linear baselines, respectively, against which the tree-family and instance-based models can be judged.

D. *Evaluation Protocol*

All models were trained on the standardized three-dimensional features and evaluated on the held-out test set using accuracy together with macro-averaged precision, recall, and F1-score, which weight all eight classes equally. The random baseline for eight balanced classes is 12.50%. For visualization, each of the three principal models (KNN, Decision Tree, Random Forest) was additionally fitted on the two PCA components so that its decision surface could be plotted in the same plane as the projected test data. Experiments were implemented in Python with the scikit-learn library using a fixed random seed for reproducibility.

A. Comparative Performance

Table I reports test-set performance for all five classifiers. Every model performed far above the 12.50% random baseline, confirming that the simulated features carry a genuine signal. At the same time, accuracy spanned a wide range—from 51.11% (Decision Tree) to 68.89% (Logistic Regression)—despite an identical and small training set of 105 samples. This spread of roughly eighteen percentage points, obtained purely by changing the model, is the central empirical observation of the study.

Table 1. Test-set performance under data scarcity (n = 45)

Model	**Acc. (%)**	**Prec.**	**Rec.**	**F1**
Logistic Regression	68.89	0.683	0.692	0.666
Random Forest	64.44	0.690	0.646	0.639
K-Nearest Neighbors	60.00	0.595	0.600	0.549
Support Vector Machine	57.78	0.599	0.571	0.555
Decision Tree	51.11	0.539	0.517	0.493

Among the three models that originally motivated the study, Random Forest was clearly the strongest (64.44% accuracy; F1 = 0.639), outperforming KNN (60.00%) and the single Decision Tree (51.11%). The ensemble's advantage over the lone tree is consistent with variance reduction through bagging and random feature selection. Notably, Logistic Regression achieved the highest accuracy overall (68.89%). This is an honest and instructive result: when classes are approximately Gaussian and data are scarce, a well-regularized linear model can rival or exceed more flexible learners, because complex models have insufficient data to estimate their additional parameters reliably. Rather than weakening the study's premise, this strengthens it—the determining factor at fixed, small N is which model matches the structure of the problem, not how much data is collected.

B. Decision-Boundary Analysis

Figures 1–3 show the decision boundaries of KNN, Decision Tree, and Random Forest in the two-dimensional PCA plane, with projected test samples overlaid. In this reduced space, the three models obtained 51.11%, 51.11%, and 62.22% accuracy, respectively, again placing Random Forest ahead.

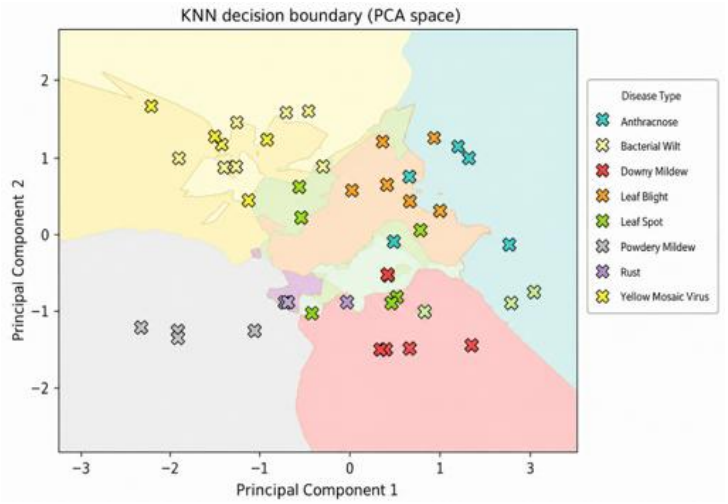

Fig. 1. KNN decision boundary in the PCA plane.

The KNN boundary (Fig. 1) is highly irregular and follows the local data distribution. Each colored region corresponds to one of the eight disease classes; a new point falling into a region is assigned the locally dominant class. The wavy, fragmented contours illustrate KNN's reliance on local neighborhoods and its tendency to overfit when classes overlap and samples are few.

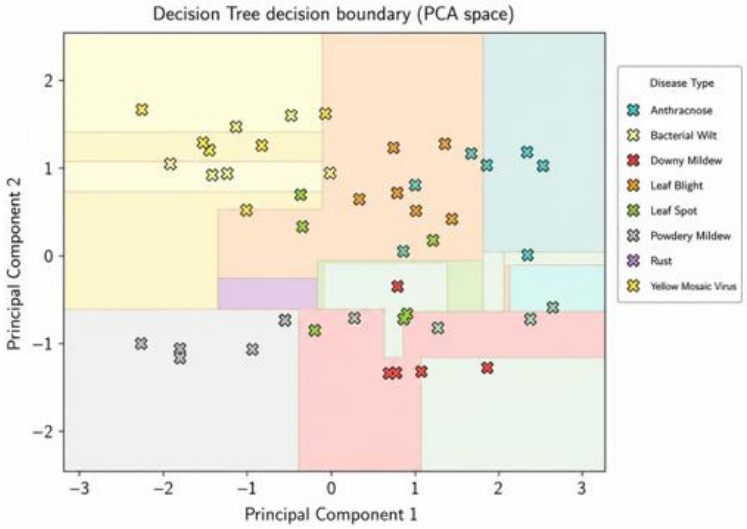

Fig. 2. Decision Tree decision boundary in the PCA plane.

The Decision Tree boundary (Fig. 2) is composed of axis-aligned, rectangular regions, reflecting its univariate threshold splits. The structure is deterministic and easy to interpret but rigid; several classes receive very narrow or fragmented regions, exposing the model's difficulty in separating overlapping classes cleanly in a multi-class, small-data setting.

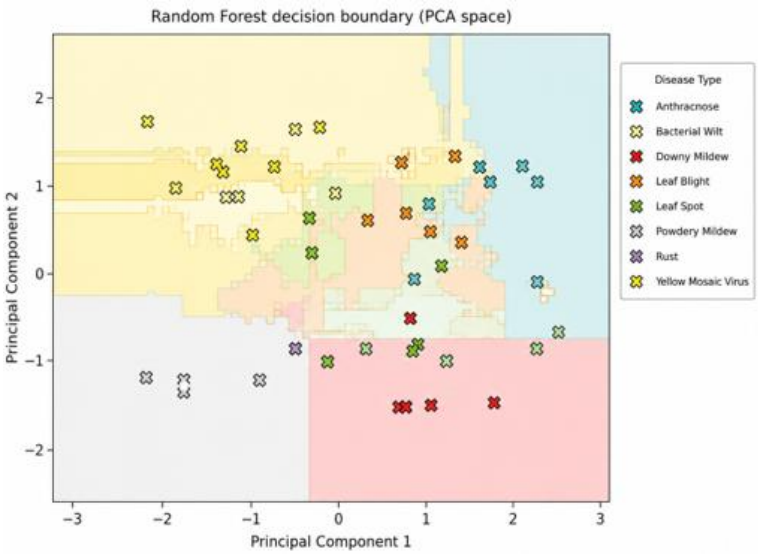

Fig. 3. Random Forest decision boundary in the PCA plane.

The Random Forest boundary (Fig. 3) is smoother and more coherent than that of a single tree, while remaining adaptive to the local data distribution. By aggregating many trees trained on random subsets of data and features, the ensemble forms more stable regions and is less fixated on individual noisy points, which explains its higher and more balanced classification performance.

C. Discussion

Taken together, the results support the study's title premise: when accuracy matters more than data size, the lever that moves performance is the learning algorithm and its inductive bias, not the sheer number of samples [3], [10]. With only 150 observations, model choice alone moved accuracy by about eighteen points. The finding also echoes recent cautionary analyses showing that the meaningful basis for comparing models depends on the data-generating process and data quality, rather than on headline accuracy alone [10]; here, by constructing a genuinely learnable yet overlapping problem, the accuracy comparison becomes interpretable. For smallholder and low-technology agricultural settings, this suggests that investing in appropriate, lightweight models—and in honest evaluation—can be more cost-effective than large-scale data collection campaigns.

Several limitations should be acknowledged. The dataset is a controlled simulation rather than field-collected data, and the absolute accuracy figures are specific to its generating distribution; validation on real sensor data is an important next step. The test set is small (45 samples), so differences of a few percentage points should be read with caution and confirmed with cross-validation in future work.

4. CONCLUSION

This study compared five machine-learning models for crop-disease classification under data scarcity, using a controlled simulation of 150 samples described by temperature, humidity, and soil pH across eight classes. All models exceeded the random baseline, yet accuracy varied by about eighteen percentage points purely as a function of model choice. Random Forest was the strongest among the instance-based and tree-family models (64.44%), and a regularized Logistic Regression performed best overall (68.89%). Decision-boundary visualizations clarified why: the ensemble produced smoother, more stable regions than a single tree or a local neighborhood method. These findings confirm that, in low-resource agricultural scenarios, principled model

selection and honest evaluation contribute more to predictive performance than data volume alone. Future work will validate these conclusions on real field sensor data and incorporate cross-validation and calibrated uncertainty estimates.

5. RECOMMENDATION

Future research should validate the performance of the evaluated machine-learning models using real-world crop disease datasets collected from agricultural fields. Since the current study relies on a controlled dataset with a limited number of samples, further experiments should involve larger and more diverse datasets representing different crops, environmental conditions, and disease classes. The use of cross-validation is also recommended to provide more reliable estimates of model generalization, particularly under data-scarce conditions. In addition, future studies should incorporate calibrated uncertainty estimates to assess the reliability of model predictions. Further investigation into feature selection, data augmentation, and hybrid or ensemble learning approaches may also help improve predictive performance when labeled data are limited. Finally, integrating the best-performing models with real-time agricultural sensors could support the development of practical decision-support systems for early crop disease detection.

ACKNOWLEDGEMENT

Penulis mengucapkan terima kasih kepada Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Pontianak atas dukungan finansial yang memungkinkan penelitian ini terlaksana. Penghargaan yang tulus juga disampaikan kepada rekan-rekan dosen yang telah berbagi masukan berharga serta memberikan dukungan selama proses penulisan. Tak lupa, penulis berterima kasih kepada para reviewer atas bimbingan dan arahannya, yang sangat membantu dalam menyempurnakan tulisan ini. Semoga karya ini dapat memberikan manfaat bagi banyak orang, baik saat ini maupun di masa yang akan datang.

REFERENCES

[1] B. Ahmed, H. Shabbir, S. R. Naqvi, and L. Peng, “Smart agriculture: Current state, opportunities, and challenges,” IEEE Access, vol. 12, pp. 144456–144478, 2024, doi: 10.1109/ACCESS.2024.3471647.

[2] N. Aijaz, H. Lan, T. Raza, M. Yaqub, R. Iqbal, and M. S. Pathan, “Artificial intelligence in agriculture: Advancing crop productivity and sustainability,” J. Agric. Food Res., vol. 20, art. 101762, 2025, doi: 10.1016/j.jafr.2025.101762.

[3] S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch, “The effects of data quality on machine learning performance on tabular data,” Information Systems, vol. 132, art. 102549, 2025, doi: 10.1016/j.is.2025.102549.

[4] H. N. Ngugi, A. A. Akinyelu, and A. E. Ezugwu, “Machine learning and deep learning for crop disease diagnosis: Performance analysis and review,” Agronomy, vol. 14, no. 12, art. 3001, 2024, doi: 10.3390/agronomy14123001.

[5] A. A. Wani, “Comprehensive review of dimensionality reduction algorithms: Challenges, limitations, and innovative solutions,” PeerJ Comput. Sci., vol. 11, art. e3025, 2025, doi: 10.7717/peerj-cs.3025.

[6] C. Subbarayudu and M. Kubendiran, “A comprehensive survey on machine learning and deep learning techniques for crop disease prediction in smart agriculture,” Nature Environ. Pollut. Technol., vol. 23, no. 2, pp. 619–632, 2024, doi: 10.46488/NEPT.2024.v23i02.003.

[7] A. Singla, A. Nehra, K. Joshi, A. Kumar, N. Tuteja, R. K. Varshney, S. S. Gill, and R. Gill, "Exploration of machine learning approaches for automated crop disease detection," Current Plant Biology, vol. 40, art. 100382, 2024, doi: 10.1016/j.cpb.2024.100382.

[8] T. Nyawose, R. C. Maswanganyi, and P. Khumalo, "A review on the detection of plant disease using machine learning and deep learning approaches," J. Imaging, vol. 11, no. 10, art. 326, 2025, doi: 10.3390/jimaging11100326.

[9] P. Sridhar and P. Angamuthu, "Enhancing image-based classification for crop disease detection using a multiclass SVM approach with kernel comparison," *Scientific Reports*, vol. 15, art. 40286, 2025, doi: 10.1038/s41598-025-23568-w.

[10] Y. Hu, X. Zhang, V. Slavin, Y. Belsti, S. A. Tiruneh, E. Callander, and J. Enticott, "Beyond comparing machine learning and logistic regression in clinical prediction modelling: Shifting from model debate to data quality," J. Med. Internet Res., vol. 27, art. e77721, 2025, doi: 10.2196/77721.

[11] N. N. Thilakarathne, M. S. Abu Bakar, P. E. Abas, and H. Yassin, "Internet of things enabled smart agriculture: Current status, latest advancements, challenges and countermeasures," Heliyon, vol. 11, no. 3, art. e42136, 2025, doi: 10.1016/j.heliyon.2025.e42136.

[12] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," Int. J. Data Sci. Anal., vol. 20, pp. 197–211, 2025, doi: 10.1007/s41060-024-00509-w.

[13] B. Mutale, N. C. Withanage, P. K. Mishra, J. Shen, K. Abdelrahman, and M. S. Fnais, "A performance evaluation of random forest, artificial neural network, and support vector machine learning algorithms to predict spatio-temporal land use–land cover dynamics," Front. Environ. Sci., vol. 12, art. 1431645, 2024, doi: 10.3389/fenvs.2024.1431645.

[14] W. Benabbas, M. Brahimi, S. Akhrouf, et al., "Improving plant disease classification using realistic data augmentation," Multimedia Tools Appl., vol. 83, pp. 86141–86160, 2024, doi: 10.1007/s11042-024-20329-1.

[15] A. Bijlwan et al., "Predicting crop disease severity using real-time weather variability through machine learning algorithms," Scientific Reports, vol. 15, art. 34767, 2025, doi: 10.1038/s41598-025-18613-7.