IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

# MACHINE LEARNING AND DEEP LEARNING FOR PLANT DISEASE CLASSIFICATION AND DETECTION

**Devansh Goyal, Rahul Kumare, Kunal Bagde, Sangita Jaybhaye**
Mentee, Mentee, Mentee, Mentor
Department of Computer Engineering,
VISHWAKARMA INSTITUTE OF TECHNOLOGY, PUNE, INDIA

***Abstract:*** Precision agriculture is a rapidly developing field aimed at addressing current concerns about agricultural sustainability. Machine learning is the cutting-edge technology underpinning precision agriculture, enabling the development of advanced disease detection and classification methods. This paper presents a review of the application of machine learning and deep learning techniques in precision agriculture, specifically for detecting and classifying plant diseases. We propose a novel classification scheme that categorizes all relevant works in the associated classes. We separate the studies into two main categories depending on their methodology (i.e., classification or object detection). In addition,we present the available datasets for plant disease detection andclassification. Finally, we perform an extensive computational study on five state-of-the-art object detection algorithms on the PlantDoc dataset to detect diseases present on the leaves, and eighteen state-of-the-art classification algorithms on the PlantDoc dataset to predict whether or not there is a disease in a leaf. Computational results show that object detection accuracy is high with YOLOv5. For the image classification task, the networks ResNet50 and MobileNetv2 have the most optimal trade-off on accuracy and training time.

## I. INTRODUCTION

Precision agriculture has become a key strategy in contemporary agriculture for tackling urgent issues related to productivity and sustainability. Technological advances are the driving force behind this methodological move towards precision, with machine learning emerging as a key technology that enables precise disease detection and classification procedures essential for preserving crop health and maximizing yields.

A thorough analysis explores the complex ways that deep learning and machine learning are used in precision agriculture, emphasizing how they are used to identify and categorize plant diseases. Within this framework, the research presents a novel classification strategy that distinguishes between object detection and classification approaches by methodically classifying pertinent studies according to the methodologies they have employed. It also provides a comprehensive analysis of the datasets that are accessible to support plant disease detection and classification research, highlighting their importance in advancing this subject. The research then conducts a thorough computational investigation, utilizing the PlantDoc dataset as the main benchmark to assess the performance of five top object identification methods and eighteen top classification techniques. The results of this study demonstrate the exceptional accuracy of the YOLOv5 algorithm in object detection tasks and highlight the superior performance of ResNet50 and MobileNetv2 in picture classification, which provide an ideal trade-off between accuracy and computational economy.

Using these thorough examinations, the writers provide insightful information that advances the development of precision farming by providing a sophisticated comprehension of the potential of machine learning methods for illness identification and categorization. This research highlights the critical role that technology plays in improving sustainability and resilience in the agricultural industry, in addition to guiding future agricultural practices.

**NEED OF THE STUDY.**

By utilizing machine learning (ML) and deep learning (DL) approaches to improve plant disease detection and classification, this study is crucial for furthering precision agriculture. The project intends to enhance early disease identification, minimize crop losses, and optimize resource allocation, ultimately promoting sustainable agricultural practices, by methodically classifying existing research and assessing cutting-edge algorithms.

## RESEARCH METHODOLOGY

To meet the urgent demand for automated disease diagnosis in the farming sector, a system for paddy plant disease detection and classification has been presented. It incorporates techniques from computer vision, image processing, machine learning, and deep learning. The system's primary method is hierarchical and includes image pre-processing, segmentation, feature extraction, classification, and prescription for a predictive treatment.

The first step in the process is picture pre-processing, which aims to improve the quality of input photographs by optimizing contrast, minimizing noise, and maintaining uniform lighting. Accurate illness detection is facilitated and ideal circumstances for further investigation are ensured in this step.
After pre-processing, the system separates unhealthy areas from healthy areas in the paddy plant photos using image segmentation algorithms. Segmentation makes it possible to precisely localize diseases, which facilitates focused investigation and categorization. It does this by drawing boundaries around the affected areas.

The next step involves extracting features from the segmented regions to obtain distinguishing qualities. Texture, colour, and form descriptors are among the traits that are extracted to successfully characterize various types of paddy plant diseases. This stage is essential to improving the system's capacity to distinguish between different illness types.

The retrieved features are then used with machine learning and deep learning algorithms for the classification of diseases. To accurately recognize and classify diseases, the system combines convolutional neural networks with support vector machine classifiers. The classifiers are trained extensively on labelled datasets and can differentiate between several paddy plant diseases based on visual cues.

Thorough testing is done with a variety of datasets that include photos of paddy plants with bacterial leaf blight, fake smut, brown leaf spot, rice blast, and sheath rot to assess the effectiveness of the suggested approach. The system's classification accuracy and resilience may be evaluated thanks to the dataset's partitioning into training and testing subsets.

Based on experimental data, the deep learning-based strategy is effective, as the suggested system achieves a high validation accuracy of 0.9145. Moreover, the predictive capacity of the system enables the recommendation of suitable treatments to counteract detected diseases, enabling agriculture-related stakeholders to take proactive steps to protect the health of paddy crops and maximize yield results.

All things considered, the suggested methodology offers a thorough and effective method for automatically identifying and categorizing paddy plant diseases, with great promise for raising agricultural output and sustainability.

### 3.1 Classification Scheme

In this section, we propose a classification scheme and categorize all relevant works in the associated classes. In addition, we present statistics of the most common crops/data types/methods/algorithms/datasets/metrics used by the reviewed papers. The classes that we use in our classification scheme are the following:

- *Crop*: the type of crop that each study uses as a case study
- *Input data*: the type of the input data used in the ML algorithms
- *Dataset*: the name of the dataset that is used if information is available in the paper
- *Models/Algorithms*: the ML algorithm that was used
- *Method*: whether classification (C) or object detection is targeted (C)
- *Metrics*: the evaluation metrics employed to measure the performance of the ML algorithms
- *Results*: a brief description of the results obtained by the ML algorithms

CNN is the most common neural network architecture used throughout the studies. There are also some works dealing with classifying diseases that use SVM, ANN, RF, KNN, and MLP algorithms.

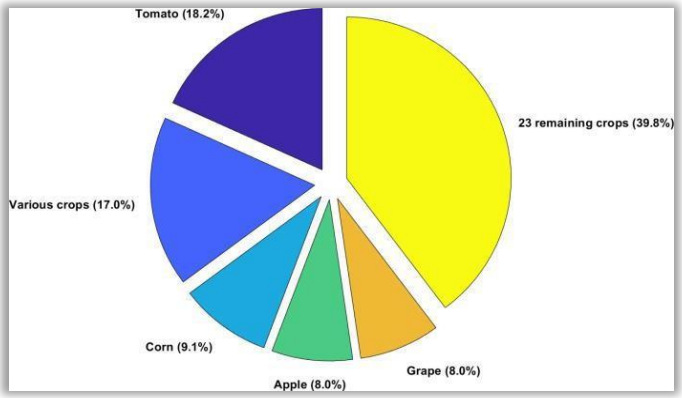


FIGURE 1: Statistics of crops.

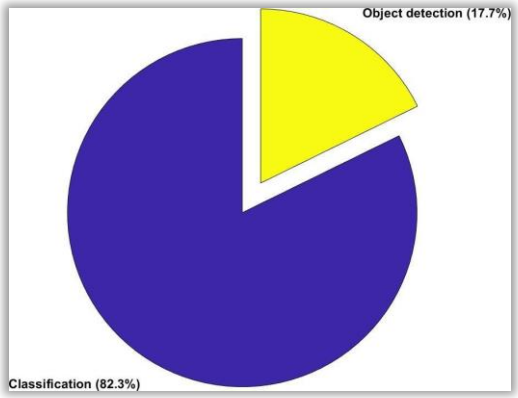


FIGURE 2: Statistics of input data types.

Regarding the method that each study targets, classification is the clear winner compared to object detection. Different metrics have been employed in the literature, with accuracy, F1-score, precision, and recall being among the most commonly used ones.

### 3.2 Data and Sources of Data

This section provides an overview of publicly available datasets. These datasets serve different purposes; some of them are used for classification to determine if a plant image is healthy or infected with a disease (discussed in Subsection 3.2-A), while others are used for object detection to identify diseases on plants (discussed in Subsection 3.2-B).

#### a) Classification

The PlantVillage dataset [10] comprises 54,303 leaf images, both healthy and diseased, categorized into 38 classes based on species and diseases. The dataset includes images of 14 crop species, such as apple, blueberry, cherry, corn, grape, orange, peach, bell pepper, potato, raspberry, soybean, squash, strawberry, and tomato. It covers 17 fungal-related diseases, four bacterial diseases, two mold (oomycete) diseases, two viral diseases, and one mite-related disease. FIGURE 6: Statistics of methods Additionally, it provides images of disease-free healthy leaves from 12 crop species.

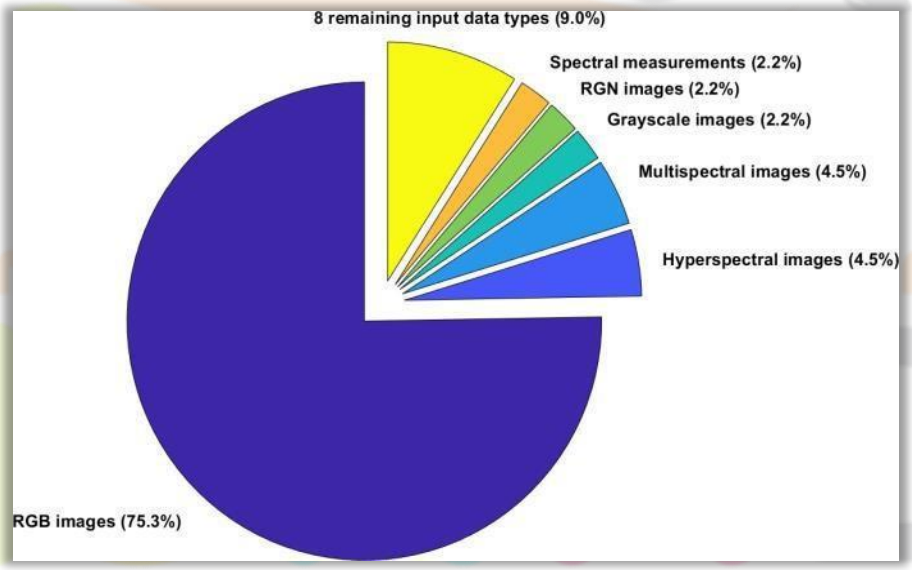


FIGURE 3: Statistics of datasets

The bean leaf image dataset comprises images of bean leaves captured from field conditions. The National Crops Resources Research Institute (NaCRRI), the national organization in charge of research in agriculture in Uganda, and the Makerere AI lab collaborated to capture these images in several areas of Uganda. Images were captured from the field or garden using a simple smartphone, which were then analyzed by NaCRRI experts who determined which illness was present in each image. The dataset consists of 1,296 images and three classes. 428 images are for the healthy class, 432 images are for the angular leaf spot, and the remaining 436 are for the bean rust.

The PlantLeaves dataset is comprised of 4,503 im- ages of plant leaves, both healthy and diseased, categorized into 22 categories based on the species and the state of health. The dataset includes 2,278 healthy leaf images and 2,225 diseased ones. The images were captured using a basic digital camera.

The PlantaeK dataset is a leaf database of indigenous plants found in Jammu and Kashmir. It comprises 2,153 images of healthy and diseased plant leaves, categorized into 16 groups by species and health status. The images feature various crop species, including apple, apricot, cherry, cranberry, grapes, peach, pear, and walnut. The dataset comprises 1,223 healthy leaf images and 934 diseased leaf images. The Plant Pathology 2020 challenge dataset [129] is a classification dataset for the foliar disease of apples.

The creators of the dataset manually captured 3,651 real-world symptoms of several apple foliar diseases with varied lighting, angles, surfaces, and noise. The dataset includes 865 healthy leaves, 187 cases of complex diseases, 1,200 cases of apple scab, and 1,399 cases of cedar apple rot.

The citrus leaf images dataset contains images of healthy and infected citrus plants with diseases such as black spots, canker, scab, greening, and melanosis. The dataset in- includes 609 images from citrus leaves, of which 58 are healthy, and 150 images from citrus fruits, of which 22 are healthy.

The Kaggle dataset contains 9, 436 annotated images and 12,595 unlabeled images of cassava leaves. The dataset contains five classes, one is the class for healthy plants and the other four are for diseases (cmd, cgm, cbsd, and cbb). NaCRRI in collaboration with the AI lab at Makerere University captured and annotated these images.

The dataset of Rice leaf images [95] includes 120 images collected from a village in India. The dataset contains 40 images of each disease, for a total of 120 images. The NLB dataset is comprised of 234 images of leaf spot disease in maize crops.

### b) OBJECT DETECTION

The PlantDoc dataset includes 2, 345 images. These images contain 13 plant species and 18 classes of diseases. This dataset is publicly available for download and it can also be used as an open dataset for benchmarks. The classes of the PlantDoc dataset are the following: Cherry leaf, Peach leaf, Cherry leaf, Peach leaf, Corn leaf blight, Apple rust leaf, Potato leaf late blight, Strawberry leaf, Corn rust leaf, Tomato leaf late blight, Tomato mold leaf, Potato leaf early blight, Apple leaf, Tomato leaf yellow virus, Blueberry leaf, Tomato leaf mosaic virus, Raspberry leaf, Tomato leaf bacterial spot, Squash Powdery mildew leaf, Grape leaf, Corn Gray leaf spot, Tomato Early blight leaf, Apple Scab Leaf, Tomato Septoria leaf spot, Tomato leaf, Soyabean leaf, Bell pepper leaf spot, Bell pepper leaf, grape leaf black rot, Potato leaf, and Tomato two-spotted spider mites leaf. PlantDoc contains annotations with an average of 3.4 annotations per image. The average image size is 0.53 mp and the distribution of the image sizes starts from 0.01 mp to 24.00 mp. The median image ratio is

800 675. The balance of the classes of PlantDoc is presented in Figure 8. The classes Blueberry leaf, Tomato leaf yellow virus, and Peach leaf are overrepresented with more than 600 images for each class and the classes Tomato leaf late blight, Tomato Early blight leaf, Apple rust leaf, Apple Scab Leaf, grape leaf black rot, Corn rust leaf, Corn Gray leaf spot, Soybean leaf, Potato leaf, and Tomato two-spotted spider mites' leaf are under-represented with less than 220 images for each class. A sample of four images with their annotations are shown in Figure 5.

The CropDeep dataset is composed of 31,147 im- ages containing over 49, 000 annotated instances from 31 different classes. The images were captured in greenhouses under various conditions using different cameras. Additionally, the IP102 dataset is a comprehensive benchmark dataset for recognizing insect pests. It comprises over 10, 000 images divided into 102 categories, with insect pests that mainly target one agricultural product grouped in the same top-level category. The IP102 dataset has a hierarchical taxonomy.

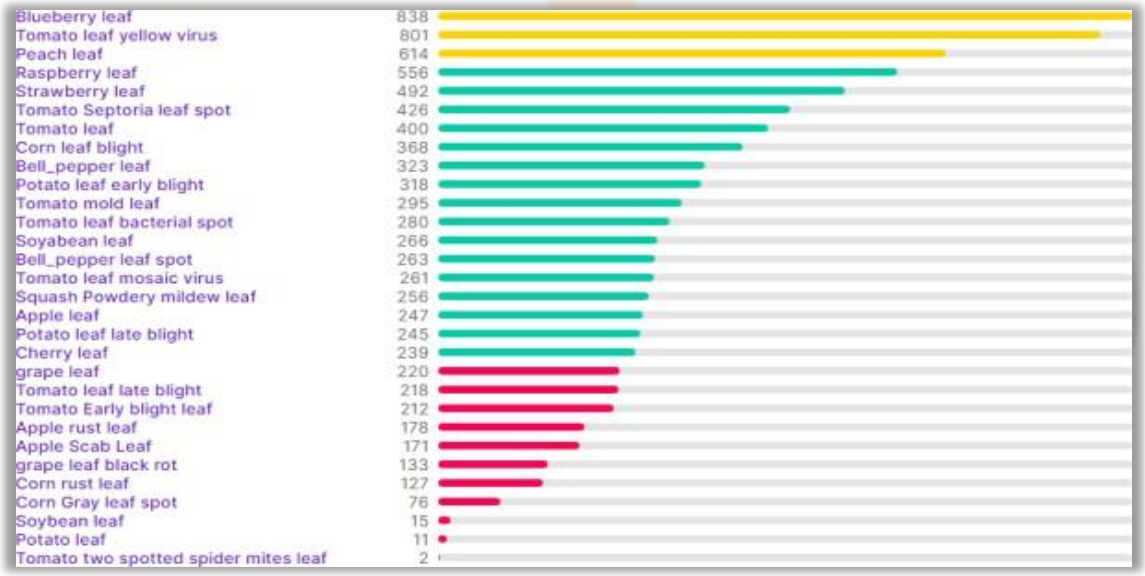


FIGURE 4: Class balance of the PlantDoc dataset


FIGURE 5: Visualization of the PlantDoc dataset with image annotations

The CropDeep dataset is composed of 31,147 images containing over 49, 000 annotated instances from 31 different classes. The images were captured in greenhouses under various conditions using different cameras. Additionally, the IP102 dataset is a comprehensive benchmark dataset for recognizing insect pests. It comprises over 10, 000 images divided into 102 categories, with insect pests that mainly target one agricultural product grouped in the same top-level category. The IP102 dataset has a hierarchical taxonomy Table 1 presents the statistics for each dataset including:
(i) the name of the dataset, (ii) the type of the crop, (iii) the number of images, (iv) the number of classes, and (v) whether classification (C) or object detection is targeted (C).

| Dataset | Crop | Images | Classes | Method |
|---|---|---|---|---|
| PlantVillage | Various crops | 54, 303 | 38 | C |
| Bean | Bean | 1, 296 | 5 | C |
| PlantLeaves | Various crops | 4, 502 | 22 | C |
| Plantae | Various crops | 2, 153 | 16 | C |
| Plant Pathology | Apples | 3, 651 | 4 | C |
| Citrus leaf images | Citrus fruits | 759 | 6 | C |
| Kaggle | Cassava | 22, 031 | 5 | C |
| Rice leaf images | Rice | 120 | 3 | C |
| NLB | Maze | 234 | 1 | C |
| PlantDoc dataset | Various crops | 2, 345 | 18 | O |
| CropDeep | Various crops | 31, 147 | 31 | O |
| IP102 | Corn | 10, 000 | 102 | O |

TABLE 1: Datasets statistics

### 3.3 Theoretical framework

The PlantDoc dataset was used in this computational work to train and assess five cutting-edge object detection algorithms and eighteen classification techniques. Annotation correction, image scaling, auto-orientation, and denoising were among the pre-processing techniques used. A system for high-performance computing was used for the experiments. EfficientDet, Faster RCNN, RetinaNet, SSD, and YOLOv5 were among the object identification algorithms evaluated; each has particular hyperparameters designed for the best results. For example, EfficientDet used a batch size of 8 and a learning rate of 5e-2, whereas Faster RCNN used a batch size of 16 and a learning rate of 2e-4. These setups attempted to strike a balance between stability and convergence speed, which helped in leaf disease categorization and detection.

| Algorithm m | LR | BS | Optimizer | Backbone |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Efficient Det | 5.00E-02 | 8 | Adam | Efficient Net |
| Faster RCNN | 2.00E-04 | 16 | SGD | ResNet-50 |
| RetinaNet | 7.00E-05 | 8 | Adam | ResNet-50 |
| SSD | 7.00E-05 | 32 | Adam | VGG16 |
| YOLOv5 | 3.00E-03 | 32 | Adam | CSPDar knet53 |

TABLE 2: Hyperparameters of the algorithms

The findings show that YOLOv5 outperforms the other algorithms in terms of accuracy and is particularly good at identifying objects in photographs of various sizes. This superiority is visible in multiple assessment criteria, such as the detection of tiny, medium, and large objects and the rigorous AP metric of the COCO challenge. Although EfficientDet, Faster RCNN, SSD, and RetinaNet show comparable accuracy, they are not as good as YOLOv5. The results are visualized, showing that all algorithms successfully identified and located the majority of leaf species and illnesses, even if some photos had less-than-ideal detections. In addition, eighteen cutting-edge CNN model architectures for image classification—AlexNet, DenseNet, MobileNetV2, ResNet, ResNeXt, ShuffleNet, VGG, and WideResNet—were trained and assessed as part of the study.

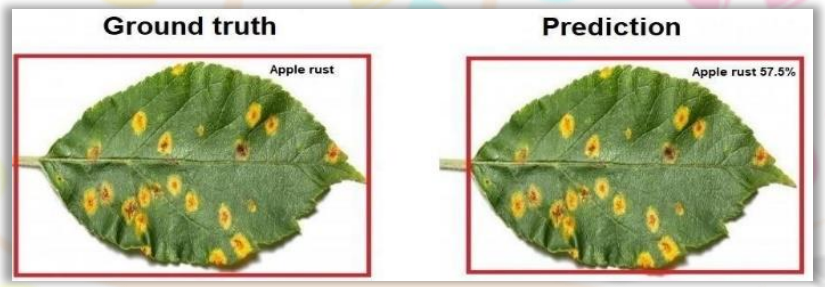


FIGURE 6: YOLOv5 ground truth vs prediction visualization.

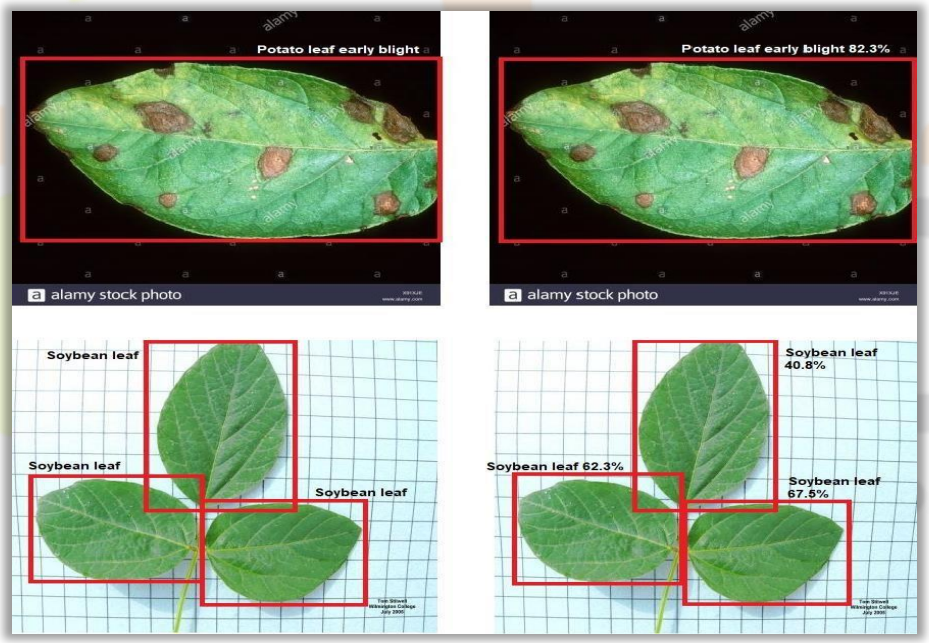


FIGURE 7: YOLOv5 ground truth vs prediction visualization.

## IV. RESULTS AND DISCUSSION

In computer vision, choosing appropriate approaches requires balancing the benefits and drawbacks of different approaches for object recognition and classification. Machine learning is based on classification techniques, which efficiently classify inputs into predetermined classes while providing simplicity and interpretability. They have trouble managing several items at once or jobs requiring exact localization due to their lack of fine-grained spatial knowledge. On the other hand, object detection techniques work well in situations where handling multiple objects and duties such as instance segmentation are required. They do this by providing

class labels in addition to accurate geographical information. However, their adoption in real-time and resource-constrained situations is limited due to their sophisticated implementation, labour-intensive data annotation requirements, and large computing resources.

High-performing disease detection algorithms are essential for early disease identification, reducing crop losses, and allocating resources optimally in precision agriculture. They accurately identify illnesses including viral breakouts and fungus infections, which lessens the need for lengthy therapies. But putting them into practice might call for specialized gear and powerful processing power. Accurate model training requires high-quality labelled datasets that capture a variety of illness symptoms and environmental factors. Farmers are empowered and trust is fostered through integration with current precision agriculture systems through standardized interfaces and user-friendliness.

Two examples of useful applications include John Deere's See & Spray technology, which uses object detection algorithms to maximize resource utilization, and PlantVillage, which uses classification algorithms for real-time disease identification. Using object identification algorithms, Blue River Technology's See & Spray solution improves crop management's sustainability and economics. IBM's Smart Agriculture solution uses AI and machine learning to combine sophisticated algorithms with a data-driven methodology to give farmers practical recommendations.

### a) CHALLENGES IN PLANT DISEASE DETECTION

1) There are only a few completely annotated open datasets. Many studies rely on the PlantVillage dataset, which was obtained in a controlled laboratory setting. Generating larger datasets under real-world conditions is crucial. Collaborative efforts are needed to create representative datasets.

2) Most works treat the disease detection problem as a classification problem, either binary classification or multi-class classification. While many works treat disease detection as a classification problem, more emphasis should be placed on object detection to identify both the disease type and affected regions in the image.

3) Most papers use a single dataset used to train and test the model. Models trained on a single dataset After the detailed review of ML and DL algorithms for plant disease detection and classification and the

detailed computational study on five state-of-the-art object detection algorithms for plant disease detection and eighteen state-of-the-art classification algorithms for plant disease classification on a widely-used dataset, we have identified several challenges in practical applications of plant disease detection:

There is a lack of models that handle non-image data. Most existing classification and object detection algorithms focus solely on image data, neglecting other relevant information such as temperature and humidity. Developing techniques to incorporate non-image data is essential for more accurate predictions. often perform poorly on different datasets. It is essential to consider diverse datasets to improve model robustness.

4) Overreliance on CNN architectures: While CNNs yield good results, exploring other neural network architectures like recurrent neural networks can enhance dis-lease detection methods.

5) Small leaf and early-stage disease recognition: Current datasets mainly consist of images with large leaves. Annotating datasets for early-stage disease detection and small leaf recognition is necessary.

6) Challenges with illumination and occlusion: Existing algorithms struggle with images under varying lighting conditions and occlusion. More robust methods are needed to address these issues.

7) Computational efficiency: Many models are computationally intensive, hindering real-time applications. Researchers should focus on improving the computational efficiency of their models.

### b) FUTURE DIRECTION IN PLANT DISEASE DETECTION

In addition to the challenges mentioned above, there are several promising directions for future research in plant disease detection:

1) Integration of non-image data: Develop models that can effectively integrate non-image data, such as environmental factors, into disease detection algorithms to improve prediction accuracy.

2) Creation of diverse and real-world datasets: Collaborate with experts to generate large, representative datasets under real-world agricultural conditions to enhance the generalizability of models.

3) Emphasis on object detection: Explore the potential of object detection methods for predicting plant diseases, which can provide more detailed information about disease localization.

4) Robustness across datasets: Develop models that perform consistently well across various datasets to ensure their practical utility.

5) Exploration of alternative neural network architectures: Experiment with different neural network architectures beyond CNNs, such as recurrent neural networks, to uncover their potential in disease detection.

6) Early-stage and small leaf recognition: Annotate datasets specifically for early-stage disease recognition and the identification of diseases on plants or leaves with small sizes.

7) Addressing illumination and occlusion challenges: Implement techniques to enhance the robustness of algorithms in the presence of variable lighting conditions and occluded images.

8) Improved computational efficiency: Focus on optimizing model architectures and algorithms to make them suitable for real-time applications.

## V. CONCLUSION

This study concluded with a thorough analysis of the use of deep learning (DL) and machine learning (ML) approaches in precision agriculture, with an emphasis on the identification and categorization of plant diseases. A new categorization scheme that distinguished between object detection and classification approaches was presented to group pertinent studies according to their methods. Furthermore, databases for the identification and categorization of plant diseases were made available, together with information about how well-suited they were for various activities. Utilizing the PlantDoc dataset, computational tests on cutting-edge algorithms showed that YOLOv5 exhibited the highest accuracy in object detection, while ResNet50 and MobileNetv2 showed the best balance between training time and accuracy in classification tasks. To further improve the precision and applicability of plant disease detection and classification systems, future research will examine new algorithms, datasets, and picture preprocessing techniques.

## REFERENCES


[1] Ramesh, S., Hebbar, R., Niveditha, M., Pooja, R., Shashank, N., & Vinod, P. V. (2018, April). Plant disease detection using machine learning. In the *2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)* (pp. 41-45). IEEE.

[2] Saleem, M. H., Potgieter, J., & Arif, K. M. (2019). Plant disease detection and classification by deep learning. *Plants*, *8*(11), 468.

[3] Shruthi, U., Nagaveni, V., & Raghavendra, B. K. (2019, March). A review on machine learning classification techniques for plant disease detection. In *2019 5th International Conference on Advanced Computing & communication systems (ICACCS)* (pp. 281-284). IEEE.

[4] Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and electronics in agriculture*, *145*, 311-318.

[5] Li, L., Zhang, S., & Wang, B. (2021). Plant disease detection and classification by deep learning—a review. *IEEE Access*, *9*, 56683-56698.

[6] Albattah, W., Nawaz, M., Javed, A., Masood, M., & Albahli, S. (2022). A novel deep learning method for detection and classification of plant diseases. *Complex & Intelligent Systems*, 1-18.

[7] Panigrahi, K. P., Das, H., Sahoo, A. K., & Moharana, S. C. (2020). Maize leaf disease detection and classification using machine learning algorithms. In *Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2019* (pp. 659-669). Springer Singapore.

[8] Applalanaidu, M. V., & Kumaravelan, G. (2021, February). A review of machine learning approaches in plant leaf disease detection and classification. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 716-724). IEEE.

[9] Nancy, P., Pallathadka, H., Naved, M., Kaliyaperumal, K., Arumugam, K., & Garchar, V. (2022, March). Deep learning and machine learning-based efficient framework for image-based plant disease classification and detection. In *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)* (pp. 1-6). IEEE.

[10] Haridasan, A., Thomas, J., & Raj, E. D. (2023). Deep learning system for paddy plant disease detection and classification. *Environmental monitoring and assessment*, *195*(1), 120.