MegaPlant: A Dataset and Modular Decision Pipeline for Autonomous In-Field Plant Disease Detection

witty/comical group name

November 12, 2025

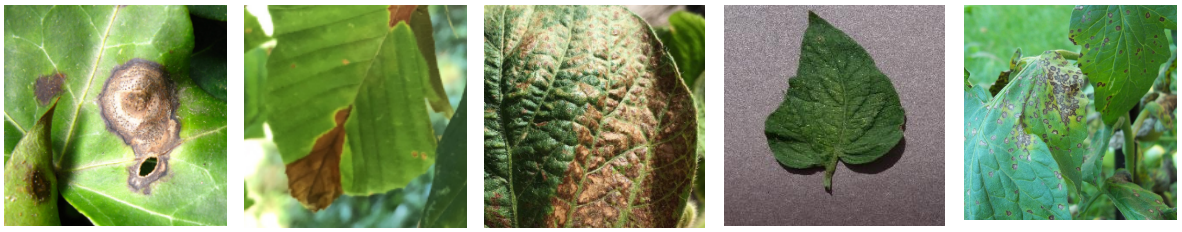


Figure 1. Images of multiple unhealthy leaves in varying conditions. (1, 2, 3) Beckerman & Creswell, 2022; (4) Hughes & Salathé, 2015; (5) Singh et al., 2020.

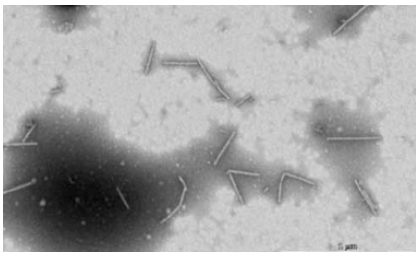
Abstract

We introduce MegaPlant, a consolidated leaf-image dataset designed to support plant disease classification models that generalize across diverse environmental conditions, from controlled laboratory settings to highly variable in-field scenarios. MegaPlant integrates multiple publicly available datasets and standardizes them into a unified taxonomy of healthy and diseased leaf categories, enabling robust training across modalities. In addition, we propose a compartmentalized decision-making framework tailored for fully autonomous, in-field agents such as UAVs and mobile scouting robots. The framework separates disease detection from symptom identification, reducing single-point failure risks and improving reliability in real-world deployments. This modular structure also enhances interpretability, allowing practitioners to diagnose which stage of the pipeline misperformed when errors occur. Together, MegaPlant and our decision framework facilitate more dependable, scalable, and transparent plant disease surveillance systems suited for modern precision agriculture.

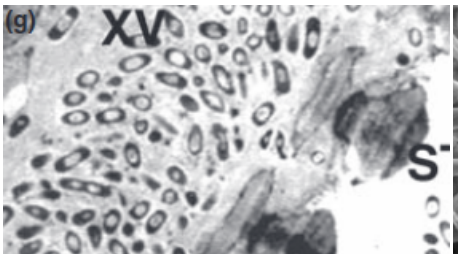
Table of Contents

1 Background

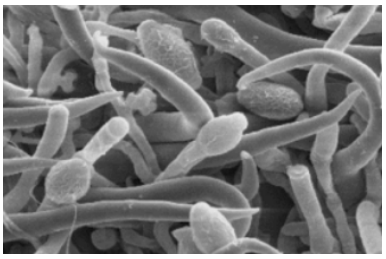
Plant diseases are abnormal changes in appearance and behaviour that progresses over time, unlike plant injury that occurs immediately (DeBusk, 2019). These are caused by pathogens such as viruses, bacteria, fungus, oomycetes (fungus-like micro-organisms), parasitic nematodes (worm-like micro-organisms), and parasitic plants. Pathogens and pests (P&Ps) account for about 20% and at least 10% of harvest yield loss in major crops (Savary et al., 2019, 1; Strange & Scott, 2005, 83).



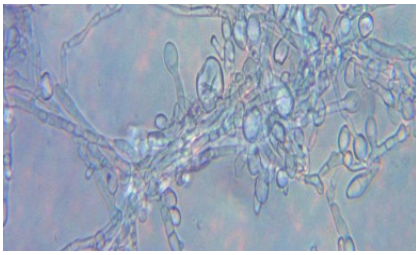
(a) Pepper mild mottle virus



(b) Cabbage afflicted with black rot marked XV

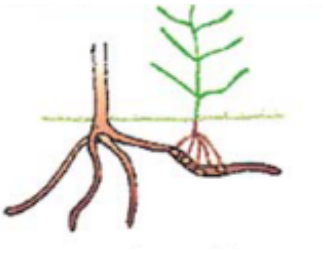


(c) fungus on an infected host flower



(d) oomycete “potato blight” *P. infestans*

(e) parasitic nematode



(f) parasitic plant “witchweed” *Striga*

Figure 2. Morphology[1] of various pathogens. (a) Colson et al., 2010, 4; (b) Dow et al., 2016; (c) Pinto et al., 2016, 258; (d) Raza et al., 2022, 8; (e) Mitiku, 2018, 36; (f) Agrios, 2009, 617.

Although most diseases are caused by pathogens or biotic factors, some are a result of direct injury or abiotic factors, also called environmental factors. These factors are drought, winter, disruptive human activities, etc. Diseases caused by abiotic factors are easier to diagnose but harder to control (Agrios, 2009, 613).

Considering that these pathogens are micro-organisms and invisible to the human eye, the method for identifying if a plant is unhealthy or infected is by identifying the symptoms and signs visually. However, identifying the exact disease-causing agent will

[1] **morphology,** in biology, is the study of the size, shape, and structure of animals, plants, and microorganisms and of the relationships of their constituent parts. - www.britannica.com

require certain procedures often done by professional plant pathologists (Strange & Scott, 2005, 96; UNH Extension, 2015).

1.1 Symptoms

To diagnose a plant immediately is by looking at the symptoms, these symptoms are reactions of the plant to the pathogen, not necessarily a sign of the particular pathogen itself. Signs of a plant disease are physical evidence of the causal agent or pathogen, signs are not symptoms (Penn State Extension, 2017).

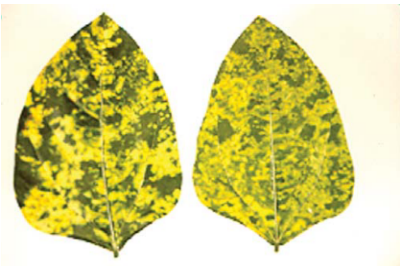
(a) localized lesions caused by virus

(b) an apple exuding bacterial ooze

(c) scorched leaf symptoms caused by bacteria

(d) canker[2] on a leaf stem

(e) tomato ring malformations



(f) mosaic pattern symptom

(g) dwarfism/stunted growth and bronzing caused by virus

(h) malformation and mosaic pattern symptoms

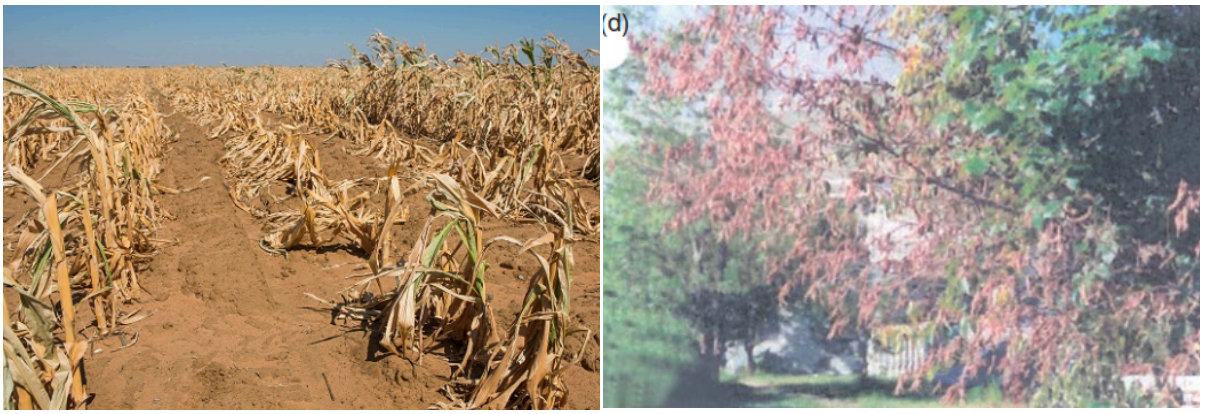
(i) tree bark afflicted with rot

Figure 2. Plants showing 9 of many symptoms observable by the human eye. (a) Colson et al., 2010, 4; (b, c, d, e, f, g, h, i) Agrios, 2009, 627-634.

[2] Sunken necrotic patch of bark (Beckerman & Creswell, n.d., 9)

Symptoms caused by abiotic factors are referred to as disorders. These symptoms are usually uniform, affecting large or evenly distributed areas of vegetation. In contrast, diseases caused by pathogens are often non-uniform and appear as scattered or irregular patches in the field (UNH Extension, 2015). With deep learning models that focus on visual inspection for plant disease detection, it may be more practical to determine if a plant is unhealthy or otherwise, rather than detecting specific diseases or disorders.

For example, a deep learning model may classify a plant as unhealthy due to bronzing during autumn, but may be harder to determine what the disease or disorder is, without additional temporal or environmental context.



(a) Total loss of a corn field due to drought

(b) Dehydrated tree due to fungi inhibiting water passage to the tree branches

Figure 4. Abiotic versus biotic induced symptoms.
(a) The Independent, 2019; (b) Agrios, 2009, 618.

Figure 4 shows the same symptoms in two different plants but caused by different factors. This shows that environmental context will be needed when diagnosing plant diseases. As such, this case study will focus only on identifying whether a plant is healthy or unhealthy based on the symptoms observed by the human eye.

1.2 Signs

(a) fungus growing as conks

(b) flagging, the insect feeding on the branch is highlighted

(c) powdery mildew appearance is composed of a fungus' hyphae

(d) web-like hyphae of a fungus

(e) rusting and spores of a fungus

(f) japanese beetles feeding

Figure 5. Signs of the P&Ps inflicting disease (Beckerman & Creswell, 2022).

Signs are physical evidence of the pathogen or pest, the real cause of the plant disease. Knowing the signs is key information to generating actions or solutions for P&P management.

Observing the signs is a sure enough method to indicate that the plant might be unhealthy. However, that will take a more complex deep learning model, to consider another piece of information, for example, the bugs scattered on the branches, or fungus hyphae fully covering the subject leaf or plant. This study's scope will only incorporate symptoms observed on the leaves and some simpler signs, like fungus mildew.

1.3 Detection Methods

Plant disease detection can be performed using a range of approaches, from traditional manual observation to advanced computational techniques. Each method varies in accuracy, cost, scalability, and practicality depending on the use case.

1.3.1 Manual and Laboratory Methods

Visual inspection is the most common and oldest method, where farmers or plant pathologists examine the visible symptoms on leaves, stems, or fruits such as spots, blight, or discoloration. Although simple and fast, this method is subjective and heavily reliant on human expertise and environmental conditions (Penn State Extension, 2017).

In laboratory diagnostics, several scientific tests are employed to accurately identify pathogens:

- Microscopy – Used to observe fungal spores or bacterial colonies.
- Culture tests – Pathogens are isolated and grown in nutrient media for species identification.
- Serological tests (e.g., ELISA) – Use antibodies to detect specific proteins associated with viruses or bacteria.
- Molecular techniques (PCR, qPCR, LAMP) – Detect pathogen DNA or RNA, offering high sensitivity and specificity (Ward et al., 2004; Schaad et al., 2003).

While these approaches are precise, they require laboratory equipment, trained personnel, and are not suitable for large-scale or real-time monitoring.

1.3.2 Image Processing

Before the advent of deep learning, plant disease detection often relied on handcrafted features derived from image processing.

Key features such as color, texture, and shape were extracted using algorithms like:

- Gray-Level Co-occurrence Matrix (GLCM) for texture analysis,
- Local Binary Patterns (LBP) for surface variation, and
- Color histograms in RGB or HSV space for spotting discoloration.

These features were then classified using traditional machine learning algorithms such as:

- Support Vector Machines (SVMs)
- k-Nearest Neighbors (k-NN)
- Random Forests
- Naïve Bayes classifiers

For example, Pydipati et al. (2006) demonstrated that SVM models using color and texture features achieved high accuracy in detecting citrus diseases. However, the performance of

these systems is limited by the need for manual feature engineering, and they often fail to generalize well to diverse environmental conditions.

1.3.3 Spectral and Sensor-Based Method

More recently, spectral imaging technologies such as hyperspectral, multispectral, and thermal imaging have been applied for early plant disease detection. These methods capture light reflectance across multiple wavelengths, including the visible, near-infrared (NIR), and thermal infrared regions. Diseased plants exhibit distinct reflectance patterns, enabling early detection even before visible symptoms appear (Mahlein, 2016).

Such approaches are commonly integrated into precision agriculture systems, where drones or UAVs collect large-scale field data. Despite their promise, these systems are often expensive, complex to analyze, and require specialized sensors, limiting their accessibility to smallholder farmers.

2 Recent Advances

Recent developments in deep learning have significantly improved plant leaf disease detection. Traditional CNNs such as VGG16 and ResNet are still widely used, but newer approaches focus on improving model accuracy, generalization, and field performance. One major advancement is the integration of attention mechanisms and Transformer-based models, which provide stronger feature extraction and robustness in real agricultural environments (Ashurov et al., 2025). These models help address challenges like varying lighting and complex backgrounds, which often reduce CNN performance in real-field conditions (Nyawose et al., 2025).

Another important trend is the development of lightweight and mobile-friendly architectures optimized for edge computing. Recent reviews highlight that compact CNN models can achieve high accuracy while being efficient enough for deployment on smartphones or IoT devices used in farms (Upadhyay et al., 2025). This allows farmers to detect diseases in real time without needing high-end hardware.

Researchers have also focused on multi-crop and large-scale datasets, enabling models to recognize multiple plant diseases across different species rather than being limited to one crop at a time (Elfouly et al., 2025). This improves the practicality of deep learning models for real agricultural use.

Overall, recent advances emphasize accuracy, generalization, interpretability, and real-world deployment, key insights that guide this project's exploration of custom CNNs and pre-trained architectures.

3 Current limitations

Although deep learning has significantly advanced plant disease detection, current innovations still face several unresolved limitations that are discussed in other research papers like Mohanty et al. (2016). Many studies rely heavily on controlled or laboratory style datasets with uniform backgrounds, making models difficult to generalize to real-world field conditions when lighting, background, or leaf appearance changes. Our approach depends heavily on visual symptoms captured from lab images, field photos and stock images, yet its performance varies across these sources due to domain shift.

Existing models from other papers also struggle with domain shift, meaning a model trained on one environment or imaging condition often performs poorly when tested in another, this reflects a common limitation noted in Ferentinos (2018). Our paper only focuses on identifying whether a plant is healthy or unhealthy based on the symptoms observed by the human eye, but we cannot reliably distinguish abiotic disorders from biotic diseases.

Studies like Mohanty et al. (2016) and Ferentinos (2018) show that accuracy can drop significantly when their models are evaluated outside their original training domain, indicating that many architectures are overfitted to specific datasets rather than truly learning robust, generalizable features. This becomes a serious problem for real-world deployment, because farmers and agricultural stakeholders often need a single system that can handle multiple crops, varying environments and different camera sources. Deploying smartphones or IoT devices used in farms (Upadhyay et al., 2025) can actually help farmers to detect diseases in real time without needing high-end hardware.

Although our dataset includes multiple image sources, it remains limited in scale compared to the large, balanced datasets required for strong deep learning generalization. Many reviews, such as Upadhyay et al. (2025) similarly highlight that insufficient data diversity leads to reduced robustness in real-world agricultural settings.

In addition to data-related issues, computational demands also represent a major limitation. Many high-performing deep learning models require powerful GPUs, large memory capacity and stable internet connectivity, resources that are often unavailable to farmers, smallholder communities and agricultural workers in developing regions. Parez et al. (2023) emphasize that without lightweight and hardware-efficient solutions, the gap between technological innovation and real-world adoption remains wide.

In conclusion, these limitations show that while our approach demonstrates the potential of deep learning in multi-condition environments, the system still requires more diverse data, improved robustness to domain variability and integration of richer visual information to achieve reliable, field-ready performance.

3.1 Known datasets

Singh et al. (2020) proposed plant disease datasets have little diversity due to the laboratory conditions the image datasets were taken in, particularly datasets like the PlantVillage dataset by Hughes & Salathé (2015). Their PlantDoc dataset improves upon this limitation but doesn't account for images with uniform background.

4 Methodology

MegaPlant integrates leaf-image subsets from PlantDoc, PlantVillage, and DiaMOS (Fenu & Malloci, 2021) to produce a model robust across varied imaging conditions. Only leaf images are included. PlantDoc and PlantVillage were obtained from their Kaggle derivatives due to issues accessing the original GitHub repositories.

(a) PlantDoc: nirmalsankalana/plantdoc-dataset
(b) PlantVillage: abdallahalidev/plantvillage-dataset
(c) The DiaMOS dataset was retrieved directly from its official repository.

Table 1. Dataset Image counts

Dataset	Reported size	Retrieved images
DiaMOS	3,901	3,006
PlantVillage	54,305	54,306
PlantDoc	2,922	2,598

We obtained different image counts across repositories due to duplicates, corrupted files, dataset inconsistencies, and untracked modifications present in derivative versions of

the datasets. To ensure consistency, we applied the constraint that only leaf images were included. Accordingly, from DiaMOS, we retrieved images only from the `leaves/` directory; from PlantDoc, only the `train/` and `test/` folders; and from PlantVillage, only the `color/` directory containing colored leaf images.

All images were consolidated and mapped into two primary classes: healthy (0) and unhealthy (1). The unhealthy category contains 12 symptom subclasses: blight, greening, malformation, powdery mildew, feeding, mold, mosaic, rot, rust, scab, scorch, and spot. We define the subclass criteria for dataset integration as follows:

(1) If a folder is labeled with a subclass, all images within are assigned to that class.
(2) Significant changes in leaf shape are classified as malformation.
(3) Changes in leaf color (hue) are classified as greening.
(4) Damage caused by insect feeding is classified as feeding.

An exception is Esca (Black Measles), which, despite its unique pathology, is labeled under the spot subclass. The final dataset was divided into train (70%), validation (20%), and test (10%) splits to enable reliable model evaluation and reduce bias.

4.1 Modeling

We evaluate two approaches for disease detection and symptom identification:

1. Single-model approach: A multiclass classifier that predicts the healthy class and the 12 symptom subclasses.

2. Two-stage approach: A binary classifier first determines whether a leaf is diseased, followed by a multiclass classifier that identifies the specific symptom for diseased samples.

The single-model approach is simpler but suffers from a single point of failure. Any misclassification immediately affects all downstream predictions. In contrast, the two-stage approach compartmentalizes decision-making, allowing flexibility of complexity in downstream models or system architectures, and improving robustness in practical deployments such as UAV-based plant disease surveillance. In many real-world applications, reliably detecting whether a plant is diseased is more critical than precisely identifying the causal agent, which is often better handled by agronomists or plant pathologists.

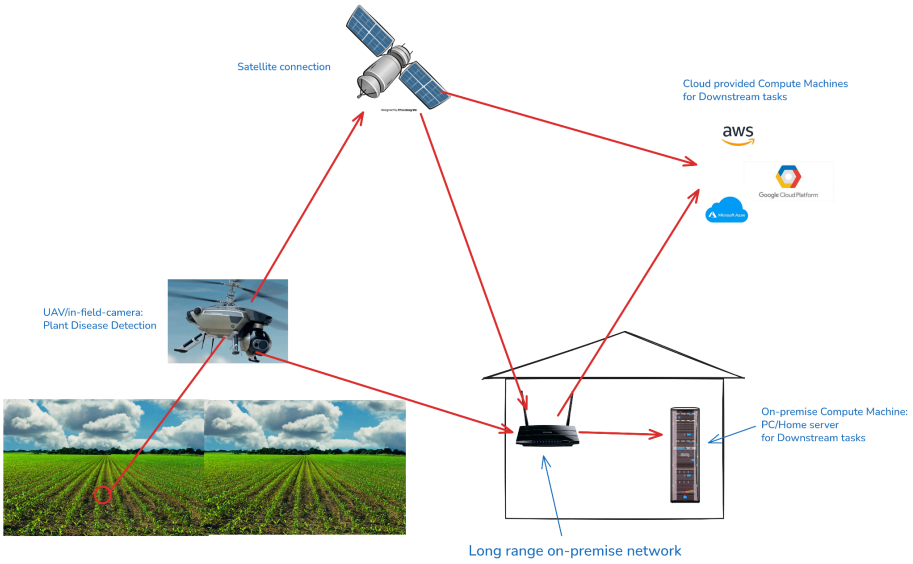


Figure 6. Imagined application of a modular decision framework in a farming business

This structure also supports additional downstream tasks, such as plant species identification or causal agent analysis, and enables clearer interpretability when the pipeline misperforms by isolating which stage produced the error. We additionally compare the proposed approaches against state-of-the-art object detection models to contextualize performance in broader plant disease detection pipelines.

Figure 6 describes a potential application of the modular framework where an unmanned aerial vehicle (UAV) is delegated the task of detecting plant diseases. If it detects any leaves with a disease, it sends the picture to a computer machine where downstream tasks such as symptom or disease identification may be performed. This approach lessens the demand for computing power on edge devices, while also compartmentalizing the decision pipeline.

5 Conclusion

We introduced the basics of plant pathology to give an idea of where our case study might be situated in the research field. We discussed what symptoms are and how our case study approaches the problem of detecting plant diseases by detecting the symptoms visually. We identified traditional and alternative methods of plant disease detection. These methods often required feature engineering steps or expensive requirement

The weaknesses and strengths of innovations in plant disease detection using deep learning were discussed and informed us of how we might tackle the problem of detecting plant diseases using deep learning, particularly on datasets of leaf images with varying conditions such as laboratory, on field conditions, and stock images.

6 Data Availability

The MegaPlant dataset is hosted in a HuggingFace dataset repository can be retrieved from this link: https://huggingface.co/datasets/chrisandrei/MegaPlant.

7 Code Availability

All relevant Python code, jupyter notebooks, notes, and references can be found in this git repository hosted on GitHub: https://github.com/iragca/DS413-final-project.

References

Agrios, G.N. (2009). Plant Pathogens and Disease: General Introduction. In M. Schaechter (Ed.), *Encyclopedia of Microbiology* (pp. 613-646). Elsevier Science. 10.1016/B978-012373944-5.00344-8

Ashurov, A. Y., Al-Gaashani, M. S. A. M., Samee, N. A., Alkanhel, R., Atteia, G., Abdallah, H. A., & Muthanna, M. S. A. (2025, January 23). Enhancing plant disease detection through deep learning: a Depthwise CNN with squeeze and excitation integration and residual skip connections. *Frontiers in Plant Science*, *15*, 1505857. 10.3389/fpls.2024.1505857

Beckerman, J., & Creswell, T. (2022). *Symptoms and Signs for Plant Problem Diagnosis: An illustrated Glossary*. https://www.extension.purdue.edu/extmedia/BP/BP-164-W.pdf

Colson, P., Richet, H., Desnues, C., Balique, F., Moal, V., Grob, J.-J., Berbis, P., Lecoq, H., Harlé, J.-R., Berland, Y., & Raoult, D. (2010, 4 6). Pepper Mild Mottle Virus, a Plant Virus Associated with Specific Immune Responses, Fever, Abdominal Pains, and Pruritus in Humans (E. Mylonakis, Ed.). *PLoS ONE*, *5*(4). 10.1371/journal.pone.0010041

DeBusk, D. (2019, 11 12). *Introduction to Plant Pathogens* [This video provides background on plant diseases and the signs and symptoms common for plant pathogens.]. YouTube. Retrieved 11 12, 2025, from https://www.youtube.com/watch?v=ZM2X-XBRKHM

Dow, M., An, S., & O'Connell, A. (2016). Bacterial Diseases. In B. Thomas (Ed.), *Encyclopedia of Applied Plant Sciences*. Elsevier Science. 10.1016/B978-0-12-394807-6.00051-4

Elfouly, M. K., AbdelAziz, A. M., Gomaa, W. H., & Abdalla, M. (2025, August 21). A deep learning-based framework for large-scale plant disease detection using big data

analytics in precision agriculture. *Journal of Big Data*, *12*(1), 205. 10.1186/s40537-025-01265-9

Fenu, G., & Malloci, F. M. (2021, October 21). DiaMOS Plant: A Dataset for Diagnosis and Monitoring Plant Disease. *Agronomy*, *11*(11), 2107. 10.3390/agronomy11112107

Ferentinos, K. P. (2018, February). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, *145*, 311-318. 10.1016/j.compag.2018.01.009

Fuentes, A., Yoon, S., Kim, T., & Park, D. S. (2021, December 10). Open Set Self and Across Domain Adaptation for Tomato Disease Recognition With Deep Learning Techniques. *Frontiers in Plant Science*, *12*, 758027. 10.3389/fpls.2021.758027

Hughes, D. P., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. 10.48550/ARXIV.1511.08060

The Independent. (2019, December 4). 130 drought affected farmers benefit from crop insurance. https://www.independent.co.ug/130-drought-affected-farmers-benefit-from-crop-insurance/

Mahlein, A.-K. (2016, February). Plant Disease Detection by Imaging Sensors – Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping. *The American Phytopathological Society*, *100*(2), 241-251. 10.1094/PDIS-03-15-0340-FE

Mitiku, M. (2018, 5 21). Plant-Parasitic Nematodes and their Management: A Review. *Agricultural Research & Technology: Open Access Journal*, *16*(2). 10.19080/ARTOAJ.2018.16.55580

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016, September 22). Using Deep Learning for Image-Based Plant Disease Detection. *Frontiers in Plant Science*, *7*, 1419. 10.3389/fpls.2016.01419

Nyawose, T., Maswanganyi, R. C., & Khumalo, P. (2025, September 23). A Review on the Detection of Plant Disease Using Machine Learning and Deep Learning Approaches. *Journal of Imaging*, *11*(10), 326. 10.3390/jimaging11100326

Penn State Extension. (2017, 7 13). *Plant Disease: Sign or Symptom?* YouTube. Retrieved 11 12, 2025, from https://www.youtube.com/watch?v=m6GoSy8RjUM

Pinto, O. R. D. O., Muniz, C. R., Cardoso, J. E., Oliveira, F. S. A. D., & Lima, J. S. (2016, 9). Morphological analyses of Pseudoidium anacardii infecting brazilian cashew plants. *Summa Phytopathologica*, *42*(3), 257-260. 10.1590/0100-5405/2101

Pydipati, R., Burks, T. F., & Lee, W.S. (2006, June). Identification of citrus disease using color texture features and discriminant analysis. *Computers and Electronics in Agriculture*, *52*(1-2), 49-59. 10.1016/j.compag.2006.01.004

Raza, W., Ghazanfar, M. U., Asif, M., Haq, I.-., Zakria, M., & Tawfeeq Al-Ani, L. K. (2022). Morphological Characterization of Phytophthora infestans and its Growth on Different Growth Media. *Sarhad Journal of Agriculture*, *38*(4). 10.17582/journal.sja/2022/38.4.1189.1202

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019, 2 4). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, *3*(3), 430-439. 10.1038/s41559-018-0793-y

Schaad, N. W., Frederick, R. D., Shaw, J., Schneider, W. L., Hickson, R., Petrillo, M. D., & Luster, D. G. (2003, September). Advances in Molecular Based Diagnostics in Meeting Crop Biosecurity and Phytosanitary Issues. *Annual Review of Phytopathology*, *41*(1), 305-324. 10.1146/annurev.phyto.41.052002.095435

Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., & Batra, N. (2020, January 5). PlantDoc: A Dataset for Visual Plant Disease Detection. 249-253. 10.1145/3371158.3371196

Strange, R. N., & Scott, P. R. (2005, 9 1). Plant Disease: A Threat to Global Food Security. *43*(1), 83-116. 10.1146/annurev.phyto.43.113004.133839

UNH Extension. (2015, 6 10). *Guidelines for Diagnosing Plant Problems* [Is your plant suffering from a disease, disorder, insect damage, or something else?... Dr. Cheryl Smith, UNH Cooperative Extension Plant Health Specialist, discusses guidelines for diagnosing plant problems.]. YouTube. Retrieved 11 12, 2025, from https://www.youtube.com/watch?v=7HnLVYhvars

Upadhyay, A., Chandel, N. S., Singh, K. P., Chakraborty, S. K., Nandede, B. M., Kumar, M., Subeesh, K., Upendar, K., Salem, A., & Elbeltagi, A. (2025, January 17). Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. *Artificial Intelligence Review*, *58*(3), 92. 10.1007/s10462-024-11100-x

Ward, E., Foster, S. J., Fraaije, B. A., & McCartney, A. H. (2004, August). Plant pathogen diagnostics: immunological and nucleic acid‐based approaches. *Annals of Applied Biology*, *145*(1), 1-16. 10.1111/j.1744-7348.2004.tb00354.x

Parez, S., Dilshad, N., Alanazi, T. M., & 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. https://doi.org/10.32604/csse.2023.037992