# The evolution of plant disease detection techniques

Martin BORDES

April 2023

**Abstract**

All around the world, diseases and pests affect a lot of major crops. It results in a loss of quantity and of quality for this crops. So, early disease detection is really important topic in agriculture. For centuries, this work was done manually. But this is really fastidious and time consuming or even immpossible for the largest crops. However, for some years, several different techniques were developped in order to automate this task. The goal of this theoretical work is to show the evolution of this techniques through time. We'll first see some techniques that do not use any machine learning. Even if this is not in my area of expertise, it can help to understand the overall problem. Then, we'll look at how it appeared for this particular task. Finally, we'll talk about the latest innovations in plant disease recognition using deep learning.

# Contents

# 1 Introduction

During the main part of the agricultural history, it was very difficult to detect plant diseases. It didn't exist any particular tool for this task. This work was done entirely manually and required some knowledge and experience. Furthermore, it is sometimes impossible to early detect diseases only manually. So, farmers were not able to monitor anything. It was probably already too late when a problem was diagnosted. Even in our modern world, diseases remain a huge problem in agriculture around the world. The estimated losses in crop yield due to pathogen infections range between 20% and 40%. In the context of a growing global population and a decreasing productivity of agriculture land area, it contributes to a global food security crisis. For the United States alone, the annual losses are around 40 billion dollars so the problem is also economic for a lot of farmers. Finally, the last issue that could be resolved by effective solutions of early disease detection is the abusive use of products such ass pesticides and fungicides. There could be used in a much proper way if the farmers where to target this solution.

To adress this major issue in agriculture, a lot of researchers began to look at it in the last century. They developped a lot of sophisticated methods, using several scientific domains. All of the techniques used to detect plant diseases can be splitted in two categories : the direct ones and the indirect ones. The first category contains all the methods that detect the pathogens, such as bacteria, fungi and viruses, while the second one contains the techniques based on plant stress profiling and plant volatile profiling, that are important indicators of pathogens. For some years, artificial intelligence is part of the second category.

# 2 Direct detection methods

Most of the direct detection methods are molecular. The most used of them has become part of our vocabulary for three years, it is the Polymerase Chain Reaction, better known as PCR. Two Nobel prizes were awarded for works using PCR, in 1984 and 1993. It is based on DNA hybridation and replication. Originally, PCR was used for highly specific human pathogens. It is still for human diseases as we all know but also for plant pathogens for several years. PCR technique provides a high sensitivity and specificity but lacks of operational robustness. It needs a very efficient DNA extraction, can be affected by inhibitors and requires designing a primer to initiate DNA replication. This limit the possible practical field applications.

Fluorescence In-Situ Hybridization (FISH) is an other molecular technique. It can recognize the presence of pathogen-specific ribosomal RNA (rRNA) sequences in plants. High sensitivity is offered by the high affinity and specificity of DNA probes. However, false positives are a common problem with FISH. This limits the technique's potency for plant disease detection.

One last molecular approach is the enzyme-linked immunosorbent assay (ELISA). It is based on antibodies and color change in the assay. The antigens from the viruses, bacteria and fungi specifically bind with antibodies conjugated to an enzyme and the detection can be visualized by color change. This technique is very practical but not sensitive enough. It can only be used for confirmation after the appearance of visual symptoms. It does not fit our early detection problem.

Immunofluorescence (IF) is an optical method used for the analyses of microbiological samples. Plant samples are first fixed to microscope slides in thin tissue sections. Then, we are able to visualize the distribution of target molecule by conjugating a fluorescent dye to the specific antibody. However, photobleaching is a common problem in IF that causes false negative results.

All the direct detection methods we talked about have different pitfalls that limit their use in practice. Furthermore, these techniques are too much time consuming, they require a lot of expertise and most of them need to be done in laboratory. They remain useful to ensure the presence of disease but it is not sufficient to deliver an efficient monitoring that can provide an early detection. We can hope to find such a solution among the indirect detection methods, where machine learning is used.

# 3 First uses of machine learning for final classification

The three main categories of direct detection methods are thermography, fluorescence and hyperspectral approch. Machine learning is used for classification in all of these domains but we'll only talk about hyperspectral techniques here. In this area, high quality optical sensors provide a multiplicity of information over the covered spectral range. These tecnhiques record large amounts of information on the object acquired at the same time. To use these data, we need powerful methods of analysis for the detection, differentiation and quantification of plant diseases. Many techniques have been developed in order to obtain maximal information from hyperspectral data and images. Simple statistical methods were first used. Then, more complex ones were introduced. We'll look at four of them.

The first classification technique we are going to see is Spectral angle mapping (SAM). **?** used them to detect late blight in tomatoes. They obtained remote sensing hyperspectral data from low-altitude flights. They first applied the Minimum noise fraction (MNF) transformation. This procedure assume that each pixel contains both signal and noise. It allows to determine which bands are more useful to discriminate the diseased plants from the healthy ones. After having removing the noise, they can perform the classification. In order to do that, they used SAM. This technique determines the similarity between two spectras by calculating the angle between them. It treats a spectra as a vector where the dimensionality is the number of bands. This way, SAM can compare the angle between the reference spectrum vector and each pixel vector in the n-dimensional space. The angle is calculated as follows :

$$SAM = \cos^{-1}\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2}}\right) \tag{1}$$

where $X$ and $Y$ are respectively the pixel vector and the reference spectrum vector. The smaller the vector angle, the closer to the reference spectrum. It allows to group them into clusters represented by the reference spectra.

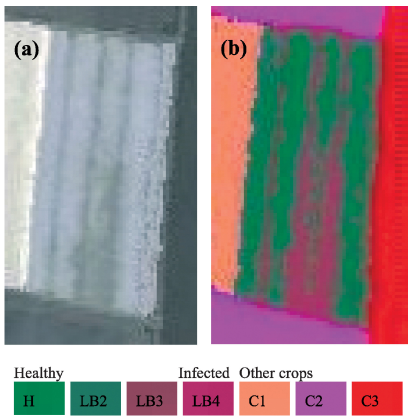


Figure 1: Results of classification to the hyperspectral image for identification of infected plants from the healthy ones (Zhang et al. (2003))

However, Bauriegel et al. (2011) showed the limits of the SAM method. Indeed, despite this technique has shown very promising result, it is a very time-consuming procedure. Furthermore, it suffers from its complexity, with 512 bands, and the need to set up regions of interests makes this method not effectively applicable outdoor. The authors proposed a new approach using only 168 wavelenghts bands, thanks to spectral subsets, and principal component analysis (PCA).

Even if we tend to see neural networks as recent methods for this task, they were already used by Moshou et al. (2004). As well as in the last paragraph, they first needed to choose which bands were the most useful to discriminate between the healthy plants and the diseased ones. In order to do that, they conducted a stepwise procedure using statistical F-tests. A waveband would only be selected when its addition to an existing set of selected wavebands would significantly increase the discriminating power of the new set. The threshold probability was fixed to 0.15. Then, the classification phase was based on a multilayer perceptron. The number of units in the input layer was the number of bands selected by the stepwise procedure. The two classes (healthy and diseased) were represented by two neurons in the output layer. There was one hidden layer. They tested to use different numbers of neurons in it, from 5 to 25. There was no improvement using more than 10 units. They compared the results to a discrimination model built by quadratic discriminant analysis, based on the Bayesian decision rule. An object is classified to the class which maximises the posterior probability. The multilayer perceptron results were way better. The correct recognition of healthy plants was improved by 7%, which means that false treatment was minimised. The main issue caused in this article was the misuse of products such pesticides. Such an improvement can really help to better target the diseased plants.

Rumpf et al. (2010) compared decision trees, artificial neural networks and support vector machines (SVM) for this task of plant disease recognition. It appeared that SVMs were the best tool for automatical classification. The article showed that SVMs use the inherent information of the vegetation indices in an optimal way. Not only the identification of diseased leaves, but also the differentiation between distinct diseases can be realised. The best results of SVMs compared to ANNs is explained by the authors from a therotical perspective. SVMs must solve a quadratic optimization problem and convexity ensures a unique and global solution. On the other hand, ANNs optimization is not convex so there are generally multiple local optima. Furthermore, by using empirical risk minimization, ANNs tend to overfit while SVMs use structural risk minimization.

The studies we reviewed until now are all at least 10 years old. And if there is one area that has experienced an impressive improvement during the last decade, this is machine learning and especially deep learning. While it was only used for final classification in the past, machine learning became the main tool in some plant disease detection approach in the last years.

# 4 Advanced machine learning techniques

For some years, machine learning for plant disease detection has been widely studied and has shown promising results. It is mainly used as part of image processing. While with the other indirect methods we saw, features were created by hyperspectral imaging, it is now possible to use raw images. Machine learning techniques allows to do feature extraction from these images after some preprocessing. The first studies in this domain were conducted in the last 2000's.

Camargo and Smith (2009) were one among the first to propose such an approach. However, they did not use a machine learning method to extract their features. They used several mathematical techniques such as fractal dimension, co-occurence matrix or lacunarity. The resulting features characterized the shape and appearance of the image, such as grey levels, connectivity and tex- ture. Then, these features were used as inputs for the classifier, which was a SVM. Even more than a decade after this paper, a lot of works on image processing and machine learning for plant disease detection nowadays use SVM as their classifier (Araujo and Peixoto (2019), Aziz et al. (2019), Abdu et al. (2020)). It can probably be explained by what was enounced in Rumpf et al. (2010), as we saw earlier. However, the superiority of SVM over neural networks is being truly challenged recently.

Indeed, deep learning is the most studied area in a lot of domains. Plant disease detection, as part of computer vision, is no exception. The main type of ANN used in computer vision is the Convolutional neural network (CNN). CNNs are simply multi-layer perceptrons, with the addition of a convolutional part. It applies some operations to the original image, such as filters and max-pooling. In fact, there is no real opposition between CNNs and SVMs (or other classical classifiers) because a lot of studies use both of them. The original images first pass into a CNN to extract some features. Those features can then be used as inputs for a classifier such as a SVM (Khan et al. (2018), Pardede et al. (2018)). They even can be used in the inverse order (Hu et al. (2019)).

However, CNNs can also obviously used alone, for both feature extraction and classification. Mohanty et al. (2016) conducted a pioneering work in this domain. They trained a deep CNN on 14 crop species and 26 different diseases. They used a very important notion in deep learning and especially in computer vision : transfer learning. It consists of taking a neural network (i.e. all its weights) already trained on a huge dataset. This neural network has to have been trained on a task close to the one you are working on. To adapt this network to your task, you need to remove the last layers and add new ones. Then, you use your dataset to train these new weights, this is fine-tuning. This technique is widely used in computer vision because there exist several cutting-edge models in this area, such as AlexNet, GoogleNet, VGGNet and ResNet50. Here, the authors tested two of them in multiple conditions : AlexNet and GoogLeNet. They also tried to train AlexNet and GoogleNet models from scratch to compare to transfer learning. The result is clear, transfer learning is way more efficient than training from scratch. The reason is that those base models were trained on huge datasets. This is impossible to find such data for a very particular task such as plant disease recognition. We can also point out that GoogLeNet performed slighly better, probably because it is more recent.

Ferentinos (2018) used an open database of 87,848 images, with 25 different plants 58 distinct classes of [plant, disease]. The objective was to determine the good [plant, disease] combination. They tested five different model architectures in order to compare them : AlexNet, AlexNetOWTBn, GoogLeNet, Overfeat and VGG. The best results were achieved by the VGG network, with a success rate of 99.53% in the classification of 17,548 previously unseen by the model plant leaves images.

Figure 2: Samples of tomatoes images in laboratory conditions (up) and in field conditions (down) (Ferentinos (2018))

Yadav et al. (2021) is also an interesting work, this time on peach leaves. As often with computer vision, they conduct a preliminary phase of data augmentation. It consists of creating new images by applying transformations to the existing ones. The transformations used in this article were rotation, color and brightness change, addition of noise, random cropping, background removal and zooming. Here, the authors combined three different transfer learning models : AlexNet, VGGNet and YOLO-v3. With this approach, they reached an accuracy of 98.75%.

As we can see, some CNNs have reached very impressive level of accuracy. We can also cite Araujo and Peixoto (2019) who reached an accuracy of 99.32% on soybean. However, they first faced overfitting because their dataset, containing 12.673 leag images, was too small. To overmcome this issue, data augmentation proved to be a very effective tool. They also used weight penalties. For this purpose, introducing dropout was more effective than using L2 regularization. Dropout consist of ignoring randomly selected neurons during training to prevent overfitting.

However, even if CNNs results are very promising, the path to a final and easily usable solution is not over. There are still several issues to overcome to solve the problems of pesticides misuse and crop losses.

# 5 Limits and future developments

Having a very effective plant disease recognition model is not sufficient for the solution to be used in practice. The first issue the researchers will have to solve is the variety of species. A lot of studies we talked about in this paper only focused on a single plant an few different diseases. For a solution to be easily used by a lot of farmers, it needs to be very generalized to as much plants and diseases as possible. Even Ferentinos (2018), that used 58 different classes of [plant, disease] and claimed to have conducted the largest plant disease identification task tackled with deep learning methodologies at the moment, present the expansion of the existing database to incorporate a wider variety of plant species and diseases as the immediate future step.

The authors also point out an other issue concerning the data. In this kind of study, the testing set comes from the same database as the training set. In a real world situation, the testing data would be provided directly by the farmers, while they would use models trained on some database that does not contain their crop. Some experiments from the authors showed that doing that caused significant reduction of their accuracy, in the range of 25-35% depending on the data. To solve this problem, the training database used should be made from a much wider variety of data sources. It should represent more geographical areas, cultivation conditions and image capturing modes. A lot of studies used common plant database, such as PlantVillage and Embrapa. According to Thakur et al. (2022), these database provide limited variance and their image capturing modes were very controlled.

The last issue concerning databases is the scarcity of in-field datasets. To be really suitable for a real world situation, plant disease detection models should be trained on in-situ images. Indeed, to be able to monitor and early detect diseases, farmers would provide in-field images to the designed solution. The model has to be trained on data closed to what they will have to monitor in practice. Some studies have already been conducted in this way like Picon et al. (2019). A real world application of a plant disease detection model would be part of the Internet of Things (IoT). All the current best methods for this task are deep CNNs that are characterized a large memory demand. These solutions do not fit an IoT application. To solve this issue, some researchers have worked on shallow networks that require less memory. Lu et al. (2017) were among the first ones to work in that direction, proposing a CNN with only 5 layers. However, the majority of studies conducted on shallow networks used small datasets. It could reinforce the issues we already discussed.

In order to improve the performance of CNNs for the plant disease detection task, researchers have started to explore attention mechanism. This technique has experienced a lot of hype in the last years in most areas and for this task as well. It allows the network to focus on some parts of the input depending of the context. This is meant to mimic the cognitive attention of the human brain. It is widely used in natural language processing. Here, the context will be defined by some features that will adjsut the weights. Some studies have already been conducted in that direction (Tang et al. (2020), Chen et al. (2021), ...) but this could the future of plant disease recognition.

# 6 Conclusion

The goal of this theoretical work was to review the evolution of techniques in plant disease recognition, an essential task in modern agriculture. In the last decades, this task has known a lot of changes. Not so long ago, the only way to detect plant diseases was manually. Then, two types of methods were developped at the end of the last century : the direct and the indirect ones. We described some of direct methods but the progress were more important for the indirect ones in the last years.

The rapid development of machine learning made possible to include more sophisticated classification modules in those techniques, instead of more simple statistical ones. In the last decade, it has gone further thanks to the impressive improvements in deep learning. Some cutting-edge models are now available for transfer learning, easily customizable to this task. This task is now seen as a computer vision problem. In this area, the dominent models are the convolutional neural networks and plant disease recognition is no exception. The technological advancements have made CNNs the main domain of interest for the researchers on this problem. For good reason, those models have reached spectacular levels of performance in the last years.

However, there still are several issues to overcome. The researchers will have to build mmore complete datasets in order to obtain a generalized solution. This solution will then have to be usable easily by farmers, probably as part of IoT. So, it has to be lighter than the current best methods.

# References

Abdu, A. M., M. M. Mokji, and U. U. Sheikh (2020): "Automatic vegetable disease identification approach using individual lesion features," *Computers and Electronics in Agriculture*, 176, 105660.

Araujo, J. M. M. and Z. M. A. Peixoto (2019): "A new proposal for automatic identification of multiple soybean diseases," *Computers and Electronics in Agriculture*, 167, 105060.

Aziz, S., M. Bashir, O. Mughal, M. U. Khan, and A. Khan (2019): "Image Pattem Classification for Plant Disease Identification using Local Tri-directional Features," in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 0973–0978.

Bauriegel, E., A. Giebel, M. Geyer, U. Schmidt, and W. Herppich (2011): "Early detection of Fusarium infection in wheat using hyper-spectral imaging," *Computers and Electronics in Agriculture*, 75, 304–312.

Camargo, A. and J. Smith (2009): "Image pattern classification for the identification of disease causing agents in plants," *Computers and Electronics in Agriculture*, 66, 121–125.

Chen, J., D. Zhang, A. Zeb, and Y. A. Nanehkaran (2021): "Identification of rice plant diseases using lightweight attention networks," *Expert Systems with Applications*, 169, 114514.

Fang, Y. and R. P. Ramasamy (2015): "Current and Prospective Methods for Plant Disease Detection," *Biosensors*, 5, 537–561.

Ferentinos, K. P. (2018): "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture*, 145, 311–318.

Golhani, K., S. K. Balasundram, G. Vadamalai, and B. Pradhan (2018): "A review of neural networks in plant disease detection using hyperspectral data," *Information Processing in Agriculture*, 5, 354–371.

Hu, G., H. Wu, Y. Zhang, and M. Wan (2019): "A low shot learning method for tea leaf's disease identification," *Computers and Electronics in Agriculture*, 163, 104852.

Khan, M. A., T. Akram, M. Sharif, M. Awais, K. Javed, H. Ali, and T. Saba (2018): "CCDF: Automatic system for segmentation and recognition of fruit crops diseases based on correlation coefficient and deep CNN features," *Computers and Electronics in Agriculture*, 155, 220–236.

Lu, Y., S. Yi, N. Zeng, Y. Liu, and Y. Zhang (2017): "Identification of rice diseases using deep convolutional neural networks," *Neurocomputing*, 267, 378–384.

Mahlein, A.-K., E.-C. Oerke, U. Steiner, and H.-W. Dehne (2012): "Recent advances in sensing plant diseases for precision crop protection," *European Journal of Plant Pathology*, 133, 197–209.

Mohanty, S. P., D. P. Hughes, and M. Salathé (2016): "Using Deep Learning for Image-Based Plant Disease Detection," *Frontiers in Plant Science*, 7.

Moshou, D., C. Bravo, J. West, S. Wahlen, A. McCartney, and H. Ramon (2004): "Automatic detection of 'yellow rust' in wheat using reflectance measurements and neural networks," *Computers and Electronics in Agriculture*, 44, 173–188.

Pardede, H. F., E. Suryawati, R. Sustika, and V. Zilvan (2018): "Unsupervised Convolutional Autoencoder-Based Feature Learning for Automatic Detection of Plant Diseases," in *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 158–162.

Picon, A., M. Seitz, A. Alvarez-Gila, P. Mohnke, A. Ortiz-Barredo, and J. Echazarra (2019): "Crop conditional Convolutional Neural Networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions," *Computers and Electronics in Agriculture*, 167, 105093.

Rumpf, T., A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne, and L. Plümer (2010): "Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance," *Computers and Electronics in Agriculture*, 74, 91–99.

Sankaran, S., A. Mishra, R. Ehsani, and C. Davis (2010): "A review of advanced techniques for detecting plant diseases," *Computers and Electronics in Agriculture*, 72, 1–13.

Tang, Z., J. Yang, Z. Li, and F. Qi (2020): "Grape disease image classification based on lightweight convolution neural networks and channelwise attention," *Computers and Electronics in Agriculture*, 178, 105735.

Thakur, P. S., P. Khanna, T. Sheorey, and A. Ojha (2022): "Trends in vision-based machine learning techniques for plant disease identification: A systematic review," *Expert Systems with Applications*, 208, 118117.

U, S., V. Nagaveni, and B. Raghavendra (2019): "A Review on Machine Learning Classification Techniques for Plant Disease Detection," 281–284.

Wallelign, S., M. Polceanu, and C. Buche (2018): "Soybean Plant Disease Identification Using Convolutional Neural Network," in *FLAIRS-31*, Melbourne, United States, 146–151.

Yadav, S., N. Sengar, A. Singh, A. Singh, and M. K. Dutta (2021): "Identification of disease using deep learning and evaluation of bacteriosis in peach leaf," *Ecological Informatics*, 61, 101247.

Zhang, M., Z. Qin, X. Liu, and S. L. Ustin (2003): "Detection of stress in tomatoes induced by late blight disease in California, USA, using hyperspectral remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, 4, 295–310.