Kristianstad University
SE-291 88 Kristianstad
+46 44-250 30 00
www.hkr.se

Machine Learning for Plant disease diagnostics

Author: Meron Habtemichael

Research methodology in computer science course

Author (style Abstract Headline)

Meron Habtemichael

Title

Machine Learning for plant disease diagnostics

Supervisor

Examiner

Charlotte Sennersten

Abstract (maximum 250 words)

Plant diseases is a principal factor of determining the yield and quality of plants and its identification can be conducted by means of digital image processing. In recent years, Deep learning has made breakthroughs in the field of computer vision, far superior to traditional methods. How to apply Machine Learning to study plant diseases has become an issue of interest among researchers. This research reviews an application of Machine Learning algorithms in order to solve the problem of plant diseases to help farmers achieve their maximum results out of farming.

Keywords

Machine Learning, Crop monitoring, Plant diseases, Deep Learning, Convolutional Neural Networks

Content

1. Introduction

1.1 Context

Artificial intelligence (AI) is currently one of the treading fields under intense research in the field of computer science. Media coverage and public debate on AI is impossible to avoid. Still, you might also have noticed that AI means different things for different people. For some, AI is all about artificial life-forms that can overcome human intelligence, and for others nearly any data-processing technology may be called AI. In simplest terms, AI is a collection of concepts, problems, and methods of solving them [2]. The use of machines to follow these methods makes the job faster in comparison to humans. This subject has been applied to many industries such as in Finance, Medicine, Geography, Agriculture, and many others to deal with complex problems containing big data and achieve effective solutions. Machine Learning is also a subset of study within Artificial intelligence that helps to solve specific tasks through learning from data and making prediction.

This research goes into some of the major applications of Machine Learning in the sphere of Agriculture. In accordance with the Food and Agriculture organization of the United Nations, the world population is projected to reach over nine billion by 2050 [1]. This issue plus diminishing farmlands, dwindling natural resources, erratic climate changes, and changing market demands are pushing the agricultural production system into a new form [1]. The new agricultural system will have to become more productive in output, effective in operation, resistant to climate change and more sustainable for generations to come. Machine Learning offers the hope in addressing the challenges of this new problem.

Machine Learning provides enormous assistance in the vast field of Agriculture such as in crop and animal management in production and protection, natural resources sustainability, food nutrition and food safety and more. This is a wide range of applications to discover but this research work dives into crop production and management with the help of Machine Learning techniques to harvest maximum results to meet market needs.

Crop Monitoring i.e., Tradition health monitoring techniques are hard labor-intensive and time-consuming. Developing a machine learning model is an efficient way to track and identify potential crop health issues or nutritional deficiencies in the soil [1]. Thanks to the help of Deep Learning, applications are currently being developed to analyze plant health patterns in farming. Plant diseases are a major threat to the environment that can be caused by different cases. They can bring massive damage to the environment, economy, and food security. This is one of the primary reasons why Machine Learning is a winning choice to put into practice. It can identify crop diseases in its initial stages to avoid horrific losses [1].

1.2 Research Aim

This research tries to identify the use and implementation of Machine Learning algorithms for plant diseases diagnostic purposes. It tries to go into Deep Learning and review the overall process on how to implement the algorithm relating computer science to the agriculture industry. The main idea is to reflect the benefits of AI in agricultural crop management to meet the growing demand of current marketplace.

1.4 Research Question

How can Machine Learning help us identify plant diseases, and which algorithm is the most preferable?

1.5 Variables

1.5.1 Plant diseases

Plant diseases are one of the biggest threats and highly correlated to crop harvesting results. It is one kind of natural disasters affecting natural growth of plants and even lead to plant mortality during the entire growth process of plants from seed development to seedling and to seedling growth. Throughout the world, their rapid identification remains difficult because of the lack of the necessary infrastructure. In the developing countries, from around the 80 percent of the agricultural production generated by smallholder farmers, more that 50% yield is lost due to pests and plant disease. This shows that there is a crucial need for better techniques of identifying a disease correctly when it first appears.

1.5.2 Crop yield

Crop yield prediction is a crucial factor in food security assessment and policy making. Crop growth and yield data monitoring is a tedious task in cultivation. Farmers are more concerned about yield prediction based on many factors like climate, soil condition, fertilizers, and diseases. Big data of these mentioned factors has been collected over the years and the job of processing them has been tedious.

Machine Learning and Data Mining are among the heavily used technologies to solve some of the main agricultural problems. Image recognition has a big role in this scenario. This research paper tries to find the application of image classification algorithms in Machine Learning to deal with identifying plant diseases in order to achieve maximum results during crop harvesting.

2. Methodology

There are various image recognition algorithms in Machine Learning and in comparison, with other methods, Deep Learning does not need to extract specific features [5]. Through interactive learning only, Deep Learning can find appropriate features, which can obtain global and contextual features of images, and has strong robustness and higher recognition accuracy [5].

2.1 Preparing and choosing dataset

In short, because of the primary task being to identify plants with diseases, the two things that may potentially go wrong are “bad algorithm,” and “bad data.” Deep learning is not there yet, meaning the technology is not as complete as it can be since it is developing every time. It requires lots of data for most Machine Learning algorithms to work properly [4]. Usually, a Large and verified dataset of both diseased and healthy plants is needed for training accurate image classifier algorithms.

2.1.1 Pre-processing

This process is working as a preprocessing for image recognition. In this step the unwanted noise, blur, varying lightening condition, shadowing effects can be removed using pre-processing techniques. Unfortunately, If the training data is not clean, it will make it more difficult for the system to detect the underlying patterns, so the entire system is less likely to perform well. Truth is that most data scientists spend a substantial part of their time doing just that. As the saying goes, garbage in, garbage out, the Machine Learning model will only learn and be successful after having the training data with enough relevant input data and not too many irrelevant one [2].

2.1.2 Splitting the data

after doing all that pre-processing on the prepared data, it is now time to split it into two sets, the training and test sets. As their names imply, the model is trained using the training set and evaluated using the test set.

2.2 Which Algorithm to use

Many efforts have been developed and worked on to prevent crop loss due to diseases. Computer vision and object recognition has made tremendous advances over the years. Deep neural networks are successfully used across many different domains as examples of end-to-end learning. Neural networks offer mapping between an input to an output and are made of nodes which are nothing but mathematical functions that take numerical inputs and deliver a numerical output as an outgoing edge [5]. Deep neural networks are merely mapping the input later to the output layer through a series of stacked layers of nodes. The main objective of creating a deep network is to design a model that can get an input and accurately move across the nodes just to map the correct output. These algorithms are trained by tuning the parameter variables to improve the mapping during the process.

CNN is favored as a deep learning method in this study. CNN, which can very easily detect and classify objects with minimum pre-processing, is effective in examining visual images and can easily detach the required features with its multi-layered structure [5]. This method reduces the number of parameters and speed up the training of the model significantly.

It is basically a partially connected layer and usually have 3 layers [3] i.e.

1. Convolution layer: This layer basically set up a feature map that summarizes the availability of the discovered features in the input.
2. Pooling layer: The goal here is to reduce the size of the image
3. Activation layer: this layer provides a curvilinear relationship between the input and output layers and affects the network performance.

Fully Connected Layer: After going through all these 3 layers, the most recent obtained matrix is fed into the fully connected layer as input. Recognition and classification are done in this layer.

2.2.1 Overfitting and Underfitting the training data

Overfitting is a case happened if the model performs well on the training data but does not generalize well. When the model is just too simple to learn the underlying structure of the data, the reverse of overfitting, is called Underfitting. Deep Neural Networks are complex models and can detect subtle patters in the data [2]. But if the training data contains a lot of is noises or we supply insufficient training data, then the model is likely to detect patterns in the noise itself. This problem can be reduced quite dramatically by carefully pre-processing the data set and supplying more training data to the algorithm.

2.2.2 Method Example

This research goes into a simple example of following a method for plant leaf disease detection and classification. We can observe some datasets from PlantVillage dataset who incidentally offers plenty of useful datasets available which can be used to train various kinds of Machine Learning models [1]. One can take 500 tomato leaf images, 400 training and 100 test images. The images are carefully selected against noises and cropped to size of 512*512 [1].

There are four intended leaf diseases to classify from the dataset, and they are:

1. Bacterial Spot: The symptoms begin as small, yellow-green lesions or as dark, water soaked, greasy-appearing lesions on leaves. This type of disease is spread very quickly and is exceedingly difficult to control. It is indeed one of the most threatening tomato diseases [1].

2. Late Blight: It is seen as large brown spots with green-grey edges. As the disease matures, the spots become even darker. In the long run this disease seriously damages the whole plant [1].

3. Septoria Leaf Spot: The symptoms appear in the lower leaves of the plant with round, yellow, water-soaked spots. Causes small brown and black spots [1].

4. Yellow Leaf Curl: This causes the leaves to bend downwards, stiff, ticker and slightly yellowish [1].

Fig.1 shows sample images of a healthy leaf and diseased leaves [1].

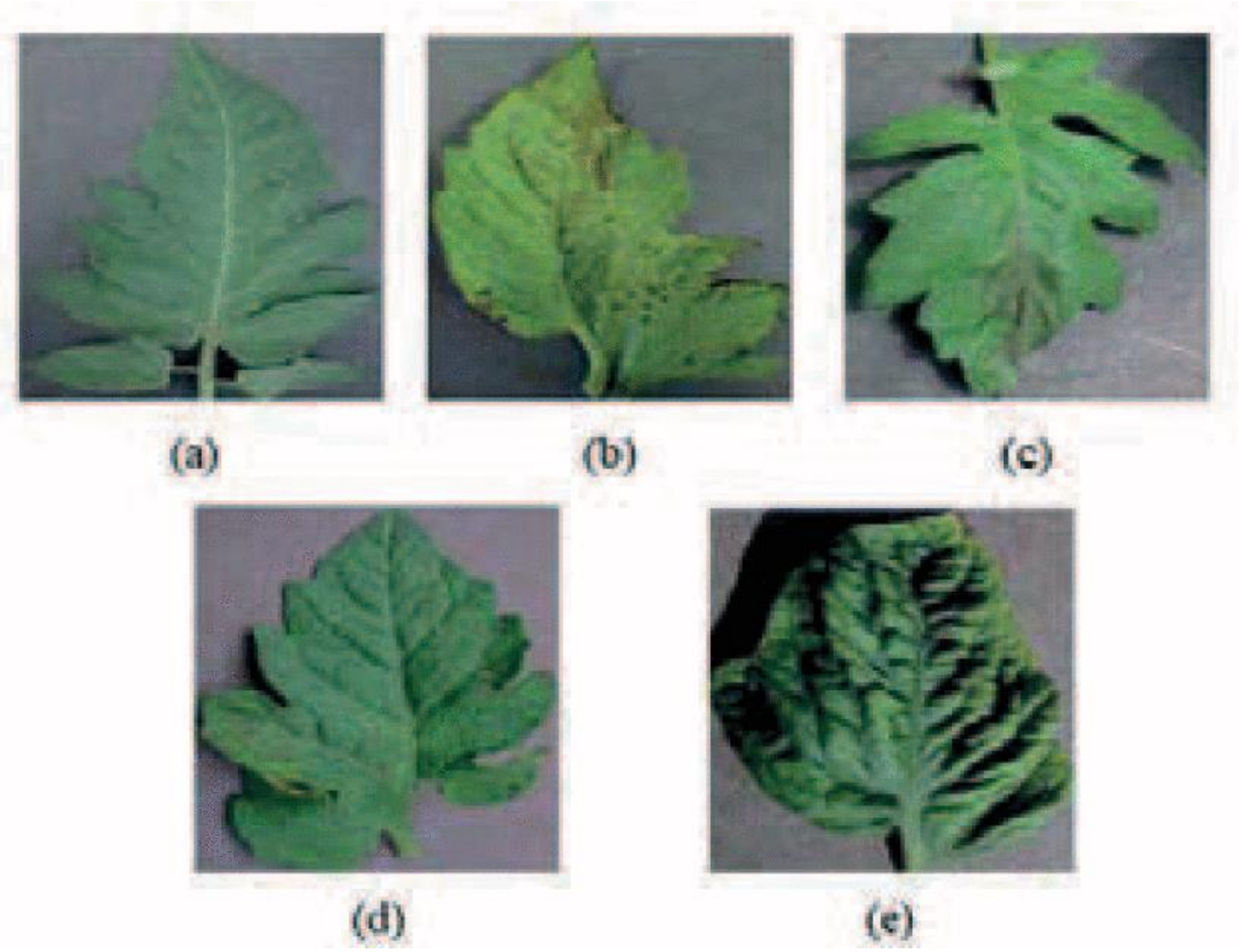

Fig.1. (a) Healthy Leaf (b) Bacterial Spot (c) Late Blight (d) Septoria (e) Yellow Leaf Curl

The next step is to visualize the data more closely and see if there are any noises or unclarity. Then we can consider going ahead and feed the data to the model. The model used the RGB notion of each pixel in a leaf image and forms a 3*3 Matrices [1]. The image is then passed across the nodes and in the final layer i.e., the Fully Connected Layer, the image is finally classified.

Fig.2. shows a typical CNN architecture [2].

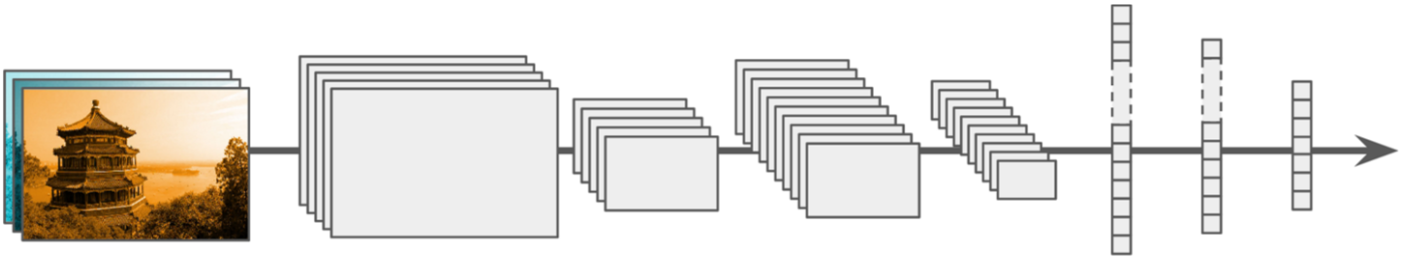

Fig.2 Typical CNN architecture.

Popular frameworks in python like TensorFlow and PyTorch can help with this theoretical concept into an actual implementation. Scikit-learn can also be used and have many utility functions like cross-validation and grid search.

2.2.3 Measurement of performance

It is vital to understand if the algorithm learns the “notion” of plant diseases or if it just learns the inherit biases in the dataset. This is identified by a set of good scientific performance testing widely used in Machine Learning. For instance, in the upper simple example, one of the main challenges can be that the leaves with different diseases are similar to each other so the algorithm may have some tough time detecting or classifying the plant disease [5]. But remember, more data is greatly beneficial in terms of the model’s performance.

3. Conclusion

In brief summary, with the development of the field of AI, the research focus of plant diseases detection based on computer vision has shifted from classical Image processing and Machine Learning methods to Deep Learning methods which gave rise solution of difficult problems that previously could not be solved via traditional methods. There is still a long way to for this technology to be heavily populated and applied across the agricultural industry but still there is a great development potential and application value. In order to completely investigate the potential of this technology, collaborative effort of experts from relevant disciplines is needed to effectively integrate the experience knowledge of agriculture with Deep Learning algorithms and models so that the methodology of plant diseases detection based on Deep Learning becomes mature.

References

1. Anon., 2018. *Plant Leaf Disease Detection and Classification Based on CNN with LVQ Algorithm.* [Online]
   Available at:
   https://ieeexplore.ieee.org/abstract/document/8566635?casa_token=4PEg9l6wpp4AAAAA:zaj1JX8S5wokMW2i2bJEQe9XYCtC4JMWQ70xw5muJy57BqpfB8rRRlEBz8jADHJ8SnUfhkUeSv3Q4w
   [Accessed 2022].

2. Géron, A., 2017. In: *Hands-On Machine Learning with Scikit-Learn and TensorFLow.* s.l.:s.n.

3. Liu, S. Y., 2020. *Artificial Intelligence (AI) in Agriculture.* [Online]
   Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9098011
   [Accessed September 2022].

4. L, S. G., 2022. *IEEE Xplore.* [Online]
   Available at:
   https://ieeexplore.ieee.org/abstract/document/9864482?casa_token=dduRXHNJGIQAAAAA:M81xj3QWk_E9uchWeycLcIQAN9-TF5e_fzPmHuxlpBeUoJwYAJ7nlb25m-Ghz2_1tldmncCctBbhug
   [Accessed 2022].

5. Wang, J. L. &. X., 2021. *SpringerLink.* [Online]
   Available at: https://link.springer.com/article/10.1186/s13007-021-00722-9#Bib1
   [Accessed 2022].

Additional Notes

Thesis plan is to work on this project and plus some additional applications of machine learning in the field of Crop Management with T-Kantor.

Reflection

Thanks for your feedback charlotte!

I have run through a grammar check but let me know if there is anything more you would like me to fix!

My thesis is now well planned and have a picture of what I want to work with. I know you are familiar with it. Andreas and Åsa has connected us with the Lund project and we met the professor (Mathematics) via Zoom. Next thing is to send him our CV and Åsa has told me that the professor has some missions for us. This puts the whole process of thesis planning successful which is the main objective of the course.