CLASSIFICATION OF PLANT LEAF DISEASES USING MACHINE LEARNING AND IMAGE PREPROCESSING TECHNIQUES

Pushkara Sharma
Department of Computer Science and Engineering
Amity University
Noida,India
pushkarasharma11@gmail.com

Pankaj Hans
Department of Computer Science and Engineering
Amity University
Noida,India
pankajhans91@gmail.com

Subhash Chand Gupta
Department of Computer Science and Engineering
Amity University
Noida,India
scgupta@amity.edu

***Abstract*— Agriculture is one of the main factor that decides the growth of any country. In India itself around 65% of the population is based on agriculture. Due to various seasonal conditions the crops get infected by various kind of diseases. These diseases firstly affect the leaves of the plant and later infected the whole plant which in turn affect the quality and quantity of crop cultivated. As there are large number of plants in the farm, it becomes very difficult for the human eye to detect and classify the disease of each plant in the field. And it is very important to diagnose each plant because these diseases may spread. Hence in this paper we are introducing the artificial intelligence based automatic plant leaf disease detection and classification for quick and easy detection of disease and then classifying it and performing required remedies to cure that disease. This approach of ours goals towards increasing the productivity of crops in agriculture. In this approach we have follow several steps i.e. image collection, image preprocessing, segmentation and classification.**

Keywords—Plant Leaf Diseases Detection, Classification, Image Preprocessing, Segmentation, K Means clustering

INTRODUCTION

Agriculture plays a very important role in the economic growth of any Country. It is the field which highly affect the GDP of the countries. Agriculture sector contributes around 16% of GDP of India. There are various factors that affects the quality and quantity of crops cultivated. Due to different weather and local conditions these plants are exposed to various diseases. And if these diseases remain undetected may cause some serious losses. In India itself around 15-25 percent of crops are lost due to diseases, pest, weeds. Also, we can take reference of the incident of Georgia (USA) in 2007 in which there was loss of around 540 USD due to plant diseases.

As the cultivational fields are quite large and have very large number of plants in that, hence it becomes very difficult for the human eye to properly detect and classify each and every plant. And doing so is very important as even single infected plant can spread the disease. Also, most of the farmers does not have proper knowledge of those diseases and actual cure for that disease. Hiring experts may cost them heavily and use of pesticides without knowledge will harm the land. Hence in order to solve this problem we have developed the Artificial Intelligence based solution.

Accuracy and speed are the two main factors that will decide the success of the automatic plant leaf disease detection and classification model. The suggested model will help the farmers to correctly detect and classify the disease by scanning the leaf and alert the farmers about the disease before it starts spreading. The model is mainly divided into four steps or phases. In first one, we collect the dataset of different plant leaves infected as well as healthy. These all images will be color images. In second step, noise from the images is removed then we will create color transformation structure for the images. In third step we segment the images using clustering techniques available. This step is performed to easily extract the foreground that is leaf. Now the image set of leaves with black background is obtained. In final step, different machine learning and deep learning algorithms like logistic regression, KNN, SVM and CNN are trained and compared on the basis of accuracy and the algorithm that performs best in training as well as testing is taken in account.

I. LITERATURE SURVEY

Paper by Saradhambal.G, Dhivya.R, Latha.S, R. Rajesh give solution to the plant disease with image classification. In their approach they collect 75 images of different diseased plant leaves such as Bacterial Blight and more. There were total of 5 classes that include 4 disease classes and one normal healthy leaf class. Removal of noise is done with some image preprocessing and then conversion into lab color model was done. They segmented the image with clustering and Otsu's method. After that some feature extraction is done on the basis of which class is determined. They have not discussed the accuracy that they have achieved as well as dataset was small [1].

Another paper named "Plant Leaf Disease Detection and Classification Based on CNN with LVQ Algorithm" clarifies that they have used CNN model for the leaf disease classification. In their methodology they have used a dataset of 500 images divided into 400 training and remaining 100 testing. Total classes for classification were 5 including one healthy class as well. Images size used was quite well that is 512*512. Three matrixes for R, G, B channels were used as input to CNN model and the output was feed into neural network known as LVQ (Learning Vector Quantization). Average accuracy of around 88 percent was achieved. Their proposed model was only for tomato related diseases [2].

"Plant Disease Classification Using Image Segmentation and SVM Techniques" by K. Elangovan, S. Nalini uses the svm for the classification purpose. In their methodology image was converted into another color space. After that image was cropped and with image preprocessing techniques noise was removed and smoothening was done and converted into

978-1-7281-2791-0/20/$31.00 ©2020 IEEE

greyscale images. Segmentation was also performed and then features were extracted. They considered color, morphology and texture as features and they were used for classification. They also does not mention about the accuracy of their suggested model [3].

II. Proposed Methodology

The model that is proposed by us to detect and classify the infected plant leaves consists of 4 phases.

Those phases are: -

- Dataset Collection
- Image Preprocessing
- Segmentation
- Selection of Classifier

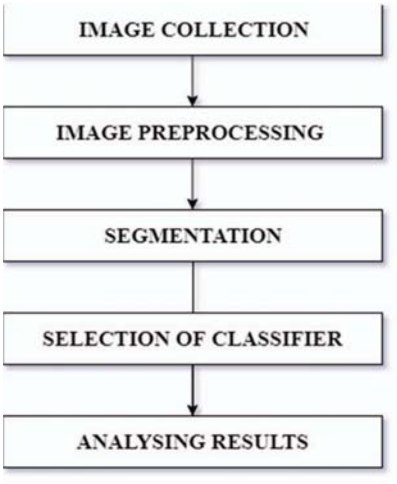
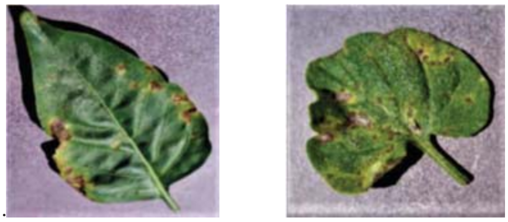


Fig. 1 Workflow Diagram of Proposed Methodology

Dataset Collection

Firstly, the images of leaves were collected from online sources such as GitHub, Kaggle and also some of the image's dataset consists of 20,000 images divided into 19 different classes. The dataset consists of both healthy and infected leaves which covers diseases like black rot, rust, bacterial spot, early blight, late blight, leaf scorch, target spot, mosaic virus of different crops like apple, potato, tomato, grape, strawberry, corn.

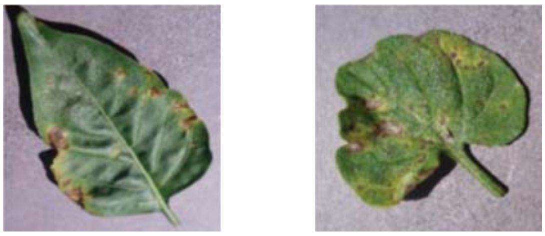


(a) (b)

Fig. 2 Sample Images from Dataset

Image Preprocessing

In this step images are resized to smaller pixel size in order to speed up the computations. The acquired images contain some noise. This noise is removed using some filtering techniques like Gaussian Blur. After that images are present in RGB format which is not appropriate for further work as RGB format is unable to separate image intensity. Hence it is converted to another color space that is HSV which separate color from intensity. Also, RGB color space is noisier than HSV.



(a) (b)

Fig.3 Images after Preprocessing

[8] RGB to HSV conversion: -

First R, G, B values are divided by max value that is 255

So, R' = R/255

G' = G/255

B' = B'/255

Then $C_{max} = \max(R', G', B')$

$C_{min} = \min(R', G', B')$

$\Delta = C_{max} - C_{min}$

Hue: -

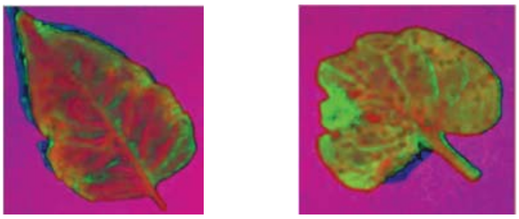
$$H = \begin{Bmatrix} 60^{\circ} * \left(\frac{G'-B'}{\Delta} mod6\right), Cmax = R' \\ 60^{\circ} * \left(\frac{B'-R'}{\Delta} + 2\right), Cmax = G' \\ 60^{\circ} * \left(\frac{R'-G'}{\Delta} + 4\right), Cmax = B' \end{Bmatrix}$$

Saturation: -

$$S = \begin{Bmatrix} 0, & Cmax = 0 \\ \frac{\Delta}{Cmax}, & Cmax \neq 0 \end{Bmatrix}$$

Value: -

$V = C_{max}$



(a) (b)

Fig.4 Images converted to HSV color space

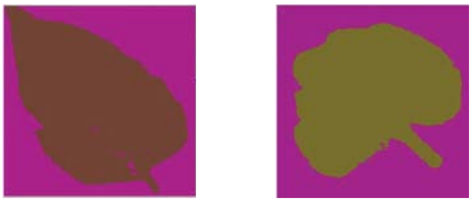
Segmentation

In this step, segmentation of images is done in order to separate the leaves from the background. Segmentation is performed using K-means clustering with 2 cluster centers, one for background and one for foreground. K-means clustering is unsupervised learning technique that is used to segregate the datapoints in the predefined number (k) of clusters or groups on the basis of their similarities.

K –Means algorithm works as follows: -

Set of inputs: - number of clusters(k), set of datapoints

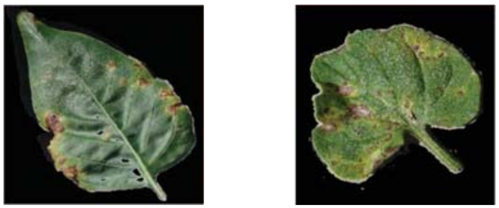
```
1. Put k centroids in random location in space.

2. Repeat the following steps until none of cluster location changes: -

    a. For every datapoint $x_i$ -
        i. Find nearest centroid $c_i$ by argmax $D(x_i, c_i)$ where $D = \sqrt{(\sum(x_i - y_i))}$
        ii. Assign $x_i$ to the cluster with nearest centroid
    b. For every cluster, new centroid is assigned by taking mean of all datapoints assigned to that cluster
```



(a) (b)

Fig. 5 Images after K-means clustering

After finding the two clusters, one with background and other one with leaf part, the clustered image is used to change the pixel value of the background of the leaf to black. By doing so the useless information from the image is eliminated which in turn increases accuracy.



(a) (b)

Fig.6 Images after removal of Background

Selection of Classifier

This is the classification problem as we have to classify the type of disease on the leaf of the plant. So, we have plenty of machine learning as well as deep learning algorithms that we can apply on this dataset.

We have decided to start with low complex algorithms and increasing the complexity level in order to increase accuracy of the model. We have selected four classifiers namely – logistic regression, KNN, SVM and CNN.

A. Logistic Regression

It is the simplest classification algorithm available but yet powerful enough to make some good results.

The logistic regression makes the use of logistic function that is sigmoid function to squeeze the output in range of 0 and 1. After training on training set, the model gives the accuracy of 66.4% on testing set which is not that bad considering complexity of algorithm and number of classes in dataset.

B. KNN (K Nearest Neighbors)

It is the algorithm can be used in both classification as well as regression problems. It is very simple and easy algorithm to implement. Here we plot all the datapoints in space and then find the k nearest neighbors of the datapoint that we want to classify by finding the distance between all other datapoints and the input datapoint. Then k datapoints are chosen which are nearest to that datapoint and their classes are taken then predicted class of input is the class with maximum occurrence. On our dataset the knn model was able to give accuracy of 54.5%.

C. SVM (Support Vector Machine)

SVM is another machine learning algorithm that we have used to classify the diseases. In this algorithm all points are mapped in space so that points of different class can be divided by gap. Gap should be as wide as possible so that boundary can separate them. This boundary is called decision boundary and the extreme data points of classes called support vectors. Kernel tricks are used for nonlinear dataset. The kernels that are available are linear, nonlinear, polynomial and RBF.

The svm in our case worked poorly as it gives accuracy of only around 53.4% after using linear kernel.

D. CNN (Convolutional Neural Network)

This is the far most complex deep learning model that we have used to classify the diseases. As it is very complex hence it requires good computational power as well. It is the most common neural network that is applied to image classification problems. CNN is a neural network which comprises of four layers namely-Convolutional layer, Pooling layer, Activation function layer and Fully connected layer as shown in figure [7]

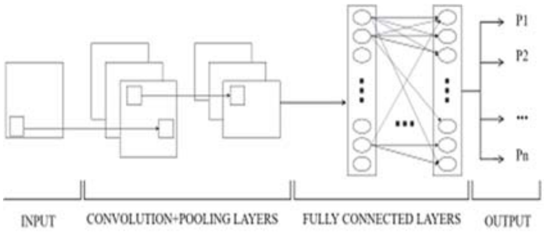


Fig.7 General CNN architecture

a) Convolutional Layer

It is most important layer in CNN model and also responsible for the naming of this network. In this layer, some mathematical operations are performed to get the features of the image. It consists of filters which have width and height less than input image and depth same as input image. If image of size 64*64*3 is feeding in the CNN and we have total of 10 filters then output of this layer will have dimension of 64*64*10. There are total of 5 convolutional layers where number of filters are 32,64,64,128,128 respectively. The kernel size is 3*3 for all layers.

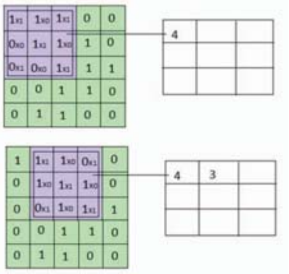


Fig.8 Convolution layer with 5*5 input images and 3*3 filters

b) Pooling Layer

This is the layer which is majorly responsible for size reduction of output of previous layer. Filters of different sizes can be used in this layer but generally 2*2 size is preferred. There are two major kind of pooling layers that are used namely- max pooling and average pooling. As name suggest max pooling take the maximum value out of filter and average pooling takes the average. In our model we have used max pooling with pool size of 3*3 and stride is default that is equal to pool size.

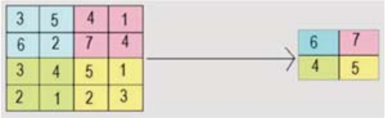


Fig. 9 Max Pooling with 2*2 filters and stride 2

c) Activation Layer

In any neural network activation layer plays an important role as it is responsible for nonlinear learning of the network. There are different types of activation functions such as sigmoid, tanh, ReLU, LeakyReLU.In our model we have used ReLU for all the layers except output layer for which we have used softmax.

d) Fully Connected Layer

After performing all the computations in previous layers, the output is feed into normal neural network for classification purpose. Model have 2 dense layers with 1024 and 19 units respectively.

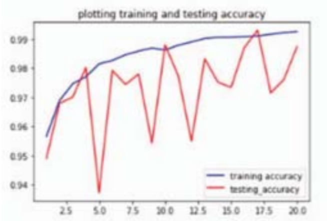


Fig. 10 Accuracy vs iterations plot of CNN

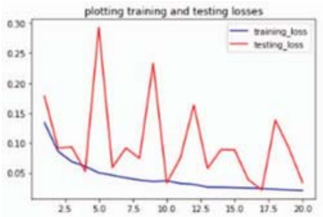


Fig.11 Loss vs iteration plot of CNN

The CNN model was able to give accuracy of 98% on testing set that is so far best among all classifiers with 0.033 loss.

III. Results

After using image preprocessing techniques along with k means clustering and comparing various classifiers available. The Logistic regression performs quite well considering number of classes and was able to give accuracy of 66.4%. KNN takes quite long time to make predictions due to large computations and only give accuracy of 54.5%. SVM also does not perform well and provide accuracy of 53.4%. CNN outperforms all other and give very good results as its accuracy score was 98%. We are able to achieve the accuracy of 98% in detecting and classifying the plant leaf disease. The accuracy score achieved by all classifiers are as follows: -

TABLE 1 – Accuracy Achieved of Classifiers

CLASSIFIER	ACCURACY (%)
Logistic Regression	66.4
KNN	54.5
SVM	53.4
CNN	98.0

IV. Conclusion

In this paper, a very accurate artificial intelligence solution for detecting and classifying different plant leaf disease is presented which makes use of convolutional neural network for classification purpose.

The presented model used the dataset that consists of more than 20,000 images with 19 total classes. The following model can be extended by using even more large dataset with more categories of diseases and the accuracy can also be improved by tuning the hyperparameters. The remedies for the classified disease can also be included in the model. The model then can be deployed on android and as well as iOS platform to reach out the farmers who can make the actual use of the proposed system.

References

[1] "Plant Disease Detection And Its Solution Using image Classification" by Saradhambal.G, Dhivya.R, Latha.S, R.Rajesh in International Journal of Pure and Applied Mathematics Vol. 119 ,no.14, pp. 879-884, 2018

[2] "Plant Leaf Disease Detection and Classification Based on CNN with LVQ Algorithm" by Melike Sardogan, Adem Tuncer, Yunus Ozen in 3rd International Conference on Computer Science and Engineering, 2018

[3] "Plant Disease Classification Using Image Segmentation and SVM Techniques" by K.Elangovan, S.Nalini in International Journal of Computational Intelligence Research ISSN 0973-1873 Vol.13 ,no.7, pp.-1821-1828, 2017

[4] Rajneet Kaur , Manjeet Kaur "A Brief Review on Plant Disease Detection using Image Processing" IJCSMC, Vol. 6, Issue 2, 2017

[5] SandeshRaut, AmitFulsunge "Plant Disease Detection in Image Processing Using MATLAB" IJIRSET Vol. 6, Issue 6 , 2017

[6] Sonal P Patel, Mr. Arun Kumar Dewangan "A Comparative Study on Various Plant Leaf Diseases Detection and Classification" (IJSRET), ISSN 2278-08882 Vol. 6 , Issue 3, March 2017

[7] "Plant Leaf Disease Detection and Classification Using Image Processing Techniques" by Prakash M. Mainkar, Shreekant Ghorpade, Mayur Adawadkar in IJIERE Vol. 2, Issue 4, ISSN-2394-5494 , 2015

[8] Conversion from RGB to HSV color space at - https://math.stackexchange.com/questions/556341/rgb-to-hsv-color-conversion-algorithm

[9] C.V. Giriraja, C. M. Siddharth, Ch. Saketa, M. Sai Kiran, "Plant health analyser", Advances in Computing Communications and Informatics (ICACCI) 2017 International Conference on, pp. 1821-1825, 2017

[10] Aparajita, Rudransh Sharma, Anushikha Singh, Malay Kishore Dutta, Kamil Riha, Petr Kriz, "Image processing based automated identification of late blight disease from leaf images of potato crops", Telecommunications and Signal Processing (TSP) 2017 40th International Conference on, pp. 758-762, 2017.

[11] N. Belsha, N. Hariprasad, "An approach for identification of infections in vegetables using image processing techniques", Innovations in Information Embedded and Communication Systems (ICIIECS) 2017 International Conference on, pp. 1-6, 2017.

[12] Trimi Neha Tete, Sushma Kamlu, "Detection of plant disease using threshold k-mean cluster and ann algorithm", Convergence in Technology (I2CT) 2017 2nd International Conference for, pp. 523-526, 2017.