Optimizing Random Forest to Detect Disease in Apple Leaf

Kamalalochana. S, Nirmala Guptha

Abstract*—Green Revolution was introduced in agriculture to meet the food scarcity. Despite the increase of agricultural production, farmers are challenged by infestations. Infestation reduced the crop yield. Traditional method involved manual inspection of plants to identify diseases. With advancement in technology, the infested plant leaves can be captured into images and subjected to processing by computing element. The computing system are being trained to process the image using Machine Learning algorithms to classify the images. Processing the image and detecting with improved accuracy is essential. Random Forest classifier is used to detect the disease in Apple Leaf. The accuracy of prediction by Random Forest can be influenced by configuring its parameters. This Paper talks about the various options that can be applied to optimize Random Forest classifier for improving the accuracy of detecting Apple Leaf disease.

Keywords*— Machine Learning Algorithm, Random Forest, Apple leaf disease detection.

I. INTRODUCTION

Green Revolution was introduced in agriculture to meet the food scarcity. The goal was to increase the agricultural production using modern techniques. Despite the increase of agricultural production, farmers are challenged by infestations. The degree of infestation could vary from mild to severe. Failure to address the infestation at the early stage could lead to wide spread of infestation to other plants in the farm. This could lead to irrecoverable loss to the farmer. Traditional method involved manual inspection of plants to identify diseases. This required expertise about disease symptoms.

The agriculture industry will need to cater to this increasing demand of food by increasing the yield while being challenged by crop infestation. This requires constant monitoring and detection of disease, classify the disease to allow farmer to treat the disease.

This is where technology advancement in capturing images with high resolution, processing the images and application of machine learning has a very key role to play in detecting the disease.

Using the publicly available dataset of healthy and diseased plants, machine learning classifiers can be used to detect diseases in plants with higher accuracy. Farmers can take necessary action to treat the diseased plants.

Popular machine learning classifiers include Support Vector Machines, K-Nearest Neighbors, Naïve Bayes and Random Forest.

This paper focusses on how the agriculture industry can incorporate tuning of Random Forest classifier to efficiently detect disease in Apple leaf.

A. *Crop Yield Loss*

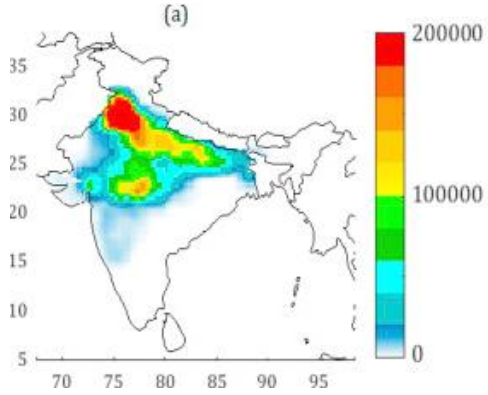

Figure 1(a) Total Wheat production (Rabi growing season) in tonnes/model grid;

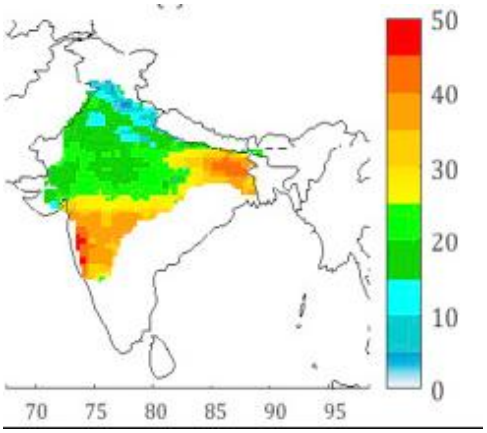

Figure 1(b). Percentage relative yield loss (%RYL);

Revised Manuscript Received on April 25, 2019
Kamalalochana. S, School of C&IT, REVA University, India.
Dr. Nirmala Guptha, School of C&IT, REVA University, India.

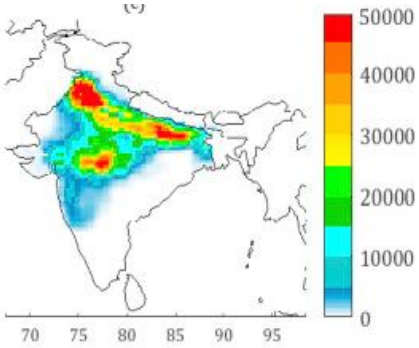

Figure 1(c)Crop production loss (CPL) for wheat in tonnes/model grid in each model grid
Source: ScienceDirect.com digital News Report – 2018

The above statistics describes the annual yield and production loss of Wheat crop. Over most of the high wheat producing regions RYL generally varies between 15 and 30%[1]

II. LITERATURE REVIEW

[1] Sukhvir Kaur, Shreelekha Pandey and Shivani Goel proposed a "Semi-automatic leaf disease detection andclassification system for soybean culture". In their work, they devised a rule based semi-automatic system using concepts of k-means and support vector machines to detect healthy leaves from diseased leaves. The best performing feature set for soya bean leaf disease detection.

[2] Ashwini Awate, DaminiDeshmankar, Gayatri Amrutkar, UtkarshaBagul and Prof. SamadhanSonavane proposed "Fruit Disease Detection using Color, Texture Analysis and ANN". In their work, they have used K-means clustering to perform image segmentation. The segmented images are labelled by maintaining a catalogue. Artificial Neural Network for pattern matching and disease classification. Diseases are categorized based on feature vectors, color, morphology, texture.

[3] Pooja V, Rahul Das, Kanchana V proposed "Identification of plant leaf diseases using image processing techniques". In their work, they propose K-means clustering for image segmentation by using Otsu's method for setting the threshold. Thesholding is used to create gray-scale image from color images. They use Support Vector machines to classify the disease. The disadvantage in this proposal is that user to manually select the region of interest.

[4] BoikoboTlhobogang and Muhammad Wannous, proposed "Design of Plant Disease Detection System: A Transfer Learning Approach Work in Progress". In their work, they propose image processing using convolutional neural network and use Transfer learning technique of machine learning for classification. They retrain the existing Googlenet Inception V3 model on a publicly available dataset. Tensorflow is used to retrain the Inception Model.

[5] Aakanksha Rastogi, Ritika Arora, and Shanu Sharma, proposed "Leaf Disease Detection and Grading using Computer Vision Technology & Fuzzy Logic". In their work, the image is processed to identify the plant using the features of the leaf using artificial neural network. The leaves are then subjected to disease classification using K-Means and artifical neural network.

[6] Rutu Gandhi, Shubham Nimbalkar, Nandita Yelamanchili and Surabhi Ponkshe , proposed "Plant Disease Detection Using CNNs and GANs as an Augmentative Approach". In their work, they use Generative Adversarial Networks (GANs) to augment the limited number of images for crating training dataset. They use Convolutional Neural Network (CNN) for classification. The use of GANs method helps to increase the variations of the available dataset.

[7] Bhavini J. Samajpati and Sheshang D. Degadwala, proposed "Hybrid Approach for Apple Fruit Diseases Detection and Classification Using Random Forest Classifier". In their work, they propose K-means clustering for image segmentation. The color and texture features are extracted from the fruit image. The color and texture features are fused together as input to the Random Forest Classifier.

III. MODERN TECHNOLOGY IN AGRICULTURE

With the advancement of technology, modern farming techniques and methods can be adopted to reduce plant infestation and reduce the crop loss. Some of the techniques that could make farming a smart system are:

- Internet of Things (IoT) – enabled by Hardware and Software
- Robotics
- Global Positioning System
- Data Analytics and Machine learning for prediction

A. *Machine Learning for prediction in agriculture*

Machine learning is an application of Artificial intelligence. It provides the system with the ability to learn from data and improve from experience without needing to modify the program. Image processing technique such as machine vision system has been proven to be an effective automated technique [2]. In Machine Learning, data is provided as input and statistical analysis is applied to predict the result. When the input data changes, the above process is repeated to update the learning algorithm. Machine Learning can be classified into

Supervised learning: Learning method in which we train the model using training (input) data that is labelled with classification of the data that it belongs to. Separate unlabeled test (input) data is used to evaluate the efficiency of the prediction. Supervised learning methods are further categorized into:

- Classification
- Regression

Unsupervised learning: Learning method that uses data that is not classified/labelled. No training is provided to the learning methods. The method has to identify hidden pattern/structures to learn.

Unsupervised learning is further classified into:

- Clustering
- Association

In this paper we focus on Random Forest – a supervised learning method.

Random Forest:

Random forest is a decision tree based supervised learning method. Unlike decision tree, Random Forest can be used for both Classification and Regression.

In decision tree, Entropy and Gini Index is used to identify the best attribute to split the tree. The leaf node in the tree consists of labelled data. The decision tree often leads to overfitting. Unlike decision trees, Random forests overcome the disadvantage of over fitting of their training data set and it handles both numeric and categorical data[3]

In Random Forest, of the M available features, N features are selected in random. from the available features/attributes. Random Forest creates more than one tree(N) by randomly selecting the features. The input data is applied to all N trees in the forest. Each tree predicts the outcome independent of the rest of the trees. The output of prediction of the Random Forest is the maximum vote received for each of the type predicted in the forest.

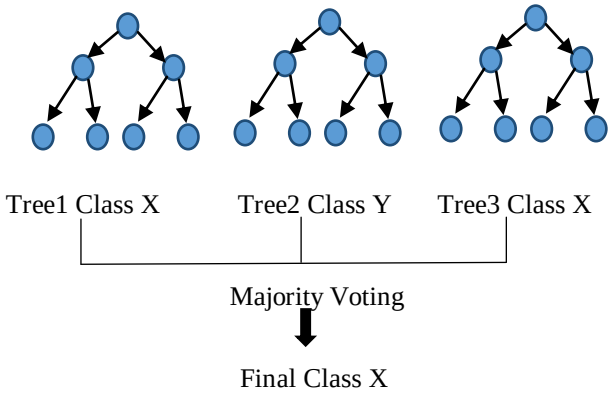

Figure 2. Random Forest Classifier

IV. HIGH LEVEL ARCHITECTURE FOR IMPLIMENTATION

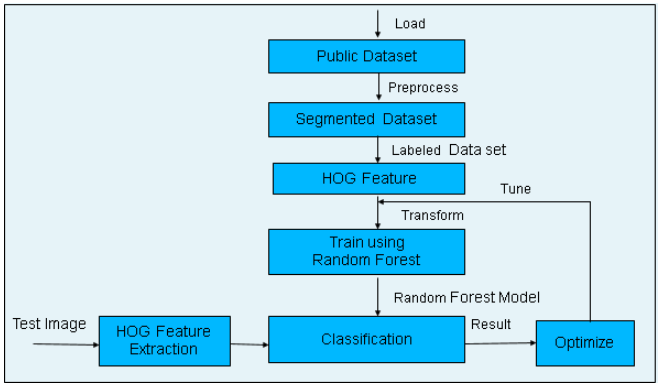

Figure 3. High Level Architecture

The classify if the leaf is infected or in good health manually requires inspection of features of leaf such as color, shape, texture. While digitally processing the image, the color, shape and texture need to be recognized. Images captured could be from different image capturing devices such as Cellular Phones, DSLR camera, etc. The devices may vary in capabilities of resolution and image size. While processing the image, images need to be transformed to a standard or predefined resolution.

Uniform Aspect Ratio – The input image from various device/s may vary is shape. It is required to bring all the input images to a standard shape. The ratio between width and height of the image is referred to as **Aspect Ratio**.

Image size of 256x256 is used for processing. All the input images shall be resized to 256x256.

Image Cropping – Input images might have irrelevant objects or elements captured. In case of plant leaf dataset, the image may contain more than one leaf in the image. Image can be cropped(trimmed) and resized to include only the sample leaf of interest. Cropping eliminates extra white spaces and completes background elimination process[4].

Image Augment – The shape and size of the leaf in the input image vary than the actual shape and size due to the angle in which the image was captured. It is practically not feasible to always capture the image at an angle of 90 degree of the leaf. The processing technique should be able to generate leaf images of varying size and shape by scaling the image. This allows to simulate the images captured from different angles which are essential for training the classifier.

Reducing the Dimensions – Color images are represented using RGB(Red Green Blue). The combination of these 3 channels of color results in 255 color combination. By varying the values of RGB, we can create various shades of a specific color. The simplest technique of dimensionality reduction is to use only 2 channel of colors Gray Scale where each pixel in the image is either black or white. The more complex approach would be to combine the shape, size and texture and transform it into 2-dimensions.

Segmenting Image Dataset: Segmentation is a critical task in image processing. The image is divided into smaller regions or cells. Each pixel in the image is assigned to one of the cells. Each cell may relate to a specific part of the object or to a different object in the image. Parts of the objects that belong to same cell having pixel values with smaller variance define connected components. The Neighboring pixels are assigned to different cells if the pixel values vary with a value larger than threshold are identified as disconnected components.Sometimes backgroundremoval techniques may also be needed in case of region ofinterest needs to be extracted [3].The segmentationcan be done using various methods like otsu' method, k-meansclustering, converting RGB image into HIS model[5]

The common approaches to segmentation are thresholding, edge based method and region based method.

Thresholding Method: In thresholding technique, each cell of the image represents range of values of pixels. In the below segmented image, all the pixels with values less than 127 are assigned to one cell and the rest in the other cell. This is called as Binary Thresholding..

Edge-based Method: Edge based method uses filter. Filter is applied. The output of the filter determines if the pixel is edge or non-edge.

Region-based Method: Neighboring pixels with similar values are grouped together into same region. Neighboring pixels with dissimilar values are form a separate region.
In this paper, Thresholding Method is used for image segmentation.

Histogram Based Thresholding

Color histogram gives the representation of the colors in the image[4]. The histogram of pixels h_1, h_2, h_3, .., h_N where h_k represents the pixel number of pixel with intensity k and N represents the maximum pixel value.

1. Initial value of threshold is chosen at random
2. With the chosen threshold value, mean value of the pixels from the two cells are computed
3. The value of threshold (T) is updated to the mid value of the two mean values computed in step 2.
4. Repeat steps 2 and 3 until the value of T computed in step 3 changes.

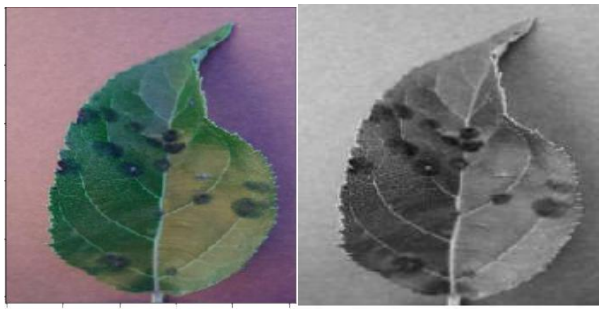

Figure 4(a) Figure4(b)

Figure 4(c) Figure 4(d)

Figure 9 (a) Color image; (b) Gray Scale Image; (c) Global thresholding (d) Adaptive Thresholding

V. PROPOSED METHODOLOGY

HOG Feature Extraction – In Histogram of Oriented Gradients, the shape and appearance of local objects within an image is described by the distribution of directions of an edge. The histogram of t directions is computed for each cell. The histograms are concatenated to define the direction. It is a common practice to normalize the histogram with intensity across larger area called blocks. The below steps are carried out in HOG Feature extraction.

- ***Gradient computation***
  Filters the colors or intensity with filter kernels [-1,0,1] and $[-1,0,1]^T$
- ***Orientation Binning***
  Based on the gradient value, weighted vote is casted by each cell to orientated histogram channel.
- ***Descriptor blocks and normalization***
  Image processing is challenged by varying illumination and contrast. The normalization of gradient strength is performed. Cells from larger spatially connected blocks are grouped to normalize the gradient strength.

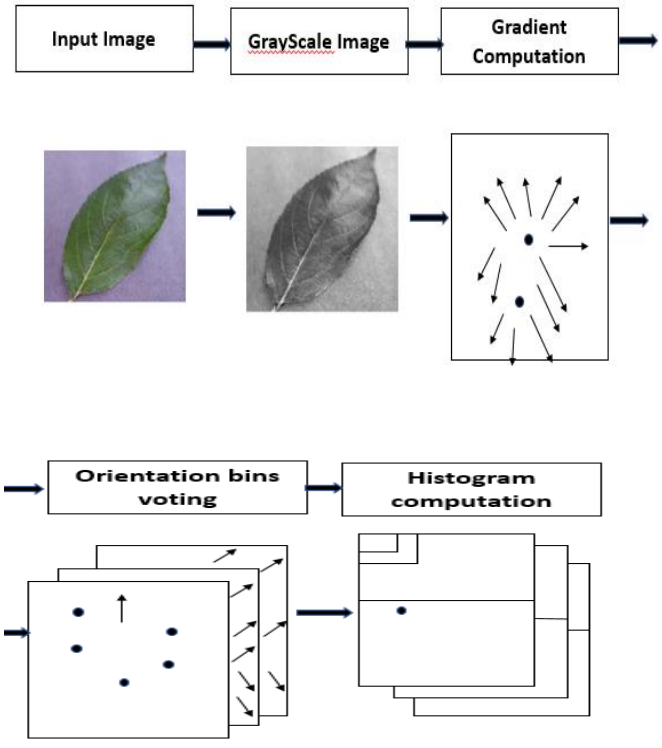

Figure 5 Image Segmentation

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

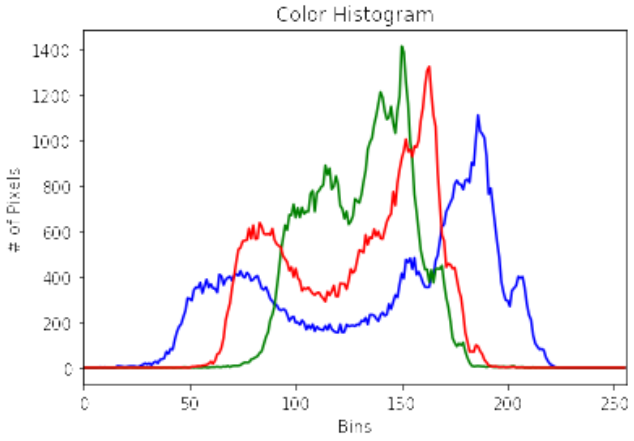

Figure 6. Color Histogram of Sample leaf image

Training Random Forest:

Training Random Forest classifier involves below procedure:

- Initialize the hyper parameter engine with different Hyper Parameter settings. Starts with default settings
- Extract features from the labelled images of Apple leaf.

1. For each hyper parameter settings
2. Split the extracted features of labelled into training and test set.
3. Input training set to the random forest along with the labels.
4. The classifier learns from training data with no hyper parameter tuning training model is generated.
5. Save the model.
6. Input the test data to classifier without labels.
7. Hyper parameter engine saves the settings if the higher accuracy achieved
8. Repeat the steps from 2 -7 for each hyper parameter settings
9. Identify the hyper parameter settings that resulted in highest accuracy.

B. Optimizing Random Forest for detecting disease in Apple leaf

Random forest employs multiple trees to predict. The maximum number of voting determines the final outcome. The outcome of Random forest is dependent on below parameters.

- Number of trees in the forest
- Size of training data
- Number of leaves in the tree
- Number of features used to determine the best split
- Random state to randomize the selection of features and samples

The above listed factors can be configured to alter the outcome of the Random Forest. These parameters are referred to as “Hyper Parameter”.

Random forest classifier chooses the number of trees and samples at random to predict. If the number of trees in the forest is too small when compared to the number of training samples, then the probability of a sample data being evaluated could be very low or never evaluated at all.

If the number of trees (estimators) is too large when compared to the training sample, some of the samples may never be evaluated at all.

Choosing a smaller value of random state reduces the variance in the forest and increases the bias of individual tree in the forest.

VI. RESULTS

The dataset from PlantVillage was used to train the Random Forest classifier. The preprocessed images are segregated into separate directories. The result of segmentation of images with different thresholding approaches are provided below.

The following parameters were used to tune the random forest classifier

- Estimators (number of trees)
- Leaf size
- Random State

Below are the experimental results obtained by running the Random Forest Classifier with a dataset of 1000 images.

Accuracy Results

- With the default settings, the accuracy is :**73.7%**
- Accuracy results with parameter tuning:

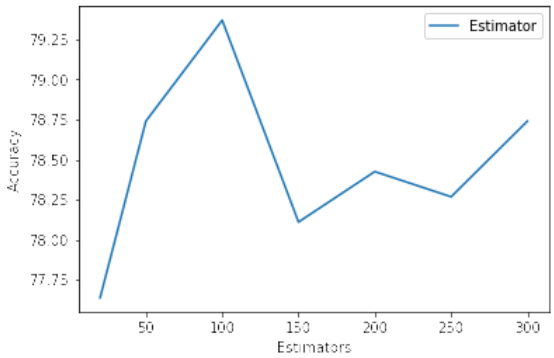

Figure 7(a) Tuning Estimators

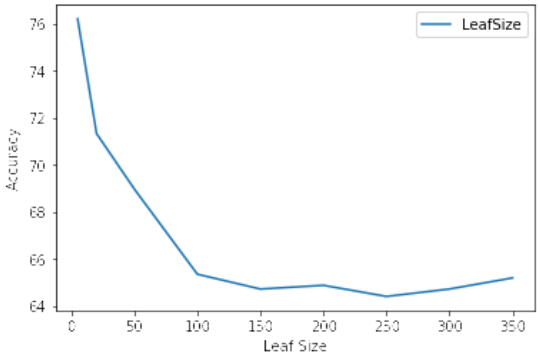

Figure 7(b) TuningLeafsize

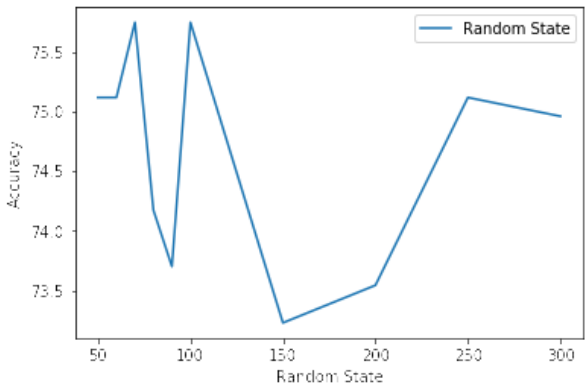

Figure 7(c) TuningRandom State

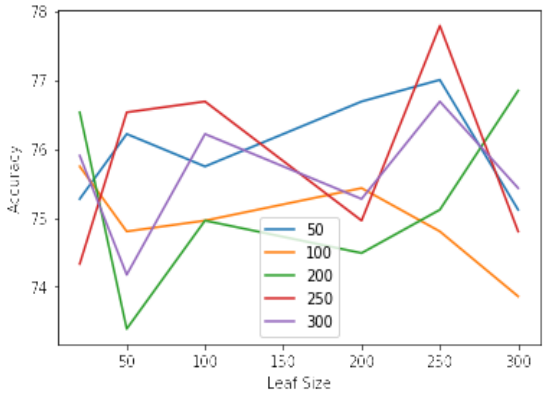

Figure 7(d) Tuning Estimators and Leaf size

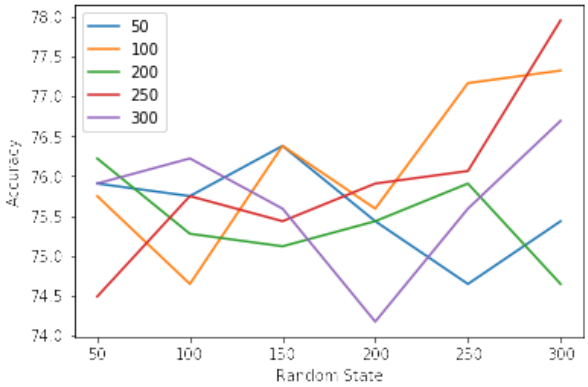

Figure 7(e) Tuning Estimators and Random State

VII. CONCLUSION AND FUTURE SCOPE

In this paper we perform image segmentation of Apple leaf using thresholding method. Histogram of Oriented Gradients is applied to extract the features of image. We evaluate different hyper parameters of Random Forest to determine the optimal hyper parameter to achieve the best accuracy. Below are the observations:

- The accuracy that can be achieved without tuning the hyper parameters is 73.7%.
- The maximum accuracy that can be achieved by tuning Estimators alone is 79.23%
- The maximum accuracy that can be achieved by tuning Random State alone is 75.74%
- The maximum accuracy that can be achieved by tuning Estimators in combination with Leaf Size is 77.00%
- The maximum accuracy that can be achieved by tuning Estimators in combination with Leaf Size is 77.32%

From the observations made, it can be concluded that tuning the Estimators hyper parameters to a value of 100 improves the accuracy by **7.5%.**The machine learning model is trained and tested using publicly available dataset. Each image has single leaf. The machine learning model can be trained with complex images with multiple leaves and improve the accuracy.

VIII. REFERENCES

[1] ScienceDirect Article - https://doi.org/10.1016/j.aeaoa.2019.100008
[2] AakankshaRastogi, Ritika Arora, and Shanu Sharma, proposed "Leaf Disease Detection and Grading using Computer Vision Technology & Fuzzy Logic"2nd International Conference on Signal Processing and Integrated Networks2015.
[3] Shima Ramesh, Mr. Ramachandra Hebbar,Niveditha M, Pooja R, Prasad Bhat N, Shashank Nand Mr. P V Vinod, "Plant Disease Detection Using Machine Learning" International Conference on Design Innovations for 3Cs Compute Communicate Control, 2018.
[4] Sukhvir Kaur, Shreelekha Pandey and Shivani Goel, "Semi-automatic leaf disease detection and classification system for soybean culture". IET Image Process., 2018, Vol. 12 Iss. 6, pp. 1038-1048
[5] Jitesh P. Shah, Harshadkumar B. Prajapati and Vipul K. Dabhi"A Survey on Detection and Classification of Rice
Plant Diseases" 2016.
[6]Sachin D. Khirade and A. B. Patil, "Plant Disease Detection Using Image Processing"International Conference on Computing Communication Control and Automation , 2015.