CNN-based Plant Disease Identification in Crops from Multilabel Images using Contextual Regularization

R. Srinivasan
Department of Computing Technologies, SRM Institute of Science and Technology, Potheri, Kattankulathur - 603203, Chengalpattu District, Tamil Nadu, India.
srinirvs89@gmail.com

C. Santhanakrishnan
Department of Computing Technologies, SRM Institute of Science and Technology, Potheri, Kattankulathur - 603203, Chengalpattu District, Tamil Nadu, India.
santhanc@srmist.edu.in

S. Iniyan
Department of Computing Technologies, SRM Institute of Science and Technology, Potheri, Kattankulathur - 603203, Chengalpattu District, Tamil Nadu, India.
iniyans@srmist.edu.in

R. Subash
Department of Computing Technologies, SRM Institute of Science and Technology, Potheri, Kattankulathur - 603203, Chengalpattu District, Tamil Nadu, India.
subashr@srmist.edu.in

Pradeep Sudhakaran
Department of Computing Technologies, SRM Institute of Science and Technology, Potheri, Kattankulathur - 603203, Chengalpattu District, Tamil Nadu, India.
pradeeps1@srmist.edu.in

Abstract

Image processing is a technique of enhancing or extracting several information from an image using various algorithms like contrast enhancement, resizing, color conversion, etc. Due to the high performance in recent years, Deep Learning (DL) algorithms have grown more popular for image categorization. The purpose of this research work to identify the healthy leaf and unhealthy leaf from the Multilabel image dataset. The next step is to identify the type of disease affect in the unhealthy leaf. The dataset used in the proposed work comprises of 13 crops leaf images that includes apple, blueberry, cherry, tomato, grapes, orange, peach, pepper, potato, raspberry, soybean, squash, strawberry, and corn. This research work classifies the plant leaf into health or unhealthy using Convolutional Neural Network (CNN) technique. CNN proves that 98.75% of accuracy obtained to the plant village dataset.

Keywords: *Convolutional Neural Network; Deep Learning; Customer Churn; Entity Extraction.*

I. Introduction

Accurate detection of leaf disease in various crops is essential to the agricultural economy [1]. Early disease detection in plants are extremely beneficial to farmers. Some diseases are visible to the naked eye, allowing farmers or plant pathologists to diagnose them early on. Plant pathologist are very limited in developing countries and availability is inconsistent. Expertise in the farmers for detecting plant leaf is time consuming process and lead to errors. Hence automatic detection of healthy or unhealthy leaf helps farmers to identify and diagnose the disease in the earlier stage. The rise in the world's population has resulted in an increase in food demand. With this increase in demand, the agricultural industry must develop a method for identifying and mitigating plant disease in order to increase plant yield and reduce pesticide use.

Disease affected to the leaves can be various reason such as, chemical composition of fertilizer, climatic conditions like temperature, humidity, precipitation and so on. Infections of plant leaves, such as fungi, bacteria, and viruses, have a negative impact on plant health and crop production. The symptoms present in the plant leaf are not visible to the naked eye, and some diseases manifest themselves in the visible spectrum. The advancement of computer vision has accelerated the development of automated systems for detecting plant diseases that are not visible to the naked eye. The quality of the agricultural product can be estimated automatically using images and different Deep Learning (DL) based approaches. Images from different parts of the plants are helpful to develop detection system. Leaves are the most important part to identify the disease detection in any types of plants. These images are helpful to classify or identify the diseases in the leaf image using Machine Learning (ML) or DL approaches.

There are many approaches have been investigated to detect the healthy leaf from the dataset using ML approaches. The challenging task in the ML approaches are, extracting proper feature, computation cost, and time. By considering all the facts the proposed work is to use DL techniques to classify the leaf as healthy or unhealthy. This paper discussed about the healthy leaf detection from the plant village dataset. This paper addresses the CNN technique to identify the leaf is healthy or not. The CNN technique makes a combination of different layers present in the existing method. The combination of different layers helps to extract features and observes patterns from the identified features. The major contributions of this paper are,

1. Applying contrast enhancement step to increase the brightness of the image. Contextual regularization helps to identify the important features present in the haze or noisy images.

2. Applying CNN technique to classify the leaves into healthy or unhealthy. The next step is to identify the disease type in the affected leaf.

II. LITERATURE SURVEY

Image Processing can be applied in various applications such as plant leaf classification, intelligent traffic monitoring [2], pattern recognition, remote sensing etc. This section briefly discussed image processing techniques to analyze plant leaf diseases from images. The author proposed a method to detect plant disease using multiclass Support Vector Machine (SVM) combined with “Active Contour Edge Detection (ACED)” [3]. The author used gaussian filters to reduce noise, contrast enhancement to convert input image to L*a*b color space. ACED and sobel edge detection method were used for edge detection technique. After applying the above mentioned methods, multiclass SVM is used for classification method.

Aditya Sinha et al., proposed a system for detecting olive spot disease in the plant leaf [4]. The authors were proposed a method to classify the similar diseases in the dataset. The proposed method test the images that has no specific geometric distribution. The image dataset had eleven images of the peacock spot diseased leaves and twelve images of the neofabrea leaf spot disease. Gray Level Co-occurrence Matrix (GLCM) was used for feature extraction and found that energy and entropy has high co-relation value with respect to infection percentage area.

Liba Manopriya J et al., analysed various machine learning techniques K-means clustering algorithm, Back Propagation Network, and Naïve Bayes (NB) [5]. The author used different plant leaf diseases to test this approach. The author used different segmentation techniques, feature extraction techniques and classification algorithm. The proposed method achieves 93% accuracy for different leaf spot diseases in various crops. G. Sambasivam et al., worked with imbalanced dataset to detect and classify diseases in Cassava plant [6]. The disease categories included Cassava brown streak virus, green mite, mosaic disease and bacterial blight. CNN was used for low cost detection which involved 3 layers. The proposed method involved 35% of time in pre-processing and loading images. 10% time in defining model, 50% in training and 5% in calculating results.

Shima Ramesh et al., present a model on papaya plant and the dataset included 120 leaf images [7]. The feature extraction task was achieved with help of Histogram of oriented gradients (HoG) technique which includes Hu moment, haralick texture and color histogram. Logistic regression, Random forest, SVM, KNN, Naïve bayes and CART were compared. Random forest achieved highest accuracy of 70.14%. A system for identifying disease in capsicum crop with a dataset consisting of 70 images which included diseases- anthracnose, bacterial spot, cercospora leaf spot, Gray leaf spot and powdery mildew [8]. The model took images of leaf as well as fruit as input for classifying the disease. K-means clustering was used for segmentation task. Kamil Dimililer et al., presented a system on maize plant detection using back propagation neural networks. Sigmoid activation function was used for activation function [9]. The output neuron classify plant using binary codes 0 and 1.

D. A. Godse et al., proposed a model for identifying the disease in Jute plant [10]. The work involved stem analysis and used color co-occurrence matrix algorithm for feature extraction. Hue based image segmentation was used. Classification was done with help of SVM technique. Mohit Agarwal et al., proposed work discussed CNN model for classifying diseases in tomato plant [11]. The input images were augmented and includes 13 convolution layers, 3 dense layers to achieve an accuracy of 91.2%. The accuracy varied from 76% to 100% with respect to classes. Santosh Adhikari et al., presented a way to identify tomato plant disease with CNN based classification [12]. The model identified late blight, Gray spot and bacterial canker. The proposed work classified into 4 classes and used stochastic gradient descent as an optimization algorithm.

Nilay Ganatra and their team used Roberts, Prewitt and sobel filters and Otsu's technique for segmentation [13]. The paper work on 14956 images. There were a category of features compared like color moment features which include mean, standard deviation, skewness and kurtosis, texture features including contrast, correlation, homogeneity, energy, entropy and variance. Shape features like eccentricity, width ratio, area, equivdiameter, roundness, convex area. Zernike moment features included AOH and Phi OH. The proposed paper compared four ML method for classification which include random forest, SVM, ANN and KNN. Gabor wavelet transform feature and Zernike moment feature for feature extraction. Random forest gave an accuracy for 73.38% SVM gave an accuracy of 67.27%, ANN gave an accuracy of 65.68% and KNN produced accuracy of 63.20%. As a result, existing approaches are used different types of filter and various ML approaches to extract the feature and trying to remove the noisy image. The proposed work helps to identify the type of disease present in the noisy data.

III. PROPOSED METHODOLOGY

A. Dataset Description

The dataset was taken from the kaggle named as “Plant Village Dataset” [14]. These image dataset was collected by Penn State University’s research and development team that empowers smallholder farmers with technology in their fight against plant disease. Table 1 depicts the number of images used for training and testing process.

TABLE 1. DATASET DESCRIPTION

Disease Types	Crop Name	Training	Test
Black_rot	Apple	444	125
Apple_scab	Apple	512	126
Cedar_apple_rust	Apple	211	55
Healthy	Apple	1208	329
Healthy	Blueberry	1212	300
Healthy	Cherry	648	170
Powdery_mildew	Cherry	824	210
Gray_leaf_spot	Corn	401	103
Healthy	Corn	992	233
Common_rust	Corn	935	239
Northern_Leaf_Blight	Corn	744	197
Healthy	Grape	393	84
Black_rot	Grape	911	236
Esca	Grape	1017	276
Leaf_blight	Grape	816	215
Haunglongbing_(Citrus_greening)	Orange	4504	1102
Healthy	Peach	277	72
Bacterial_spot	Peach	1783	459
Bell_healthy	Pepper	1381	295

Bell_Bacterial_spot	Pepper	779	200
Healthy	Potato	112	31
Early_blight	Potato	811	200
Late_blight	Potato	204	102
Healthy	Raspberry	183	77
Healthy	Soybean	1301	256
Powdery_mildew	Squash	445	291
Healthy	Strawberry	558	290
Leaf_scorch	Strawberry	231	167
Early_blight	Tomato	782	265
Mosaic_virus	Tomato	1452	280
Late_blight	Tomato	298	111
Bacterial_spot	Tomato	324	180
Septoria_leaf_spot	Tomato	780	278
Yellow_Leaf_Curl_Virus	Tomato	765	292
Spider_mites Two-spotted_spider_mite	Tomato	872	170
Leaf_Mold	Tomato	1240	285
Target_Spot	Tomato	404	160
Healthy	Tomato	654	210
No images	None	541	154

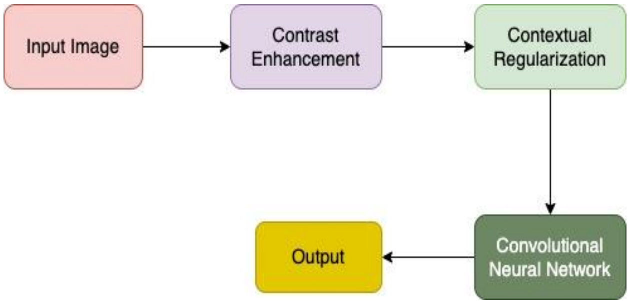
B. Architecture of Proposed Methodology

Figure 1 describes the architecture used in the proposed work. The architecture consist of data preprocessing techniques and apply the CNN model to identify the disease in the given dataset. The combination of convolutional layer, activation layer, batch normalization and max-pooling layers are repeated applied as five iterations to extract the features from the plant village dataset. Later, the features are made as flatten, fully connected dense layer, dropout layer, and then applied the activation function to predict the output.

C. Input Image

Input image may consist of various color spaces such as Grayscale, HSV (Hue, Saturation, Value), RGB (Red, Green, Blue), CMYK (cyan, magenta, yellow, and key (black)), etc. This dataset consists of images that are represented in RGB color spaces.

Fig. 1. Architecture of Proposed work



D. Contrast Enhancement and Contextual regularization

Image pre-processing is an essential step in any ML or DL methods. Image Pre-processing helps to remove the noisy images, resizing the images and removal of redundant images in the dataset. The reason to include the enhancement technique in this dataset to increase the contrast of background noise images, resizing the images (256 * 256), and also normalize the individual pixel of the image. This plant leaf dataset consist of background noise in the form of dust, dew drops and reflection of light from leaf surface. Resizing and contrast enhancement methods helps to increase the contrast and make the size as uniform in all images. The state of the art approaches are always eliminate the noisy images from the dataset. This research work primarily focused to increase the contrast on the targets and the background in the noisy images. The contrast enhancement helps to increase the contrast and helps to avoid the elimination of noisy images.

Contextual regularization helps to remove the hazes present in the image [15]. The reason to occur hazes are image capture in bad weather condition that suffers visibility. The bad weather condition might be occurred due to foggy, reflection of light, image degraded due to atmospheric particles. Mostly those images are considered as noisy image and removed from the dataset. The proposed work helps to identify those images and apply the contrast enhancement and contextual regularization method to extract the features and able to detect the disease affected in the plant leaf. This results helps to avoid the removal of noisy leaf present in the dataset. On the other side, it helps to remove the noise present in the dataset. As a result, noisy data are cleaned and considered to be the part of the dataset.

E. Convolution Layer

The goal of the convolution operation is to extract high-level features from the input image. Convolutional networks should not be restricted to a single convolutional layer. Convolutional layers' primary function is to capture simpler features such as horizontal edges, color, vertical edges, gradient orientation, and so on. The advancement of the layers aids in the capture of complex features that provide human-like understanding of the images. The convolution layer is made up of two operations.

1) matrix representation

2) A kernel (K) or a filter (F)

The proposed method consists of an image with 256 * 256 * 3 pixels. The proposed work employs the CNN model's default filter. CNN has numerous filters that are included in the Python package. The proposed work employs 32 kernels. Edge detection, identity, sharpen, gaussian blur, and other operations are available in the kernel. The output of the convolutional layer represent as “feature map”. Feature map represents the numerical value of an input image represented in the form of matrix. Padding and strides are used in the filter operation. The proposed work is to apply the strides with value of 1. The images are present in the corner of the area, padding might be useful.

F. Activation Layer

The proposed work uses two activation function named as Rectified Linear unit (ReLu) and softmax. The activation function helps to find the neuron to be activated or not based on the value of weights and bias. The objective of using activation function is to introduce the non-linearity to the output. The activation function updates the value based on back-propagation method. ReLu activation is used to make the negative value as zero, whereas softmax is used in the output layer to identify the multiclass classification method.

G. Batch Normalization

Batch normalization layer helps to normalize the output of the previous layer value. Batch normalization layer is a normalization technique that can be included more frequently in the DL model. The proposed method includes the batch normalization layer at every iteration that's helps to avoid the overfitting.

H. Max Pooling

Pooling helps to reduce the dimensionality but retain the important features. The proposed method includes the max pooling at every iteration of the CNN layer. The combination of convolutional layer and max pooling provides the complete one cycle of CNN model.

I. Dropout layer

Dropout layer can be included after the flattened layer. Mostly, Dropout layers are included in the hidden layers to deactivate the features. Dropout layers should not be included after the convolutional layer. Dropout layer works randomly to deactivate the neurons during forward or back-propagation concept. The major aim of dropout layer is to reduce the interdependency among the neuron in the hidden layer. Dropout layer is also a normalization technique that helps to reduce the overfitting.

J. Flatten Layer

Flattening layer coverts the feature map into one dimensional vector. The one dimensional vectors are considered as the input layer and can include many more hidden layers.

K. Fully Connected Dense Layer

The fully connected dense layer acts as a hidden layer in the DL network. The proposed method methods includes two hidden layer to identify the disease affected in the plant leaf image dataset.

IV. EXPERIMENTAL RESULTS

Experimental results have to be evaluated by using the standard metrics [14].The leaves present in the dataset are divided into 39 classes and 13 crop species. The dataset consist 54,309 images that are used for training and testing data. Out of 39 classes, 26 types of disease are discussed in the dataset. Some disease are common in the different crop species. The targeted crop species are, tomato, apple, soybean, cherry, corn, squash, peach, potato, raspberry, blueberry, grape, strawberry, and orange.

Fig. 2. Epoch vs Training Data

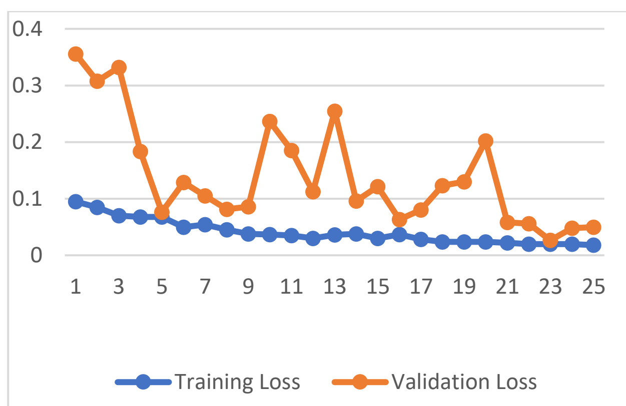


Fig. 2 depicts the graph between epochs versus training loss. X-axis represent epoch and y-axis represents training loss. The training loss in the each epoch is constantly decreasing till 21st and then become constant. Based on the training loss, the number epoch's value can be determined. The reason for evaluating the training loss is to verify whether the model is underfitting or not. The error loss is high in training data, the model is underfitting. Fig. 2 helps to identify the model is underfitting or not in the earlier stage. This clearly explains the current enhanced CNN model is not underfitting. The number of images used in the training phase are discussed in the Table 1.

Fig. 3. Epoch vs Test Data

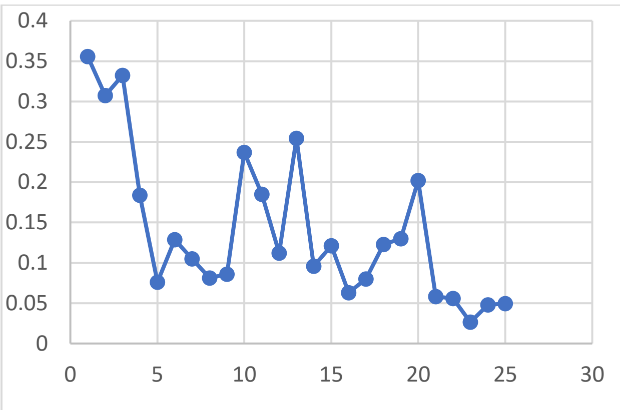


Fig. 3 depicts the graph between epochs versus test loss. This graph helps to identify how the new data is fit in the current model. The model is not fit in the validation data, then the model suffers overfitting. The researchers have proved that, the model works completely fine in training data, but does not outperform in the test data. The current enhanced CNN model clearly depicts that the loss value is getting decreased on different epochs. The test data value shows the abnormality in fig 3 due imbalanced data that occurs in testing. The maximum loss at epoch 1 is 35% and minimum loss is 2% at epoch 23. Figure 3 depicts the graph between epochs versus test loss. The value in the test data also decreases on different epochs.

Fig 4. Comparison among Training Loss and validation Loss

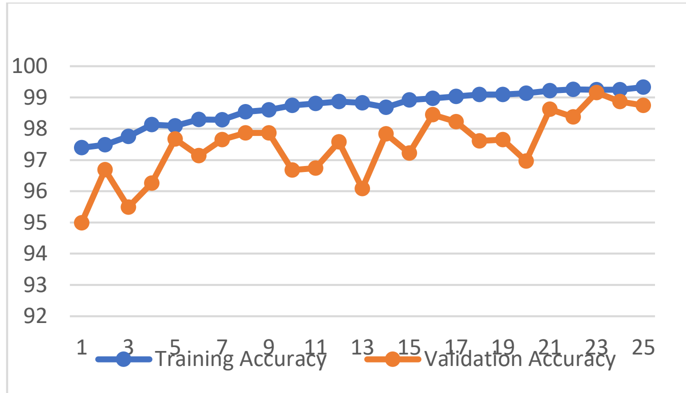


Fig 4. depicts the comparison between the loss value between the training data and validation data. Fig 2 and Fig 3 represents the individual loss compared on different epochs. Fig 4 compared the result of both the loss value. The validation loss is higher than training loss that concludes the model is underfitting. The validation data is continuously decreasing and some point starting increasing again. The, sudden increase in the loss value at some point leads to overfitting. The model should balance the individual loss value and comparison has to be happen in both validation and testing loss. Fig 4 clearly depicts that the model is neither overfitting nor underfitting. This clearly depicts the data is properly fit for train data as well as test data and leads the model is good fit. At epoch 1, validation loss is very high and later on, it constantly decreasing. At epoch 23 both the validation loss and testing loss is almost equal. The model is stopped at the epoch 25.

Fig. 5. Accuracy of the Test data

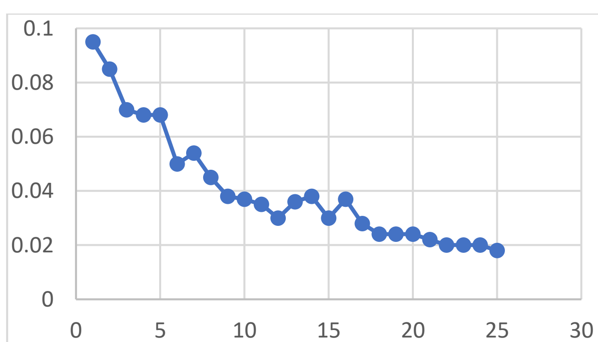


Fig. 5 depicts the graph between epochs versus accuracy of the test data. Validation data helps to find the suitable model among several models. This research work is not compared with several models, but it verify the CNN works well for this dataset. This research work applies the CNN model to the test data and identify the performance of the model. The accuracy of the test data gets increased on different epochs.

The minimum accuracy at epoch 1 is 94.99 and the best epoch accuracy is 98.75% achieved from the CNN model. Though the accuracy is good, some of the class label not perform well due to imbalancing problem.

Fig. 6. Comparison of accuracy score in training and testing

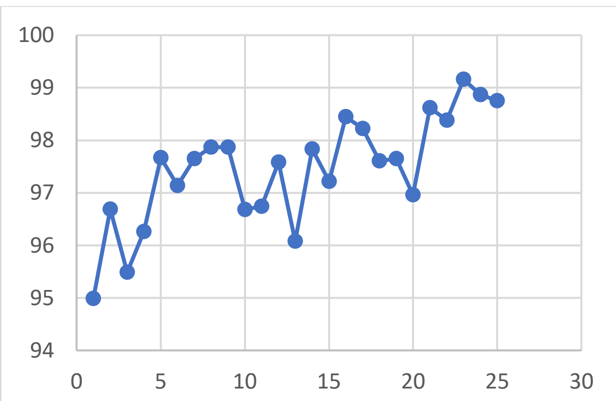


Fig 6. Clearly depicts the accuracy is compared between the training and testing data. Though the accuracy is good, the dataset suffers the imbalance problem. This paper helps to perform the contrast enhancement and contextual regularization done on the dataset. Contextual regularization helps to achieve the good accuracy in the noisy image dataset.

V. CONCLUSION

The task of plant leaf classification and identifying type of disease from the dataset is quite challenging. This paper addresses the disease identification present in the plant leaf image dataset. This paper applied an effective CNN technique to identify the 39 types of disease present in the 13 crop species. The proposed work introduced a CNN technique to extract the features from the plant village image dataset and produces an accuracy of 98.75% at the epoch 25. The limitations of plant leaf images are not tested under the conditions of lighting, illuminations, different pose variations, etc. In future, different parameters are introduced to reduce the complexity of CNN model.

References

1) Vizcarra Aguilar, Gerson & Bermejo, Danitza & Mauricio, Antoni & Zárate Gómez, Ricardo & Dianderas, Erwin. (2021). The Peruvian Amazon forestry dataset: A leaf image classification corpus. Ecological Informatics. 62. 101268. 10.1016/j.ecoinf.2021.101268.

2) Gulati, I., & Srinivasan, R. (2019). Image processing in intelligent traffic management. Int. J. Recent Technol. Eng, 8(2), 213-218.

3) Varalakshmi, P. & Aravindkumar, S. (2019). Plant disorder precognition by image based pattern recognition. Procedia Computer Science. 165. 502-510. 10.1016/j.procs.2020.01.018.

4) Sinha, Aditya & Shekhawat, Rajveer. (2020). Olive Spot Disease Detection and Classification using Analysis of Leaf Image Textures. Procedia Computer Science. 167. 2328-2336. 10.1016/j.procs.2020.03.285.

5) Liba Manopriya J, Arockia Jansi Rani and M.Asha Paul ,Machine Intelligence for the detection of plant diseases using Image processing, EasyChair Preprint, October 11, 2019.

6) G. Sambasivam, Geoffrey Duncan Opiyo, A

predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural network, Egyptian Informatics Journal,2020.

7) Shima Ramesh, Mr. Ramachandra Hebbar, Niveditha M, Pooja R, Prasad Bhat N, Shashank N, Mr. P V Vinod, Plant Disease detection using machine learning, 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control.

8) Mr.V Suresh, D Gopinath, M Hemavarthini, K Jayanthan, Mohana Krishnan, Plant disease detection using image processing, International Journal of Engineering Research & Technology (IJERT) Vol. 9 Issue 03, March-2020

9) Kamil Dimililer , Ehsan Kiani, Application of back propagation neural networks on maize plant detection, 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017, 24-25 August 2017, Budapest, Hungary

10) Prof. Dr. D. A. Godse, Nalini Tripathi, Reetika Wali, Richa Singh, Sunidhi Swaraj , Detecting Jute plant disease using Image processing and machine learning, International Journal of Current Engineering and Scientific Research (ijcesr), 2394-0697, Volume 5, Issue 5, 2018

11) Mohit Agarwal, Abhishek Singh, Siddhartha Arjaria , Amit Sinha, Suneet Gupta, ToLeD: Tomato Leaf Disease detection using convolution neural network, International Conference on Computational Intelligence and Data Science (ICCIDS 2019).

12) Santosh Adhikari, Bikesh Shrestha, Bibek Baiju, Er. Saban Kumar K.C, Tomato plant Diseases detection system using image processing, 1st KEC Conference Proceedings| Volume I, ISBN 978-9937-0-4872-9 September 27, 2018.

13) Nilay Ganatra and Atul Patel, A multiclass plant leaf disease Detection using image processing and machine learning techniques, International Journal on Emerging Technologies 11(2): 1082-1086(2020).

14) https://www.kaggle.com/emmarex/plantdisease/metadata

15) G. Meng, Y. Wang, J. Duan, S. Xiang and C. Pan, "Efficient Image Dehazing with Boundary Constraint and Contextual Regularization," 2013 IEEE International Conference on Computer Vision, 2013, pp. 617-624, doi: 10.1109/ICCV.2013.82.