xIlkogretim Online - Elementary Education Online,2021; Vol 20 (Issue 1): pp.3067-3076
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.01.344

Automated Disease Detection and Classification from Plant Leaf Images

S.Iniyan, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India.Email: iniyans@srmist.edu.in
R.Jebakumar, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India.Email: jebakumr@srmist

Abstract- India's economy is vastly based on agriculture. Increase in the productivity of agricultural fields affects the GDP of the nation. Early and automated crop disease detection is an important step to increase the productivity of these farms. This paper proposes a model using machine learning classifiers for the detection and classification plant disease and is based on tomato plant but similar results can also be obtained using other crops. Experimental results also reveal that 76 % and 72% of accuracy level obtained using Extremely Randomized Tree classifiers and Random forest outperformed other machine learning classifiers. Though our model not produced higher accuracy when compared to other pre-trained Convolutional Neural Network (CNN) models such as Res net and Image Net, the difference in time complexity made our model less memory extensive.

Keywords: Agriculture, Disease Identification, Machine Learning, Image Processing, Random Forest, K-Nearest Neighbours.

I. INTRODUCTION

Agricultural Technologies has evolved magnificently in the past few decades. In recent years, early detection of diseases in Plant still remains a challenging task in Machine Learning. Tomato is among one of the world's largest cultivated vegetable category representing almost 16%. Indian farmer cultivates about 11% of total worldwide production of tomatoes which make India 3rd largest Tomato Producer in the World [1]. Due to the vast Plantation area, it is not feasible to inspect all the Plants individually by some experts, and Machine Learning plays a vital role here in the early detection of diseases. Diseases and viruses in Tomato plants are a major threat to the income of farmers as well as impose a health risk to the consumers also.

The most common types of disease that can be seen in Tomato Plants are:-

1. Bacterial Spot
2. Early Blight
3. Late Blight
4. Leaf Mould
5. Septoria Leaf Spot
6. Two Spotted Spider Mites
7. Target Spot
8. Mosaic Virus
9. Yellow Leaf Curl Virus

A. Bacterial Spot

Bacterial spot is a devastating disease in tomato plants as it is very difficult to control and also it spreads rapidly. The four major bacteria responsible for this disease are X. euvesicatoria, X. perforans, X. gardneri and vesicatoria (All four are the species of Xanthomonas). The symptoms of Bacterial spot can be a circular brown spot on the leaves

B. Early Blight

Early Blight is a common fungal disease. Early blight can cause complete defoliation in tomato plants especially in the regions with high humidity, heavy rainfall and high temperature (24°–29°C) [2]. It is caused by Alternaria Solani fungi. The symptoms of early blight is the formation of small dark spots on the foliage that is near the ground.

C. Late Blight

Late Blight appears in the later stages of the growing season. Phytophthora infesters is the pathogen that causes late blight in tomato plants. The initial symptoms of Late Blight disease are irregular and light brown lesions covered with white mycelial growth on leaf [3].

D. Leaf Mold

Leaf Mold is also caused due to pathogens of the Fungi Kingdom i.e. Passalora Fulva which infect the oldest leaf of Tomato Plant first. Except for Tomato, it is not pathogenic to other plants. The Initial symptoms of Leaf Mold are Pale greenish-yellow spots on the top of the leaf and leaf begins to wither.

E. Septoria Leaf spot

Septoria leaf spot is also caused by fungus can be seen in tomato plant between early to mid-season, commonly in humid weather. Septoria lycopersici fungus is responsible for Septoria leaf spot in the tomato plant. The initial symptoms of Septoria leaf spot are irregular spots (usually small) with a grey centre on the lower leaves of the plant which gradually advance upwards.

F. Two Spotted Spider Mites

Two-spotted spider mites (TSSM) or 'Tetranychus Urticae Koch' are the most common spider mites species. Due to TSSM attack, the surface of the leaves becomes speckled and dull in appearance, later the leaves turn yellow and fall eventually.

G. Target Spot

Target Spot is also a foliar fungal disease found in Tomato Plant. Target Spot is caused by the fungus named Corynespora Cassiicola. The early symptoms of target Spot are almost similar to the early symptoms of early blight and bacterial Spot diseases that's why for the identification of target spot disease lab tissue tests are recommended.

H. Mosaic Virus

Mosaic virus is a common and dangerous type of plant virus which has no cure. Since this ToMV (Tomato Mosaic virus) can survive for more than 50 years, any healthy plant can get infected by ToMV from the debris of infected plant or . Yellowing and stunting of Tomato plant are one of the effects of ToMV (Tomato Mosaic virus) which eventually reduces the yield.

I. Tomato Yellow Leaf Curl Virus

TYLCV or Tomato Yellow Curl virus is most damaging DNA virus responsible for the most destructive disease in the tomato plant. TYLCV is transmitted by 'Whitefly Bemisia Tabaci' which is also known as 'Silver leaf whitefly'. Crinkling of Tomato's leaf, pale, and stunting of the plant are the symptoms of Tomato Yellow Leaf Curl Virus

In the Proposed model, we have trained different Machine Learning Classifiers such as K-Nearest neighbour, Decision tree, Random Forest Classifier and Extra Tree classifier with various features that are extracted from the Images of Tomato's leaf. In the Proposed work, we are classifying nine different types of disease that are found in Tomato Plants.

We have organized this paper as follows, In section 1 we have discussed about the diseases which are most common in Tomato, In section 2 recent work related to this Proposed model is discussed, In section 3 we have discussed our Proposed Model, In section 4 we have explained the results of the proposed model, In section 5, we have concluded our Proposed work and discussed about scalability of project and future work related to it.

II. RELATED WORKS

In [4] Qing Yao et.al, proposed a model using image processing techniques and SVM with a radial basis kernel for the detection of rice diseases. They have taken three diseases into account namely rice bacterial leaf blight, rice sheath blight and rice blast. They have extracted a number of shape and colour features from the images in 4 orientation angles (0º, 45º, 90º, 135º). They trained 3 models using subsets of the features and got an accuracy score of 97.2%, 88% and 82% respectively.

R. Pydapati et.al in [5] investigated the use of computer vision in citrus plant disease detection. They used four different types of citrus diseases for the task. They used colour-co-occurrence method to extract colour and texture features to get some unique features that best describe the image. They used SAS discriminant analysis to reduce the number of attributes and got an accuracy of 81% using a traditional statistical classifier.

A. Meunkaew jinda et.al in [6] has proposed a model for plant disease detection in grapes. They have use self-organizing maps to segment the images and then used SVM to classify the disease. They have used images belonging to 3 classes including a healthy class. They have achieved an average accuracy of 86.03%

J.K.Patil et al. in [7] have proposed a Content Based Image Retrieval (CBIR) system to retrieve diseases leaf images of soyabean plant. They have used color, texture and shape features. Color features are extracted using HSV color histogram. They have used 5 different diseases of soybean plant and achieved an overall accuracy of 80% for the top 5 retrieval and 72% accuracy for top10 retrieval.

Shanwen Zhang et.al in [8] has proposed a model for leaf disease detection in cucumber using sparse representation classification. They have tried to improve upon the existing models that use color, texture feature in which all the features are given equal importance. After using K-means to segment the leaf image, features are extracted using lesion information after which SR space is used to effectively reduce the computation cost and achieve an accuracy of 85.7%.

In [10], K-Means has used for clustering the affected regions of apple from Gray level co-occurrence matrix (GLCM). In [11], used neutrosophic set has been used to identify the affected and un affected leaves through the true and false sets, the intermediate set has been further processed through SVM to obtain the précised output. In [12], detection and classification of disease from rice crop using Scale Invariant Fourier Transform has been utilized to extract the corner as well as interesting points from the input.

III. PROPOSED MODEL

After going through the research papers in our survey paper, we propose the following model for the identification of the disease from images of the leaf. The whole process can be broken into four parts or sub-models named below.

A. Pre-processing

Pre-processing is the major step before starting the training of the classifier. The images taken from various devices have some dependency on the configuration of that device or on the environment in which the images are taken which can affect the accuracy of any machine learning algorithm so, it is very important to make the images device and environment independent. This Proposed model includes the conversion of Raw RGB images (Input of our projects) to LAB or CIELAB (Lightness, Green-Red, and Blue-Yellow) color space. Suppose the pictures with uniform background is taken under non-uniform lighting (different lighting) conditions then there will be an increase or decrease in R, G, and B channels (In case of RGB color space) but in case of LAB or CIELAB color space, only the Lightness (L) channel will be affected. After converting the image to LAB color space, the Fourier Transform is used to partially remove the slight 'L' or lightness gradients with band masking in the frequency space.

B. Noise Reduction and Segmentation

Noises can be defined as the unwanted information present in leaf images which can produce strange results in classification. Also, many undesirable effects are produced by these noises such as blurred

objects, disturbs background scenes, corners and unrealistic edges [9] which can affect the accuracy of Machine Learning Classifiers. The 'L', 'a' and 'b' channels of CIELAB colour space is blurred slightly to decrease the amount of noise.

Segmentation is defined as the process of partitioning of images into different regions and removing the region which contains the background or shadow by keeping the only that region of the image which contains leaf. This region is our ROI (region of interest) [a] or the useful part of the image.

In Proposed model, Masking is used for the image segmentation. A mask is an image which contains only two types of pixels i.e black and white. While creating the mask, two different masks are created.

1. Color mask
2. Shadow mask

Color mask selects only those pixels whose color is close to the color of the background of the image while shadow mask selects those pixels which are not too dark and also its color is not too far from the color of the background (i.e shadow).

Shadow and color mask are combined in such a way that the shadow area gets removed from the color mask. After that if any isolate white or black pixels are present in the mask then it will be removed. Now the mask is ready and the corrected raw image is masked to remove the background and shadow from the image. As show in Figure 1 raw and segmented image

Figure 1. Raw and segmented image

C. Feature Extraction

A feature is a parameter on which the classification algorithm classifies the class label. Feature Extraction is the process of extracting various features from the Images. Selection of features for the analytical procedure determines the precision of Machine learning algorithm. In this proposed methodology, various features are extracted from the Leaf Image and fed into the machine learning algorithm for training and testing purpose.

Total of 10 features have been extracted from each leaf image of Tomato Plant such as:-

1) Mean
2) Standard Deviation
3) Contrast
4) Correlation
5) Inverse Difference Moment
6) Entropy
7) Energy
8) Yellow Ratio
9) Skewness
10) Kurtosis

Mean

Mean is the average intensity of the image pixels.

$$\bar{x} = \sum_{B} \alpha \quad (1)$$

Where, $\sum a$= the sum of a, B= Number of data

Standard Deviation

Standard Deviation is the measure of variability. The value of standard deviation provide the extent of variability exists from the mean value

$$\text{SD}=\sqrt{\frac{\sum|a-\bar{a}|^2}{b}} \quad (2)$$

Where $\sum$= summation of, a =each value in image dataset, $\bar{a}$=mean values in the image dataset, b= Total number of values in the image dataset.

Contrast

Contrast is the measure of the pixel's intensity i.e difference between luminance reflected from two adjacent surfaces. Contrast is also the difference in visual properties (like the difference in colour or brightness) that makes the two objects distinguishable or object distinguishable from background.

$$\text{Contrast}=\sum_{a,b} C_{a,b} \cdot (a-b)^2 \quad (3)$$

Where a,b are adjacent image pixels, $(a-b)^2$ is the weighing squared term, $C_{a,b}$ is the largest weighing factor

Correlation

Correlation is defined as the measure of dependency (Linear) of gray levels of neighboring pixels. Correlation can be measured by the given formula

$$\sum_{a,b} \frac{(k-\mu)(l-\mu)s(a,b)}{\sigma_a \sigma_b} \quad (4)$$

Where a, b are gray pair values, $\sigma_a \sigma_b$ is standard deviation, $s(a,b)$ is occurrence probability of gray values pairs a,b , $(k-\mu)(l-\mu)$ is the difference in gray-levels between two distinct pixels.

Inverse difference moment

Inverse difference moment can be defined as the measure of local homogeneity. Inverse difference moment (IDM) can be calculated by the given formula

$$\text{IDM}=\sum_{a=1}^{B_g} \sum_{b=1}^{B_g} \frac{1}{1+(a-b)^2} g_{ab} \quad (5)$$

Where $(a-b)^2$ is the difference in gray-levels between two distinct pixels, $g_{a,b}$ is occurrence probability of gray values pairs a,b

Entropy

Entropy can be defined as the measure of the degree of the Image's randomness. The texture of Image can be characterized using Entropy. Entropy is calculated by formula -sum(p.*log2(p))

$$\text{E}=-\sum_{a=1}^{B_g} \sum_{b=1}^{B_g} g_{ab,d} \log_2(g_{ab,d}) \quad (6)$$

Where $g_{ab,d}$ is occurrence probability of gray values pairs a,b separated by a distance vector d

Energy

Energy can be defined as the measure of textural uniformity. Mathematically Energy is the sum of squared elements in the GLCM or Gray-level co-occurrence matrix

$$\sum_{a=0}^{G-1} \sum_{b=0}^{G-1} [q(a,b,d)]^2 \quad (7)$$

Where $q(a, b, d)$ occurrence probability of gray values pairs a,b separated by a distance vector d

Yellow Ratio

Yellow Ratio is can be calculated by dividing the number of yellow pixels in the image by the difference between the total pixels and background black pixels.

$$\text{YellowRatio}=\frac{Number\ of\ Yellow\ Pixels}{Totalpixels-Number\ of\ background\ pixels} \quad (8)$$

Skewness

Skewness is the measure of Symmetry in the distribution. Skewness is calculated by dividing the 3rd central moment by the cube of standard deviation. For a symmetric distribution, the value of Skewness is zero.

$$\text{Skewness}=\frac{\frac{1}{B}\sum_{i=1}^{B}(b_i-b')^3}{(\frac{1}{B}\sum_{i=1}^{B}(b_i-b')^2))^{\frac{3}{2}}} \quad (9)$$

Where b_i is mean of ith image , b' is mode , $\sum_{i=1}^{B}(b_i-b')^2)$ is Standard deviation.

Kurtosis

Kurtosis is the measurement of how pointy the distribution is from a normal distribution. The Positive value of kurtosis means the distribution in Histogram is very pointy and Negative value of Kurtosis means the distribution is very flat. Kurtosis is calculated by dividing 4th central moment by the fourth power of standard deviation.

$$\text{Kurtosis}=\frac{\sum_{i=1}^{B}\frac{(A-\bar{A})}{X}}{C^4} \quad (10)$$

Where, $\bar{A}$ is the mean, C is the standard deviation, X is the sample size

D. Classification

Classification was one of the most important aspects of our research. With so many models available out there for classification, model selection was a huge issue. We tried using various classification algorithms available in sci-kit learn library to train and test our model. Among the various models, ensemble models such as random forest, extra trees classifier gave some of the best results. Some other classification algorithms we used were Naïve Bayes and KNN. The data was split in 80:20 ratio for training and testing the models. A pandas data-frame was used to store the csv file created from the image dataset using the feature extraction process after segmentation and pre-processing.

E. K-Nearest Neighbours

Figure 2 shows K-Nearest Neighbours is a basic supervised classifier algorithm. The underlying principle to this classifier is that the class of a given data point is that of the most common class among the nearest K data points. K is a positive integer value which specifies the number of nearest neighbours the classifier checks to find the most common class. If the value if K is one, the classifier just predicts the class of the nearest neighbours. The nearest neighbours can be found using the Euclidean distance. Sometimes K-D tree is used to reduce the number of required distance calculations and hence reducing the complexity.

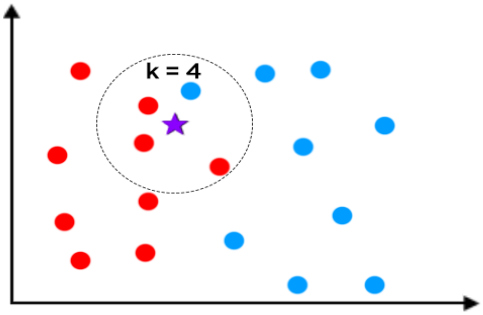

Figure 2. K- Nearest Neighbours

F. Decision Tree

Decision Tree is a tree like structure used for classification problems. Each of the node in the tree represents a feature and the branches emerging from it represent the possible values for the given feature. The leaf nodes in the tree represent the classes. They are simple and generally use a feature selection attribute such as information gain to select the feature as splitting criteria. This is a white box model and can be easily understood. In spite of all the advantages, decision trees are highly susceptible to over fitting when the trees grows large. So, random forest are used to overcome this problem.

G. Random Forest

Figure 3 shows random forest is an ensemble module that can be used for classification as well as regression tasks. It comprises of vast number of decision trees which in our case is used to classify a data point into one of the classes. The trees in the forests are sensitive only to some selected feature dimensions. It solves one of the most important problem in decision tree, i.e. over fitting to the training set. A random forest can gain accuracy by growing without over fitting to the training set. The goal of random forest is to boost the performance of the model by averaging various decision trees using majority vote.

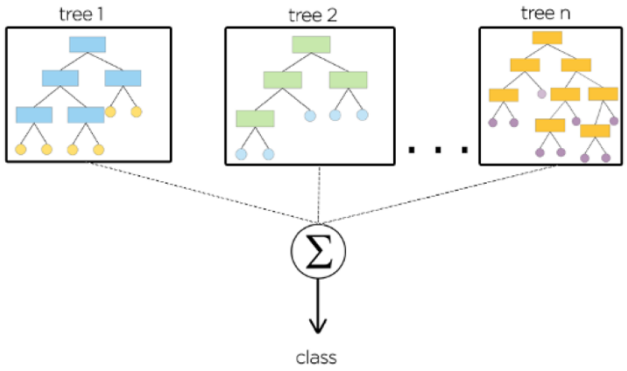

Figure 3. Random Forest Tree

H. Extra Trees Classifier

Extremely Randomized Trees classifier [1] is a slightly different version of random forest. In Extra Trees Classifier, every tree uses the whole dataset for training of the tree instead of selected feature dimensions. Also, at each split in the tree, the features used are selected randomly after checking all the splits possible and the one giving the highest accuracy is chosen to be the split node. Default number of parameters used at each split is $\sqrt{n}$ where n is the number of features.

IV. RESULTS

we trained the classifiers on data extracted from tomato plant leaves. More than 18000 images were used for the training and testing of the classification models. The train-test data split was 80:20. So around 14500 images were used to train the classifiers. The dataset contained images of 9 disease affected plants and healthy leaf images, so a total of 10 classes.

Few of the best accuracy score obtained by the some models are given the Table 1 Extra Trees classifier and random forest clearly has a greater accuracy than the other models. Though the accuracy of 76% is not the best but we think it can be improved considerably using better quality images and adding a few more features.

Table 1. Accuracy Scores

Classifiers	*Accuracy score*
Extra Trees Classifier	0.760187225
Random Forest Classifier	0.729900881
K Nearest neighbour(k=10)	0.609030837
Decision Tree	0.597742291

The following Table 2 gives the importance of each feature in the Extra Trees Classifier

Table 2. Feature scores of Extra Trees

Feature	Importance in Extra Trees
Mean_Red	0.06564679367817937
Mean_Green	0.057490833287763436
Mean_Blue	0.07994734881347378
Standard_Deviation_Red	0.07362962910145658
Standard_Deviation_Green	0.07478981764459179
Standard_Deviation_Blue	0.08658056992912504
Contrast	0.0854336779814096
Correlation	0.08158429400853892
InverseDifference Moment	0.051880538157214634
Entropy	0.057865540828931555
Energy	0.0563032311783971
Yellow Ratio	0.06120045088040656
Skewness_X	0.034971742181265954
Skewness_Y	0.04879301129548205
Kurtosis_X	0.04534294234062141
Kurtosis_Y	0.03853957869314217

Here, we can see that standard deviation of blue pixels is the most important feature in the ExtraTrees classifier followed by contrast and correlation. This may be due to prevalent yellow colour in the diseased plants whereas mean of green had a low importance because of the color being present in all the leaves irrespective of the class. By removing half of the features having the lowest feature importance gives us an accuracy of 71.4%. That's a very less difference in the accuracy given that half of the features were dropped.

The following diagram shows the change in the accuracy score with the value of k while using KNN classifier.

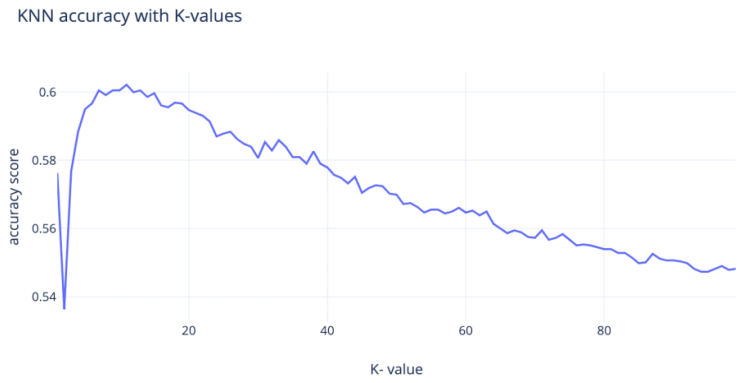

Figure 4. KNN Accuracy with K-values

As shown in the figure 4, the accuracy is the maximum at the value if k = 10, after which it steadily decreases.

V CONCLUSION

Among the various classifiers used in our proposed method, random forest and extremely random forest have the highest accuracy score. Even though they have a slightly lower accuracy of about 76% compared to other ANN models such as Resnet and other CNN architectures, the difference in time complexity makes our model far less memory extensive. Even comparing it to Multiple SVM which has a worst time complexity of $O(n^3)$ and best time complexity $O(n^2)$, random forest have a worst time complexity of O(n_estimate*mtry*nlog(n)) where mtry is the number of variables sampled at each node and n_estimate is the number of trees used which are constant and the best time complexity is O(log(n)). This makes our project easily scalable to any other crop once a dataset is procured. Future work on this project might include.

1. Increasing the number of features
2. Using higher quality images for better feature extraction
3. Adjusting the hyper parameters
4. Predicting the various stages of the disease
5. Pesticide recommendation system

REFERENCES

1. Tm, Prajwala, Alla Pranathi, Kandiraju SaiAshritha, Nagaratna B. Chittaragi, and Shashidhar G. Koolagudi. "Tomato Leaf Disease Detection Using Convolutional Neural Networks." In 2018 Eleventh International Conference on Contemporary Computing (IC3), pp. 1-5. IEEE, 2018
2. Chaerani, Reni, and Roeland E. Voorrips. "Tomato early blight (Alternaria solani): the pathogen, genetics, and breeding for resistance." Journal of general plant pathology 72, no. 6 (2006): 335-347.
3. Tripathi, A. N., K. K. Pandey, B. R. Meena, A. B. Rai, and B. Singh. "An emerging threat of Phytophthora infestans causing late blight of tomato in Uttar Pradesh, India." *New Disease Reports* 35 (2017): 14-14
4. Yao, Qing, Zexin Guan, Yingfeng Zhou, Jian Tang, Yang Hu, and Baojun Yang. "Application of support vector machine for detecting rice diseases using shape and color texture features." In 2009 international conference on engineering computation, pp. 79-83. IEEE, 2009.
5. Pydipati, R., T. F. Burks, and W. S. Lee. "Identification of citrus disease using color texture features and discriminant analysis." Computers and electronics in agriculture 52, no. 1-2 (2006): 49-59.
6. Meunkaewjinda, A., P. Kumsawat, K. Attakitmongcol, and A. Srikaew. "Grape leaf disease detection from color imagery using hybrid intelligent system." In *2008 5th international conference on electrical engineering/electronics, computer, telecommunications and information technology*, vol. 1, pp. 513-516. IEEE, 2008.
7. Patil, J.K. and Kumar, R., 2017. Analysis of content based image retrieval for plant leaf diseases using color, shape and texture features. Engineering in agriculture, environment and food, 10(2), pp.69-78.

8. Zhang, Shanwen, Xiaowei Wu, Zhuhong You, and Liqing Zhang. "Leaf image based cucumber disease recognition using sparse representation classification." Computers and electronics in agriculture 134 (2017): 135-141.
9. Boyat, Ajay Kumar, and Brijendra Kumar Joshi. "A review paper: Noise models in digital image processing." arXiv preprint arXiv:1505.03489 (2015).
10. Agarwal, A., Sarkar, A., Dubey, A.K.: "Computer vision-based fruit disease detection and classification": Smart Innovations in Communication and Computational Sciences. Springer (2019) 105-115
11. Dhingra, G., Kumar, V., Joshi, H.D :" A novel computer vision based neutrosophic approach for leaf disease identification and classification". Measurement.135 (2019)782-794
12. Bashir, K., Rehman, M., Bari, M.: "Detection and classication of rice diseases: An automated approach using textural features". Mehran University Research Journal of Engineering and Technology 38(1) (2019) 239-250.