Classification of leaf disease based on Multiclass SVM classifier

Pooja Kulinavar	**Vidya I. Hadimani**
Department of Computer Science & Engg, KLE DR M S	*Department of Computer Science & Engg, KLE DR M S*
Sheshgiri College of Engg. & Tech, Belagavi	*Sheshgiri College of Engg. & Tech, Belagavi*
poojakulinavar17@gmail.com	*vidya4279@gmail.com*

Abstract: India, the country where the main source of income is from agriculture. Farmers grow a variety of crops based on their requirement. Since the plants suffer from the disease, the production of crop decreases due to infections caused by several types of diseases on its leaf, fruit, and stem. Leaf diseases are mainly caused by bacteria, fungi, virus etc. Diseases are often difficult to control. Diagnosis of the disease should be done accurately and proper actions should be taken at the appropriate time. Image Processing is the trending technique in detection and classification of plant leaf disease. This work describes how to automatically detect leaf diseases. The given system will provide a fast, spontaneous, precise and very economical method in detecting and classifying leaf diseases. This paper is envisioned to assist in the detecting and classifying leaf diseases using Multiclass SVM classification technique. First, the affected region is discovered using segmentation by K-means clustering, then features (color and texture) are extracted. Lastly, classification technique is applied in detecting the type of leaf disease. The proposed system effectively detects and also classify the disease with an accuracy of 92%.

Keywords: Image Processing, Leaf diseases detection, K-means clustering, feature extraction, Multiclass SVM Classification.

I. INTRODUCTION

India is the land of agriculture. Farmers have an option to select required crops and then find appropriate pesticides for the plant to decrease the disease and increase the production. The cultivated plants will not always be healthy. In-order to increase the production with good quality the plant need to be monitored frequently because the plant disease leads to a reduction of the product. For successful cultivation, one should monitor the health as well as the disease of the plant. Diseases in plant cause heavy loss of the product. Hence the disease needs to be identified at the early stages, recommending farmers to avoid the harm in the production of the crop to increase the yield.

Plants suffer from diseases like Alternaria alternate (fungal), Anthracnose, Bacterial Blight (bacteria), and Cercospora Leaf Spot. Plant disease will be basically identified by observing different patterns on the parts of the plant like leaf, fruit, and stem. The indications on the leaf are taken into consideration for detecting the disease.

The disease can be categorized as bacterial, viral, fungal etc. The proposed work emphases on identifying and categorizing the disease like Alternaria Alternata, Anthracnose, Bacterial Blight, and Cercospora Leaf Spot which are basically found on pomegranate, rice, soya bean, carrot, rose, watermelon, mango etc., using the technique called as image processing. It automatically detects leaf diseases. This system will provide a fast, spontaneous, precise and very economical method in detecting and classifying leaf diseases.

II. SYSTEM DESIGN

The steps involved are
1) Training – In training all the collected images are trained to the model and all six features are extracted and stored in the database.
2) Classification – After training, the SVM will classify the given new input as which type of disease is affected.

The system design mainly consists of
1) Image collection
2) Image Preprocessing
3) Image segmentation
4) Feature extraction
5) Training

6) Classification using Multiclass SVM

Image Collection

The sample images of the diseased leaves are collected and are used in training the system. To train and to test the system, diseased leaf images and fewer healthy images are taken. The images will be stored in some standard format. In this study, the available images from the internet are also taken. The leaf images that are infected by Alternaria Alternata, Anthracnose, Bacterial Blight, Cercospora Leaf Spot and Healthy leaf are also included
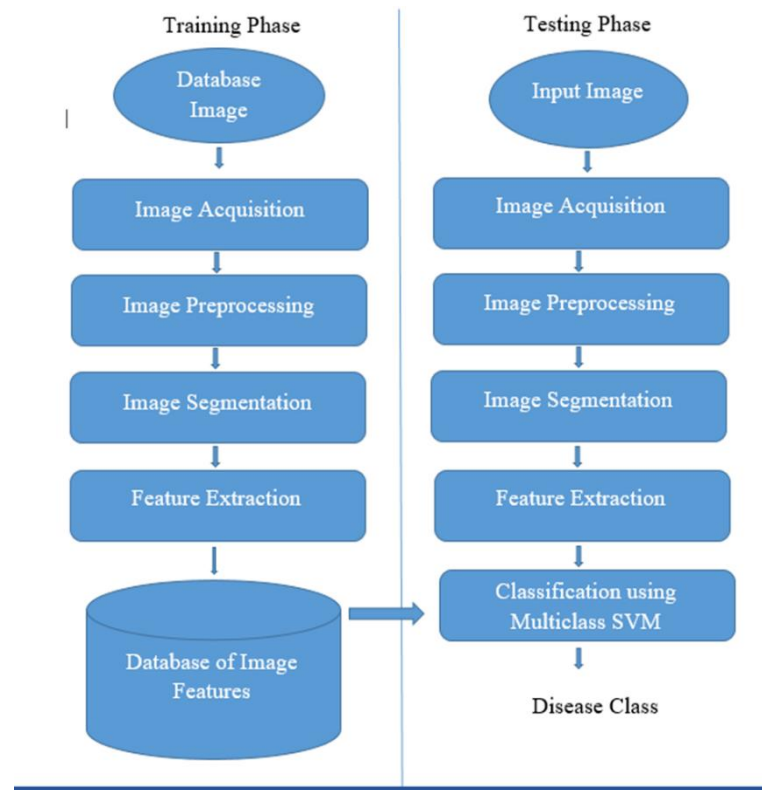

Fig 2.1 System Design

Pre-processing

Image pre-processing is significant for genuine data that are frequently noisy and uneven. During this phase, the transformation is applied to convert the image into another image to improve the quality that better suits for analyzing. This step represents a crucial phase in image processing applications because the effectiveness of subsequent tasks (e.g., features extraction, segmentation) depends highly on images quality. Also, it significantly improves the effectiveness of data mining techniques.
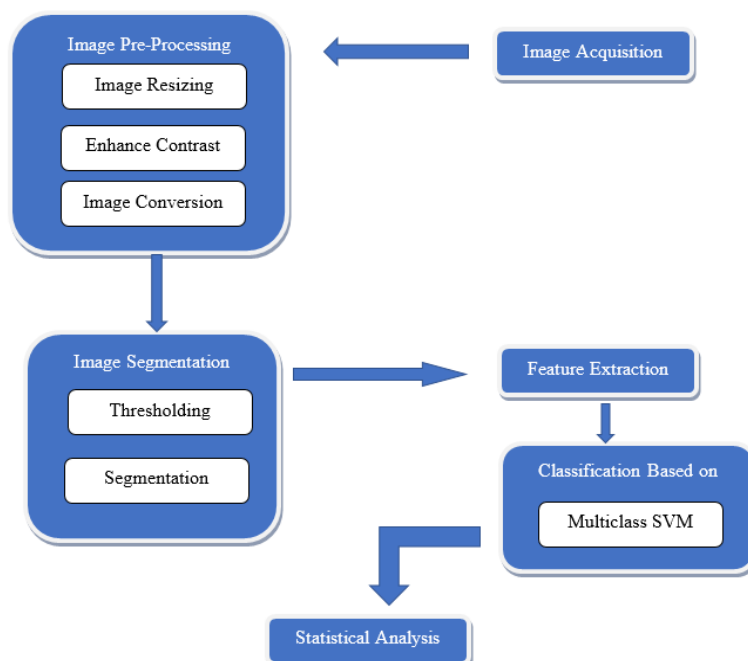

Fig 2.2 Work Flow

Properties like boundaries and edges are better viewed in black and white images; statistical properties related to intensities are observed in greyscale format, and the information related to color is seen well in RGB and other color formats of the image.

In this system, the image will be resized to 256x256 and thresholding is done using Otsu's method which converts the intensity image to binary image. Convert the RGB image format to a gray-scale image. Input image's histogram is used to compute the mean of the distribution and then scaled to a normalized value between 0 and 1.

Image Segmentation

During image segmentation, the given image is separated into a homogeneous region based on certain features. Larger data sets are put together into clusters of smaller and similar data sets using clustering method. In this proposed work, K-means clustering algorithm is used in segmenting the given image into three sets as a cluster that contains the diseased part of the leaf. Since we have to consider all of the colors for segmentation, intensities are kept aside for a while and only color information is taken into consideration. The RGB image is transformed into LAB form (L-luminous, a*b-chromous). Of the three dimensional LAB, only last two are considered and stored as AB. As the image is converted from RGB to LAB, only the "a" component i.e. the color component is extracted.

Properties and process of K-Means Algorithm are as follows:

Properties

   i   K number of the cluster should be present always.
   ii   In each given cluster, at least one item should be present.
   iii   Overlapping of clusters should never happen.
   iv   Every participant of the single cluster should be close to its own cluster than any other cluster

Process

1. The given data set should be divided into K number of clusters and data points need to be assigned to each of these clusters randomly.
2. For each data point, the distance from data point to each cluster is computed using Euclidean distance The Euclidean distance is nothing but the distance between two-pixel points and is given as follows:
   Euclidean Distance=$\sqrt{(x1-x2)^2 + (y1-y2)^2}$ where (x1, y1) & (x2, y2) are two pixel points (or two data points).
3. The data point which is nearer to the cluster to which it belongs to should be left as it is.
4. The data point which is not close to the cluster to which it belongs to should be then shifted to the nearby cluster.
5. Repeat all the above steps for entire data points.
6. Once the clusters are constant, clustering process needs to be stopped.

Feature Extraction

From the input images, the features are to be extracted. To do so instead of choosing the whole set of pixels we can choose only which are necessary and sufficient to describe the whole of the segment. The segmented image is first selected by manual interference. The affected area of the image can be found from calculating the area connecting the components. First, the connected components with 6 neighborhood pixels are found. Later the basic region properties of the input binary image are found. The interest here is only with the area. The affected area is found out. The percent area covered in this segment says about the quality of the result.

The histogram of an entity or image provides information about the frequency of occurrence of certain value in the whole of the data/image. It is an important tool for frequency analysis. The co-occurrence takes this analysis to next level wherein the intensity occurrences of two pixels together are noted in the matrix, making the co-occurrence a tremendous tool for analysis.

From gray-co-matrix, the features such as Contrast, Correlation, Energy, Homogeneity' are extracted. The following table lists the formulas of the features.

Table 2.1 List of formulas

No	Features	Formula
1	Contrast	$\sum_i \sum_j \lvert i - j \rvert^2 p(i, j, d, \theta)$
2	Correlation	$\sum_{i,j} \dfrac{(i - \mu_i)(j - \mu_j)p(i,j)}{\sigma_i \sigma_j}$
3	Energy	$\sum_i \sum_j p(i, j, d, \theta)^2$
4	Homogeneity	$\sum_i \sum_j p(i, j, d, \theta)/(1 + \lvert i - j \rvert)$

Using the statistical MATLAB commands the other properties are found out. Those are Mean Standard Deviation, Entropy, RMS, Variance, Smoothness, Kurtosis, Skewness, and IDM.

Mean: Average or mean value of the array. Mean is given by

$$\text{Mean} = \left(\frac{1}{N}\right) Xi$$

Where Xi->pixel intensity, N->a total number of pixels of an image.

Standard Deviation: Standard deviation is computed using the below formula:

$$\text{Standard Deviation} = (1/N) \, \Sigma \, (Xi - \mu) \, ½$$

Where μ-> mean.

Entropy: Entropy is a statistical measure of randomness that is used to characterize the texture of the input image. Entropy is defined as

$$\text{Entropy} = -\text{sum} \, (p.*\log2 \, (p))$$

Where p-> histogram counts.

Variance: Variance is computed using

$$\text{Variance} = (1/N) \, \Sigma \, (Xi - \mu)^2$$

Variability is measured using variance.

Skewness: The image surface is judged with the Skewness.

$$\text{Skewness} = \frac{\frac{1}{N} \sum_{i=1}^{N} (x_i - x')^3}{\left(\frac{1}{N} \sum_{i=1}^{N} (x_i - x')^2\right)^{\frac{3}{2}}}$$

The same feature set is used for training the SVM as well to identify the class of the input image.

Training
1. Start with images of which classes are known for sure.
2. Find the property set or feature set for each of them and then label suitable.
3. Take the next image as input and find features of this one as new input.
4. Implement the binary SVM to multi class SVM procedure.
5. Train SVM using kernel function of choice. The output will contain the SVM structure and information of support vectors, bias value etc.
6. Find the class of the input image.
7. Depending on the outcome species, the label to the next image is given. Add the features set to the database.
8. Steps 3 to 7 are repeated for all the images that are to be used as a database.
9. Testing procedure consists of steps 3 to 6 of the training procedure. The outcome species is the class of the input image.
10. To find the accuracy of the system or the SVM, in this case, random set of inputs are chosen for training and testing from the database.

Two different sets for train and test are generated. The steps for training and testing are same, however, followed by the test is performed.

Classification

The binary classifier which makes use of the hyper-plane which is also called as the decision boundary between two of the classes is called as Support Vector machine (SVM). Some of the problems of pattern recognition like texture classification make use of SVM. Mapping of nonlinear input data to the linear data provides good classification in high dimensional space in SVM. The marginal distance is maximized between different classes by SVM. Different kernels are used to divide the classes. SVM is basically binary classifier which determines the hyper plane in dividing two classes. The boundary is maximized between the hyper plane and the two classes. The samples that are nearest to the margin will be selected in determining the hyper plane are called as support vectors.
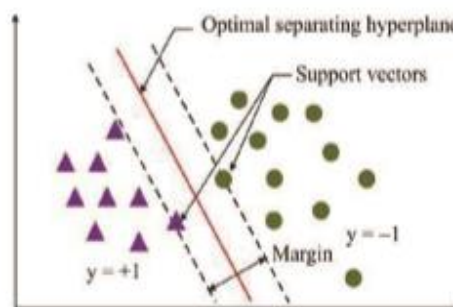

Fig 2.3 Linear SVM

Fig 2.3 shows the concept of support vector machine. Multiclass classification is also possible either by using one-to-one or one-to-many. The highest output function will be determined as the winning class.

Classification is performed by considering a larger number of support vectors of the training samples. The standard form of SVM was intended for two-class problems. However, in real life situations, it is often necessary to separate more than two classes at the same time. In this Section, we explore how SVM can be extended from binary problems to multi classification problems with k classes where k > 2. There are two approaches, namely the one-against-one approach and the one-against-all approach.

In fact, multi-class SVM converts the data set to quite a few binary problems. For example, in one-to-one approach binary SVM is trained for every two classes of data to construct a decision function. Hence there are k (k−1)/2 decision functions for the k-class problem. Suppose k = 15, 105 binary classifiers need to be trained. This suggests large training times. In the classification stage, a voting strategy is used where the testing point is designated to be in a class having the maximum number of votes. The voting approach is called the "Max Wins" strategy.

In one-against-all approach, there will be one binary SVM for each of the class to isolate the members of one class from the other class.

III. RESULTS AND PERFORMANCES

To have a closer look at the results and interim outcomes of this work variety of inputs are necessary. Some of the variations in same class images are to be observed.
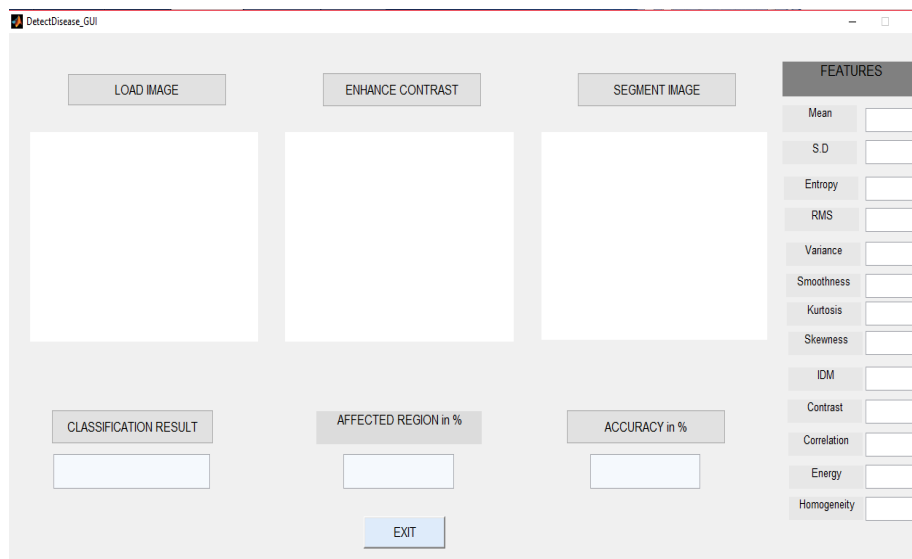

Fig 3.1 Home Page

In the above figure the input image should be loaded, then the contrast of the image will be enhanced. The image then undergoes segmentation.
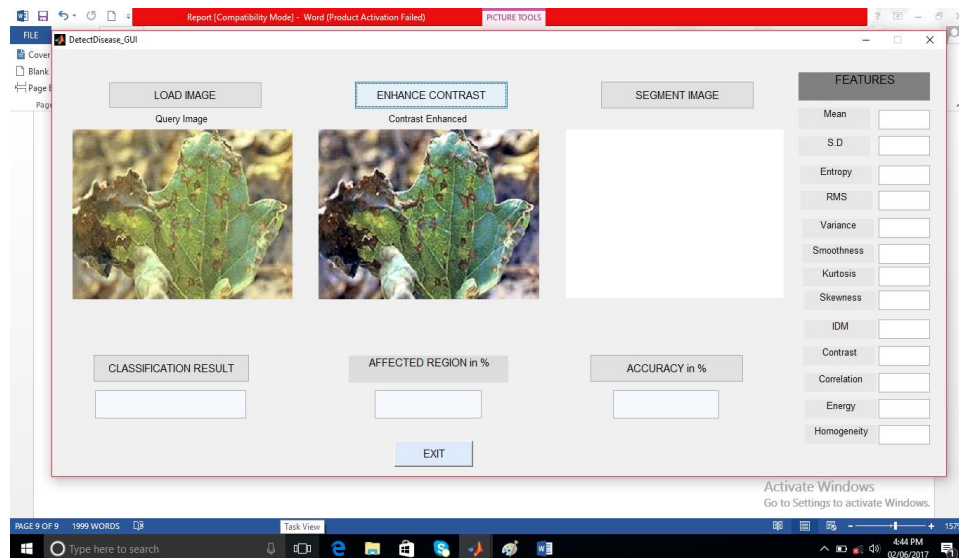

Fig 3.2 Contrast Enhanced Image

The input image is first preprocessed and the contrast is enhanced so that the minute details are visible on a larger scale. Then the enhanced image is converted into gray format representation. The second image in Fig 3.2 shows the improved version of the input image. The image will be contrasted and then displayed. Now select the Segment Image option.
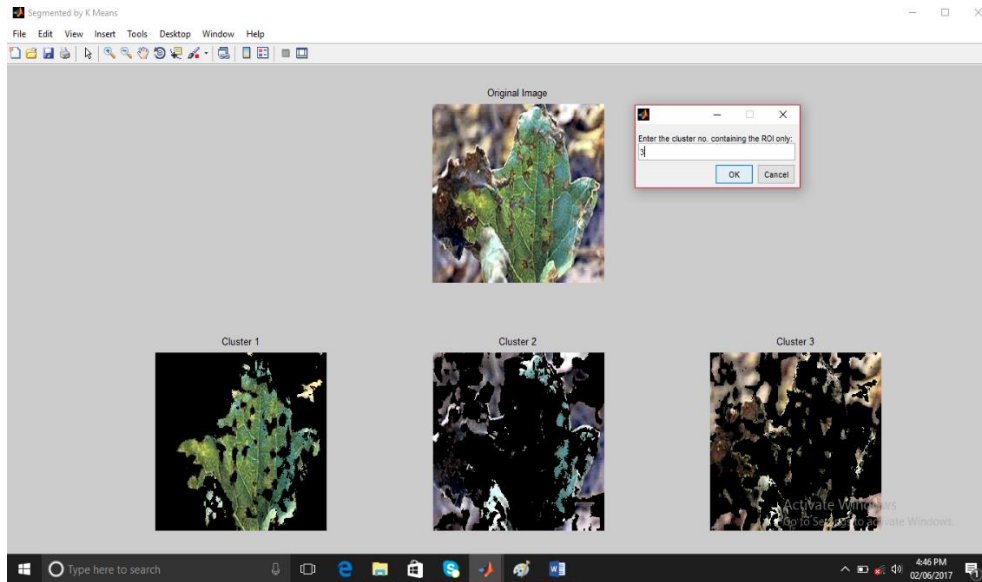
Fig 3.3 Segmented images

Select one among the segmented images in which the disease can be identified clearly. The selected image will then be used for further processing.
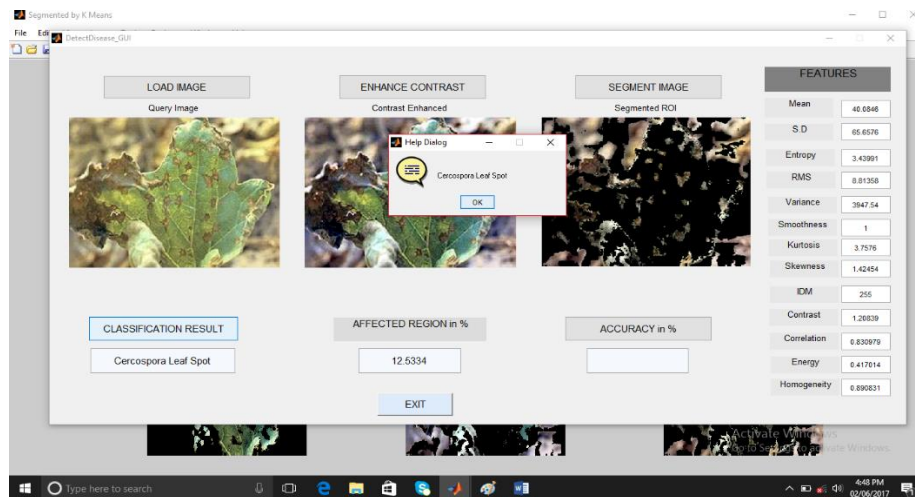

Fig 3.4 Classification of disease

In fig 3.4 among the segmented images, one image is selected and classified based on th ROI,. The classifier detects that the input leaf image belongs to the Cercospora Spot disease type.
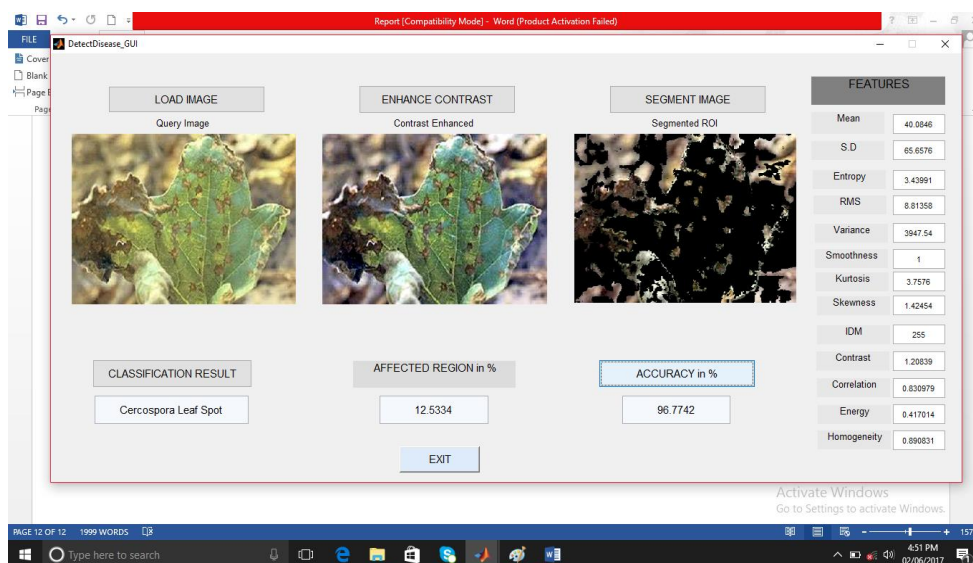

Fig 3.5 Accuracy Computation

The computed accuracy will be displayed. For accuracy calculation, the kernel function is changed and the cross validation along with class performance is used.

Accuracy of Linear Kernel with 500 iterations is: 96.7742%

Accuracy of RBF Kernel with 500 iterations is: 95.1613%

Accuracy of Polynomial Kernel with 500 iterations is: 96.7742%

The accuracy for different kernel emphasizes upon the spread of space vector (points) in space. That is since the accuracy for the linear kernel is highest it means the space vector describing the features set is more linear than radial basis function than the polynomial spread.

FUTURE SCOPE

SVM, though a binary classification technique, with a simple manipulation, can be used for a multiple class case. This provides more space not just to classify but to identify the diseases. Presently the system is semi-automatic. This can be completely automatic by choosing ROI based on criterion such as principal components analysis, or choosing the cluster with larger disease area etc.
With the proper database, this method can be applied to more diseases. Example: liver diseases, skin cancer, breast cancer identification and classification etc.

CONCLUSIONS

The world is moving more towards technology dependent era. Every day we keep hearing owes of farmers that even after using costly fertilizers the leaves were eaten away by various diseases. One of the most sensitive and costly treatments in India in terms of leaf concerned is that of pomegranate. The expertise in this field is rarely available. Since the opinion of an expert can vary from that of a novice, for the benefit of all it is advisory to make the most use of the technology available to infer or conclude for treatments.
The machine learning methods bring this aspect to reality, by observing the database and helping the botanists in the diagnosis of diseases where a lot of precision is required. And one of the machine learning technique, SVM is used in this project for classification of leaf diseases.
The accuracy results in an available range from mid-90 to top 90%. This can be bettered by increasing the database. However, the results obtained from real life images are very encouraging.

ACKNOWLEDGEMENT

REFERENCES

[1]"Classification of diseases plant leaves using Neural Network Algorithms" K. Muthukannan1, P. Latha2, R. Pon Selvi1 and P. Nisha1 1Department of ECE, Einstein College of Engineering, Anna University, Tirunelveli, India 2 Department of CSE, Government College of Engineering, Anna University, Tirunelveli, India
[2] S. S. Sannakki, V. S. Rajpurohit, V. B. Nargund, and P. Kulkarni, "Diagnosis and Classification of Grape Leaf Diseases using Neural Networks", IEEE 4th ICCCNT, 2013.
[3] P. Chaudhary, A. K. Chaudhari, Dr. A. N. Cheeran and S.Godara, "Color Transform Based Approach for Disease Spot" International Journal of Computer Science and Telecommunications Volume 3, Issue 6, pp.65-70, June 2012
[4] H. Al-Hiary, S. Bani-Ahmad, M. Reyalat, M. Braik, and Z. ALRahamneh, "Fast and Accurate Detection and Classification of Plant Diseases", IJCA, Vol-17, No.-1, pp. 31-38, March 2011.
[5] Suman T. and Dhruvakumar T., "Classification of paddy leaf diseases using shape and color features", IJEEE, Volume 07, Issue 01, PP.239250, Jan- June 2015.