Tea Leaf Disease Detection

Using

Digital Image Processing

By
Debjit Paul
Roll No: 15010109
&
Noor Md. Ayesh
Roll No: 15010118

Supervisor:
Md. Zainal Abedin
Assistant Professor
Dept. of CSE
FSET, USTC

Department of Computer Science and Engineering
Faculty of Science, Engineering and Technology
University of Science and Technology Chittagong
Chittagong, Bangladesh

November 2019

Tea Leaf Disease Detection

Using

Digital Image processing

By
Debjit Paul
Roll No: 15010109
&
Noor Md. Ayesh
Roll No: 15010118

A report Submitted in Partial Fulfillment of the Requirements of the University of Science and Technology Chittagong for the Degree of Bachelor of Science in Computer Science and Engineering

Supervisor:
Md. Zainal Abedin
Assistant Professor
Dept. of CSE
FSET, USTC

Department of Computer Science and Engineering
Faculty of Science, Engineering and Technology
University of Science and Technology Chittagong
Chittagong, Bangladesh

November 2019

Abstract:

Tea is a popular beverage all around the world, and in Bangladesh the cultivation of tea plays a vital role. Many diseases affect the proper growth of tea leaves leading to its reduction, thus hindering of the production of tea. However, if the disease is identified at an early age it would solve all the above mentioned problems through the application of appropriate treatment, or through the pruning of the diseased leaves to prevent further spread of the disease. To solve this problem image processing is the best option to detect and diagnose the disease. The main goal of this research is to develop an image processing system that can identify and classify most widespread tea leaf diseases in Bangladesh, namely Brown Blight/Grey Blight, Leaf Rust, disease and the algal leaf disease, red rust, from a healthy leaf. Disease identification is the first step; there are many methods that have been used for identifying the leaf disease. In this paper, Support Vector Machine classifier (SVM) is used to recognize the diseases. Thirteen features are analyzed during the classification. These features are then used to find the most suitable match for the disease (or normality) every time an image is uploaded into the SVM database. When a new picture is uploaded into the system the most suitable match is found and the disease is recognized. The approach is novel since the number of features compared by the SVM classifier is reduced by three features compared to previous researches, without adversely sacrificing the success rate of the classifier, which retains an accuracy of more than 90%. This also speeds up the identification process, with each leaf image taking 300ms less processing time compared to previous research using SVM, thus ensuring a greater number of leaves can be processed in a given time frame. The proposed solution increases in efficiency of the detection, identification, and classification process will enable the tea industry in Bangladesh to become more competitive globally, by reducing the losses suffered due to diseases of the leaf, and thus increasing the overall tea production rate.

Keywords—Image processing, Disease detection, Disease recognition, Feature Extraction, Support Vector Machine

Declaration

We hereby proclaim that this thesis is based on the results we found by our hard work. Contents of the work found by other researcher(s) are motioned by references. This thesis has never been previously submitted for any degree neither in whole nor in part.

Signature of Supervisor:

_____

Md. Zainal Abedin

Assistant Professor

Faculty of Science, Engineering & Technology

University of Science & Technology Chittagong

Signature of Students:

_____	**_____**
Debjit Paul	**Noor Md. ayesh**
[Student ID: 15010109]	**[Student ID: 15010119]**

Approval Certificate

The research entitled "Tea Leaf disease Detection By Digital Image processing" submitted by Debjit Paul(Roll: 15010109) and Noor Md Ayesh (Roll:15010118) has been accepted as satisfactory in partial fulfillment of the requirements for the degree of B.Sc. Engineering in Computer Science and Engineering on February, 2019.

Board of Examiners

_____

Mr. Kazi Nur-e Alam Siddique
Head, Department of CSE
Faculty of Science, Engineering and Technology
University of Science and Technology Chittagong

_____

Md. Zainal Abedin
Assistant Professor
Department of CSE
Faculty of Science, Engineering and Technology
University of Science and Technology Chittagong

_____

Mr. Sukanta Paul
Lecturer, Department of CSE
Faculty of Science, Engineering and Technology
University of Science and Technology Chittagong

Acknowledgement

First and foremost, we would like to thank to the God for his mercy for which we are able to complete our project successfully in time.

We want to thank our supervisor Mr. Zainal Abedin, Assistant Professor, Department of Computer and Engineering of USTC for giving us the opportunity to work under him and also give us all kind of support to guide us properly for making the research successful. Without his support, suggestions, and advice it would not be possible for us to learn so many things about data processing and various topic related with our thesis.

At last, we would like to express our special thanks to all of my well-wishers, friends and all who has contributed in any way to complete this research.

“Dedicated to our family for their support, during our research”

Table of Contents

Page

List of Figures

List of Tables

Chapter-1

1. Introduction:

In 2012, Bangladesh recorded its highest production of tea, at 63.85 million kilograms. The country has over 56,846 hectares of land under tea cultivation, up from 28,734 hectares in 1947.The government has begun to promote small-scale tea growers, particularly in the Chittagong Hill Tracks. Tea plants tolerate elevated levels of heat and shade, and thus, areas where tea plantations are typically found are characterized by warm climates and abundant rainfall. However, these regions are also very conducive to the growth and reproduction of diseases that have severely decreased tea quality with the gradual increase in tea production. Consequently, tea diseases are a limiting factor hindering robust tea production. Plant disease diagnosis is typically based on the characteristic appearances of diseases. However, trained tea plant pathologists are scarce, and limitations in the background knowledge of tea growers leads to an inability to identify disease events in a timely and effective manner. Therefore, development and implementation of a diagnostic framework for tea plant diseases would help ensure the accurate and timely identification of tea plant diseases by agricultural producers. Such improvements would lead to better control methods that would economically and effectively restore losses due to diseases. Moreover, these advancements would help ensure higher tea qualities while reducing costs of labor and agricultural production, and importantly, thereby improving yields and the sustainable development of tea production. The current methods for diagnosing plant diseases primarily include microscopic identification in addition to molecular biological and spectroscopic techniques. Microscopic identification is time consuming and can be subjective, where even experienced plant pathologists may incorrectly diagnose diseases. Molecular biological and spectroscopic identification are more accurate but are labor intensive and require specialized and expensive instrumentation. The rapid development of intelligent agriculture and precision agriculture in recent years has led to the widespread use of computer image processing technologies to solve diverse problems within agricultural sciences. For example, these technologies have been used to estimate plant nutrient content, classify plant species, and identify plant diseases. In particular, deep neural network and genetic algorithms have been used in combination to estimate nitrogen content in wheat leaves, which represents a considerable improvement over other existing methods. Artificial neural networks (ANNs) and Support Vector Machines (SVMs) have been used in this capacity. ANNs reflect biological neural networks and autonomously learn, progressively improving their knowledge and capacities. The multi-layer perceptron (MLP) model is a multi-layer feed forward artificial neural network model that exhibits superior performance when analyzing nonlinear systems. MLPs typically comprise the input, hidden, and output layers, where each layer contains several neurons, the sigmoidal linear activation functions was used to map the sum of the weighted inputs to the output of a neuron in the hidden layer. SVMs are an effective type of classification algorithm that has been widely used for many pattern recognition tasks, including machine learning methods based on statistical learning theory. The primary goal of SVMs is to identify a separating plane to evaluate different class memberships. SVMs were initially used to classify two-class problems in the analysis of linear separable cases. In the event of linear inseparability, nonlinear mapping algorithms can be used to transform linearly inseparable samples of low-dimensional input space into high-dimensional feature space in order to render them linearly separable. The technique is based on the structural risk minimization theory that informs the construction of an optimal hyperplane in feature space, such that the learner is globally optimized and the expectation in the entire sample space

meets a certain upper bound with a certain probability. These two methods require smaller sample sizes and an appropriate train rule, which have led to their widespread use in image classification and recognition. Convolutional neural networks (CNNs) were developed in the 1980s and are a type of deep neural network which were used to recognize handwritten digits. CNNs could learn to extract features from images by themselves through stacked layers of convolutional filters. Typical CNNs are hierarchical neural networks that are primarily composed of multiple convolution layers, a pooling layer, and a full connection layer. Local receptive fields, weight sharing, and spatial sub sampling are the primary hierarchical aspects within the networks. These attributes result in a high invariance of CNNs for translation, scaling, shifting, or other forms of deformation. Moreover, CNNs directly take the image as a network input, thus avoiding the extraction of complex features and the need for data reconstruction, as in traditional image recognition algorithms. Meanwhile, the high recognition accuracy of CNNs leads to wide implementation in fields related to computer vision, where development is occurring rapidly.The rapid development of computer vision technology in recent years has led to increased usage of computational image processing and recognition methods to identify diseases. The most widely used current method relies on extracting global features including color features, shape features, texture features, or some combination of the above features. The local features of the disease spots are then processed using various algorithms including local feature SIFT (scale-invariant feature transform), SURF (speeded-up robust features), HOG (histogram of oriented gradient), DSIFT (dense scale-invariant feature transform), and PHOW (pyramid histograms of visual words) Lastly, the extracted feature parameters are used in various classifiers including ANNs and SVMs. A significant drawback of these methods is the need to artificially extract features in advance. In contrast, CNNs learn data characteristics from convolution operations, which is better suited for pattern recognition of images. Consequently, CNNs have been used to detect and diagnose plant diseases.Following these previous studies, CNNs were developed and adopted here to improve the diagnosis and classification of tea diseases. In addition, the classification performances of traditional machine learning algorithms were evaluated relative to manual classifications and the proposed CNNs. Among the former, the HOG, SURF, and PHOW algorithms did not exhibit better invariance towards image rotation and scaling than SIFT, while the SIFT algorithm performed reliably with affine transformations, viewing angle variation, and noise. Moreover, the SIFT algorithm exhibited strong scalability, that when combined with other algorithms could be used as a highly optimized algorithm. Consequently, SIFT was used here as a feature descriptor in a traditional machine learning algorithm. Although SIFT features can describe images, each SIFT represents a 128-dimensional vector, and images contain hundreds or thousands of SIFT features, thereby leading to very computationally intensive operations. To greatly reduce computational processing, a bag of visual words (BOVW) model was constructed based on these vectors, wherein each image was represented by a numerical vector.

1.1 Motivation:

Agricultural productivity is that thing on which Economy of Bangladesh is highly dependent. This is the one of the reasons that disease detection in plant plays an important role in agricultural field, as having diseases in plants are quite natural. If proper care is not taken in this area then it causes serious effects on tea plants and due to which respective product quality, quantity or productivity is affected. Detection of plant disease through some automatic technique as it reduce a large work of monitoring in tea garden.

1.2 Related Works:

Soybean is a very essential crop which plays a key role in the complete food chain. But due to the different reasons like diseases, pest attack & suddenly changing weather conditions, the productivity of soybean crop decreases qualitatively and quantitatively. To minimize the loss in productivity of soybean crop the early stage disease detection is required preventive measure. An application of Image processing technique, in the field of agriculture is emerging exponentially which is ranging from the field management to crop disease detection. This paper presents study of soybean leaves colored images for disease detection by inspecting the visual symptoms of particular disease by using segmentation. Soybean is prone to many diseases and traits. To control these diseases farmers are experiencing so much toughness while making the one disease control policy to another. Always insecticides are not more efficient to control diseases, as the diseases on soybean leaves causes major loss in production and economic loss in the soybean crop. Hence to take the corrective action, digital image processing technique is used to detect diseased leaves. In this work the kinds of methods are used to observe and learn the soybean leaves diseases using digital image processing.

1.3 Tea leaf:

Tea, *Camellia sinensis*, is a tree or small shrub in the family Theca grown for its leaves which are used to make beverages. The tea plant is branching with alternate elliptical leaves. The leaves are leathery in texture, matte green in color and have serrated edges. The tea plant can take the form of a tree with a bowl-shaped canopy but is usually pruned under cultivation to be smaller and shrub-like. The plant produces fragrant white flower singly or in small clusters. Tea tree can reach up to 15 m (49 ft.) in height and can live anywhere between 30 and 50 years. The plant originates from China.

1.3.1 Causes of leaf diseases:

From the viewpoint of plant protection (phytomedicine),the demand for tea free from pesticide residues is in conflict with the demand for high quality from the consumer, and with the demand for high yield and low labor input, from the producer. High yield in tea has been mainly achieved by the elimination of tree shade, and reduction of losses due to diseases and pests with pesticides. Elimination of shade changed the agro-environment, with increased growth of weeds, higher input of fertilizers, and with increased susceptibility to certain diseases and pests (mites). However, shade also can be detrimental to yield with unsuitable tree species, or too much shade in the rainy season which encourages the incidence of blister blight and some insect pests (tea mosquito bug). Economic loss of tea due to diseases is higher compared to animal pests (pests), the blister blight being the main disease. Pressure of diseases and pests on tea depends also on the control strategy and the climatic environment. Reducing pesticides may be feasible by reducing the threat of diseases and/or pests through cultivation in Jess disease and pest prone environment (altitude, shade), in choosing disease tolerant clonal tea-varieties, by choosing pesticides with low interference on natural enemies of pests and diseases, and by applying pesticides according to economic threshold . The long standing time of the tea bushes favors slow developing root diseases. These are reviewed and the integrated control is discussed in detail with eradication, with pesticides, and with improvement of soil micro flora.

1.3.2 Overview of tea leaf diseases:

Algal leaf spot:

Pathogen*: Cephaleuros virescens*

Symptoms: Leaves develop lesions that are roughly circular, raised, and purple to reddish-brown.

Life cycle:

The alga produces microscopic, rust-colored, spore-like bodies on the surface of the leaf spots, giving them a reddish tinge. The "spores" are dispersed by wind or rain. The alga may spread from leaves to branches and fruit. Poor soil drainage, imbalanced nutrition, and exposure to relatively high temperature and humidity predispose tea plants to infection by algal leaf spot, so it is important to strengthen the plant through proper cultivation and fertilization. Most algal spots develop on the upper leaf surface. Older infections become greenish-gray and look like lichen. Cephaleuros usually does not harm the plant.

Brown blight, grey blight

Pathogens: *Colletotrichum sp. Pestalotiopsis sp.*

These fungi are considered weak pathogens and usually only affect plants that have been weakened by improper care or adverse environmental conditions. The disease is favored by poor air circulation, high temperature, and high humidity or prolonged periods of leaf wetness. When young twigs of susceptible cultivars are cut and used to root new plants, latent mycelium in the leaf tis-sue may start to invade nearby cells to form brown spots, and this may lead to death of leaves and twigs.

Symptoms: Small, oval, pale yellow-green spots first appear on young leaves. Often the spots are surrounded by a narrow, yellow zone. As the spots grow and turn brown or gray, concentric rings with scattered, tiny black dots be-come visible and eventually the dried tissue falls, leading to defoliation. Leaves of any age can be affected.

Life cycle:

The tiny, black spots on the lesions contain the fungal spores. Rain splash transports the spores from one plant or site of infection to another. If the spores land on a leaf, they germinate to start a new leaf spot or a latent infection.

Leaf Rust

Pathogens: *Exobasidium vexans*

Leaf Rust is the most serious disease affecting shoots of tea and is capable of causing enormous crop loss. The disease is endemic to most tea-growing areas of Asia but is not known to occur in Africa or the Americas. Cloudy, wet weather favors infection. Shan or Indian varieties of tea are somewhat resistant to this disease.

Symptoms: Small, pinhole-size spots are initially seen on young leaves less than a month old. As the leaves develop, the spots become transparent, larger, and light brown. After about 7 days, the lower leaf surface develops blister-like symptoms, with dark green, water-soaked zones sur-rounding the blisters. Following release of the fungal spores, the blister becomes white and velvety. Subsequently the blister turns brown, and young infected stems become bent and distorted and may break off or die

Life cycle:

The disease cycle repeats continuously during favorable (wet) conditions, and the spores are readily dispersed by wind. Spores that land on a leaf with adequate moisture will germinate and infect it, producing visible symptoms within 10 days. The fungus can directly penetrate the leaf tissue. The basidio spores have a low survival rate under conditions of drought or bright sunlight. The life cycle of the fungus is 3–4 weeks.

Red rust:

Red Rust is a very popular disease of the tea plant (Camellia sinensis). Orange brown, velvety areas appear on the leaves of infected plants. The disease is caused by algae of the genus Cephaleuros.

- •Causal organisms: Cephaleuros parasiticus
- •It's the most important algal disease in tea.
- •Vigorous growing and high-yielding tea cultivars has been found highly attacked.

Nature of Pathogen

•Cephaleuros infection on tea has been called as 'Red rust'.

•They are the member of Trentepohliales, the genus chlorophyte which contains the photosynthetic organisms called green algae.

•They are aerophilic, filamentous green algae.

•They consists of branched filaments that comprises a thallus in the form of irregular discs.

•They are capable of both asexual and sexual reproduction. Asexual reproduction is much more important to typical infection and disease process.

•Unlike majority of pathogen Cephaleuros species it penetrates epidermis of plants.

Symptoms:

•Red rust causes severe damage to young tea by attacking and killing stem and leaf tissues in patches.

•Affected patches on the stem become most noticeable when the alga produces fruitification which appears during the month of April to July.

•The patches are oval or oblong in shape.

•During August- March the lesions appear purplish in color there are no fruiting hair at this time but longitudinal cracks may be seen on surface

•It causes chlorosis and variegation.

•It causes defoliation, reduce photosynthesis, loss of fruit marketability, twig dieback and tissue necrosis.

Red Spider:

Symptoms: abnormal colors, abnormal leaf fall, external feeding, necrotic areas

Chapter-2

2. Machine Learning for Classification:

Machine learning classifications are processes of predicting the class of given data points using machine learning approaches. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

For example, spam detection in email service providers can be identified as a classification problem. This is s binary classification since there are only 2 classes as spam and not spam. A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known spam and non-spam emails have to be used as the training data. When the classifier is trained accurately, it can be used to detect an unknown email.

Classification belongs to the category of supervised learning where the targets also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.

Classifiers can be:

Binary classifiers: Classification with only 2 distinct classes or with 2 possible outcomes

Example: Male and Female

Example: classification of spam email and non-spam email

Example: classification of author of book

Example: positive and negative sentiment

Multi-Class classifiers: Classification with more than two distinct classes.
Example: classification of types of soil

Example: classification of types of crops

Example: classification of mood/feelings in songs/music

2.1 Support Vector Machine:

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (*supervised learning*), the algorithm outputs an optimal hyper plane which categorizes new examples. In two-dimensional space this hyper plane is a line dividing a plane in two parts where in each class lay in either side.

Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

More formally, a support-vector machine constructs a hyper plane or set of hyper planes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks like outlier's detection. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.

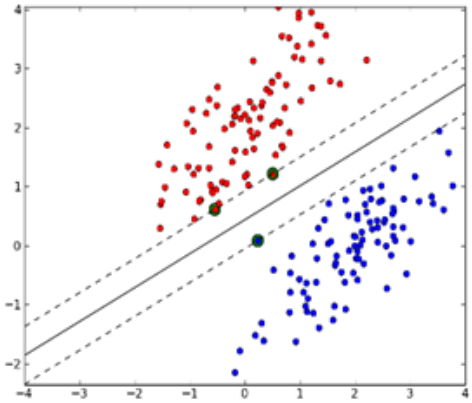

Figure 2.1: Support Vector Machine

Description of diagram:
H1 is not a good hyperplane as it doesn't separate the classes
H2 does but only with small margin
H3 separates them with maximum margin (distance)

Whereas the original problem may be stated in a finite-dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products of pairs of input data vectors may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function selected to suit the problem. The hyper planes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant, where such a set of vectors is an orthogonal (and thus minimal) set of vectors that defines a hyper plane.

SVMs can be used to solve various real-world problems:

SVMs are helpful in text and hypertext categorization, as their application can significantly reduce the need for labeled training instances in both the standard inductive and transductive settings. Some methods for shallow semantic parsing are based on support vector machines.

Classification of images can also be performed using SVMs. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback. This is also true for image segmentation systems, including those using a modified version SVM that uses the privileged approach as suggested by Vapnik. Hand-written characters can be recognized using SVM.

The SVM algorithm has been widely applied in the biological and other sciences. They have been used to classify proteins with up to 90% of the compounds classified correctly. Permutation tests based on SVM weights have been suggested as a mechanism for interpretation of SVM models. Support-vector machine weights have also been used to interpret SVM models in the past. Post hoc interpretation of support-vector machine models in order to identify features used by the model to make predictions is a relatively new area of research with special significance in the biological sciences.

2.2 **K-means Clustering in Machine Learning**

K-means clustering is a partitioning method. The function *k*-means partitions data into *k* mutually exclusive clusters and returns the index of the cluster to which it assigns each observation. *K-means* treats each observation in your data as an object that has a location in space. The function finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. You can choose a distance metric to use with *k*-means based on attributes of your data. Like many clustering methods, *k*-means clustering requires you to specify the number of clusters *k* before clustering.

Unlike hierarchical clustering, *k*-means clustering operates on actual observations rather than the dissimilarity between every pair of observations in the data. Also, *k*-means clustering creates a single level of clusters, rather than a multilevel hierarchy of clusters. Therefore, *k*-means clustering is often more suitable than hierarchical clustering for large amounts of data.

Each cluster in a *k*-means partition consists of member objects and a centroid (or center). In each cluster, *k*-means minimizes the sum of the distances between the centroid and all member objects of the cluster *k*-means computes centroid clusters differently for the supported distance metrics.

2.3 Multilayer Perceptron:

A multilayer perceptron (MLP) is a class of feed forward artificial neural network. A MLP consists of, at least, three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

Multilayer perceptron are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer.

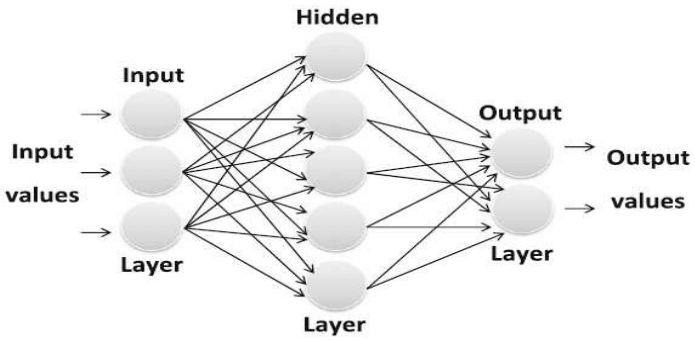

Figure 2.2: Multilayer Perceptron classification of ANN

The MLP consists of three or more layers (an input and an output layer with one or more *hidden layers*) of nonlinearly-activating nodes. Since MLPs are fully connected, each node in one layer connects with a certain weight to every node in the following layer.

The term "multilayer perceptron" does not refer to a single perceptron that has multiple layers. Rather, it contains many perceptron that are organized into layers. An alternative is "multilayer perceptron network". Moreover, MLP "perceptron" are not perceptron in the strictest possible sense. True perceptron are formally a special case of artificial neurons that use a threshold activation function such as the Heaviside step function. MLP perceptron can employ arbitrary activation functions. A true perceptron performs binary classification (either this or that), an MLP neuron is free to either perform classification or regression, depending upon its activation function.

The term "multilayer perceptron" later was applied without respect to nature of the nodes/layers, which can be composed of arbitrarily defined artificial neurons, and not perceptron specifically. This interpretation avoids the loosening of the definition of "perceptron" to mean an artificial neuron in general.

MLPs are useful in research for their ability to solve problems stochastically, which often allows approximate solutions for extremely complex problems like fitness approximation. MLPs are universal function approximation as showed by Cybenko's theorem, so they can be used to create mathematical models by regression analysis. As classification is a particular case of regression when the response variable is categorical, MLPs make good classifier algorithms.

MLPs were a popular machine learning solution in the 1980s, finding applications in diverse fields such as speech recognition, image recognition, and machine translation software, but thereafter faced strong competition from much simpler (and related) support vector machines. Interest in back propagation networks returned due to the successes of deep learning.

Chapter-3

3.1 Proposed Methodology:

To identify the affected area the images of various leaves are taken with a digital camera or similar device. Then to process those images various image processing technique are applied on them to get different and useful features required for later analyzing purpose.

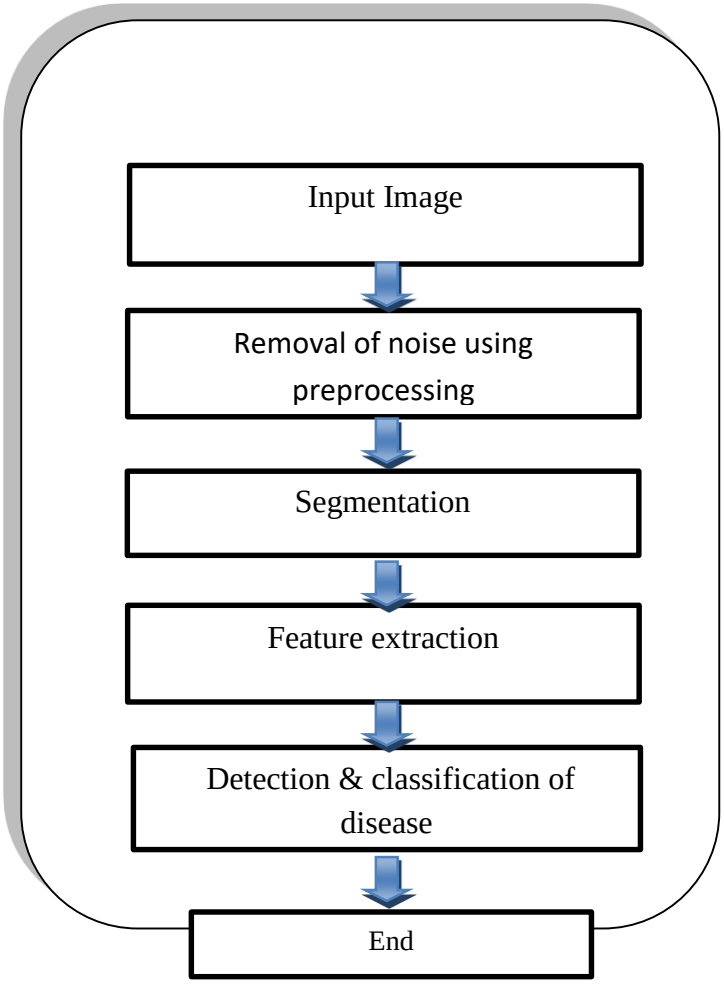

Figure 3.1: Proposed Methodology

3.2 Disease Dataset:

Images showing tea leaf diseases were all captured using **Nikon D3300 and D3100** camera in the natural environments of Sree Mongol, Sylhet The images were taken ~20 cm directly above the leaves and captured in the auto-focus mode at a resolution of 4000×3000 pixels. A total of 1402 tea leaf images were used that showed symptoms for seven different diseases, as identified by phyto-pathologists (Figure 1).The identification criteria used for the tea tree diseases were based on previously described identification schemes [35, 36]. All images in the present manuscript were resized to 256×256 pixels. In order to improve the classifier's generalization ability, we increased the size of the dataset, which is more advantageous to the training of the network. Three different methods were used to alter the image input and improve classification (Figure 2).

Finally, there are 1705 images in the database. Table 1 shows the number of images for every class used as training, validation and testing datasets for the disease classification model. The disease classification datasets that were used in these analyses are shown in Table 1.The 80/20 ratio of training/test data is the most commonly used ratio in neural network applications.

In addition, a 10% subset of the test dataset was used to validate the dataset.

Class	Number of images from the Dataset used for Training	Number of images from the Dataset used for Validation	Number of images from the Dataset used for Testing
Leaf Rust	**304**	**30**	**32**
Red Rust	**180**	**18**	**20**
Red spider	**210**	**21**	**25**
Gray Blight	**304**	**30**	**35**
Brown Blight	**100**	**10**	**20**
Algal Leaf Spot	**304**	**30**	**32**
	Total 1402	**Total 139**	**Total 164**

Table 3.1: The image dataset comprising of six different diseases

3.3 Performance Measurements:

The accuracy and mean class accuracy (MCA) indices were used to evaluate algorithm performances, as previously described. CCRk is first defined as the correct classification rate for class k, as determined by Equation (1):

$$CCRk = \frac{ck}{Nk} \quad (1)$$

Where, Ck is the number of correct identifications for class k and Nk is the total number of elements in class k. Classification accuracy is then defined by Equation (2):

$$Accuracy = \frac{\sum k\ CCR_K \times N_K}{\sum K\ N_k} \quad (2)$$

Lastly, MCA is determined using Equation (3):

$$MCA = \frac{1}{k} \sum_k CCR_k \quad (3)$$

	Leaf Rust	Red Rust	Red Spider	Gray Blight	Brown Blight	Algal Leaf Spot	Sensitivity	Accuracy
Leaf Rust	**79**	**11**	**0**	**2**	**1**	**5**	**67.52%**	
Red Rust	**12**	**89**	**0**	**4**	**10**	**3**	**74.79%**	
Red Spider	**2**	**4**	**59**	**23**	**19**	**2**	**53.15%**	**90.00 %**
Gray Blight	**0**	**0**	**5**	**13**	**11**	**6**	**63.06%**	
Brown Blight	**0**	**2**	**19**	**17**	**73**	**1**	**63.48%**	
Algal Leaf Spot	**9**	**10**	**12**	**13**	**4**	**54**	**51.43%**	

Table 3.2. Error matrix showing the classification accuracy of the SVM

3.4 Materials and Methods:

This research proposes a system which is based on Support Vector Machine (SVM) classifier, since all concepts for any vision related approach for image classification remains the same. A digital camera is used to capture a tea leaf, and image processing techniques are applied to these images to extract various features. The useful features are used to train the SVM which performs the classification as shown in Figure 1.

3.4.0 Image Acquisition:

A Nikon D3300 and D3100 was used to take pictures of many tea leaves from Bangladesh Tea Research Institute (BTRI) located in Sri Mongol. A large numbers of image samples were collected, some which were affected by diseases (brown blight, and algal) and others that were unaffected or otherwise healthy. Figure 2 shows a representation of these three types of leaves.

Figure 3.2: Leaf Rust

Figure 3.3: Red Rust

3.4.1 Image preprocessing:

The aim of pre-processing is to improve the image data and suppresses unwanted distortion or enhances some image features which would be important for further processing. We enhance the image so that we can get the better features as shown in Figure 3.

Figure 3.4: Contrast Enhancement

3.4.2 Image processing:

This consists of two steps: Image Normalization and Color space Conversion. In image normalization, the captured image has been converted into the normalized image (i.e. the image size have been made constant). Then the color space of the normalized image has been converted into grayscale image as shown in Figure 4.

Figure 3.5(a): **Normal image**

Figure 3.5(b): normalized image

Figure 3.5(c): Grayscale Iamge

Figure 3.5: Image Preprocessing

3.5.0 Clustering method:

K-means clustering is used to partition the leaf image into three clusters in which one or more clusters contain the disease in case when the tea leaf is infected by more than one disease. K means clustering algorithm was developed by Macqueen (1967) and then by Hartigan and Wong (1979). The k-means clustering algorithms tries to classify objects (pixels in our case) based on a set of features into K number of classes. The classification is done by minimizing the sum of squares of distances between the objects and the corresponding cluster or class centroid (Macqueen, 1967; Hartigan and Wong, 1979).

In present experiments, the K-means clustering is set to use squared Euclidean distances.

It is observed from Fig.3.5 that cluster 3 contains infected object of early scorch disease. Furthermore, clusters 1 and 2 contain the intact parts of leaf, although they are distinct from each other. However, cluster 3 represents the black background of the leaf which can be discarded primarily. Finally, the image in (f) facilitates the segmentation procedure followed in K-Means algorithm.

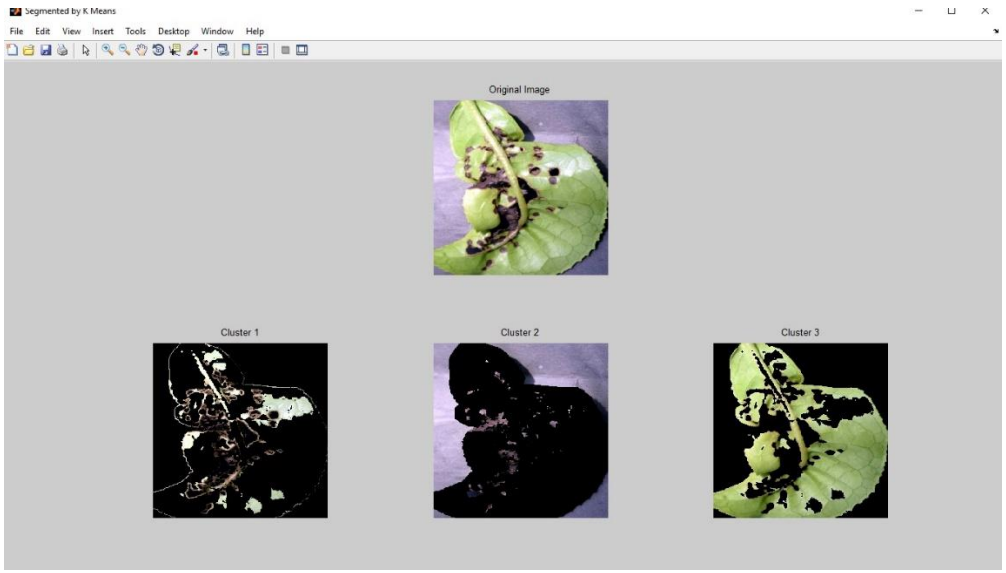

Figure 3.6: Clustered Image

3.5.1 Feature extraction:

The method followed for extracting the feature set is called the color co-occurrence method or CCM method in short. It is a method, in which both the color and texture of an image are taken into account, to arrive at unique features, which represent that image. In feature extraction method features such as color, texture, morphology and structure are used in tea leaf disease detection. Color co-occurrence method is used in which the texture and color of the image are considered. The methods used in color co-occurrence are firstly the RGB image of the leaves are converted into HIS color space representation. For generation of color co-occurrence matrix each pixel map is applied which results into three color co-occurrence matrix one for each of H, S, I.

$$X = 0.5\ \{(R-G) + (R-G)\}$$

$$Y = \sqrt{(R-G)(R-G) + (R-B)\ (G-B)}$$

$$\theta = \text{acos}(x/y)$$

$$H = \begin{cases} \theta & \text{if } B<G \\ 360-\theta & B>G \end{cases}$$

$$S = 1-3/(R+G+B) * [\min (R,G,B)]$$

$$I = (1/3)(R+G+B)$$

3.5.2 Co-occurrence methodology for texture analysis:

The image analysis technique selected for this study was the CCM method. The use of color image features in the visible light spectrum provides additional image characteristic features over the traditional gray-scale representation

The CCM methodology consists of three major mathematical processes. First, the RGB images of leaves are converted into HSI color space representation. Once this process is completed, each pixel map is used to generate a color co-occurrence matrix, resulting in three CCM matrices, one for each of the H, S and I pixel maps. Hue Saturation Intensity (HSI) space is also a popular color space because it is based on human color perception (Stone, 2001). Electromagnetic radiation in the range of wavelengths of about 400 to 700 nanometers is called visible light because the human visual system is sensitive to this range. Hue is generally related to the wavelength of a light and intensity shows the amplitude of a light. Lastly, saturation is a component that measures the colorfulness in HSI space (Stone, 2001).

The color co-occurrence texture analysis method was developed through the use of spatial gray-level dependence matrices or in short SGDM's. The gray level co-occurrence methodology is a statistical way to describe shape by statistically sampling the way certain grey-levels occur in relation to other grey-levels.

These matrices measure the probability that a pixel at one particular gray level will occur at a distinct distance and orientation from any pixel given that pixel has a second particular gray level. For a position operator p, we can define a matrix Pij that counts the number of times a pixel with grey-level i occurs at position p from a pixel with grey-level j. The SGDMs are represented by the function P(i, j, d, θ) where i represents the gray level of the location (x, y) in the image I(x, y) and j represents the gray level of the pixel at a distance d from location (x, y) at an orientation angle of θ. The reference pixel at image position (x, y) is shown as an asterix. All the neighbors from 1 to 8 are numbered in a clockwise direction. Neighbors 1 and 5 are located on the same plane at a distance of 1 and an orientation of 0 degrees

The RGB image is converted to HIS and then we calculate the feature set for H and S, we dropped the intensity (I) since it does not give extra information. However, we use GLCM function in MatLab to create gray-level co-occurrence matrix. The number of gray levels is set to 8 and the symmetric value is set to true and finally, offset is given a 0 value.

Chapter-4

4.1 Results and Analysis

In this experimental study, **1402** samples for training and another **164** samples for testing has been used in the system. Furthermore, 13 features have been taken to train the classifier as shown in Table 4.1. Nonessential features have been eliminated. Previous researches used 13 features, however; three features including homogeneity, smoothness, and IDM were determined to not show any difference between healthy, brown blight disease, and algal disease using statistical analysis. Therefore these three features have been excluded from the SVM classification. However, as can be seen from Table II and III, after elimination of these features, the accuracy has not degraded very much but the system has become faster than before. Our algorithm has been tested by processing 300 tea leaves and has been found to be approximately 1.5 minutes faster than previous researches. That averages to speed up of 300ms per leaf! There has also been the removal of manual clustering. The designing of the system is in such a way that it can select the best cluster automatically. From the graph in Figure 3.7, it is shown that as the number of training cycle's increases, the error rate decreases and after 300 training cycle the curve becomes steady, which implies that the classifier do not require more than 300 samples to become fully trained.

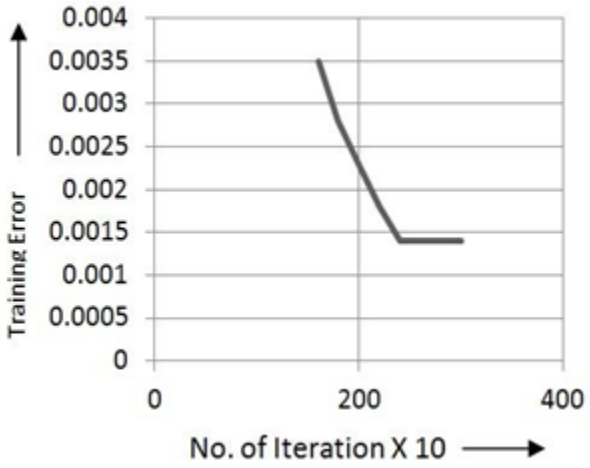

Figure 3.7: Training Error vs. Number of Iteration

Average Features of system for leaves,

Features	Brown Blight	Algal Leaf Spot	Leaf Rust	Red Rust	Red Spider
Contrast	0.1850	0.1435	0.1523	0.1190	0.1550
Correlation	1.0217	1.0309	1.0999	1.0808	1.0416
Energy	0.2029	0.2248	0.2210	0.2314	0.2429
Mean	101.9871	88.0292	90.6521	88.0929	101.9832
Std. Deviation	65.2077	67.8828	63.2132	66.3421	65.3277
Entropy	7.6262	7.4902	7.2123	7.4353	7.9862
RMS	16.7164	16.6964	15.347	16.7841	16.7874
Variance	2.6127e+03	2.4773e+03	2.6127e+03	2.4773e+03	2.4327e+03
Kurtosis	2.2952	2.4351	2.2764	2.8735	2.5652
Skewness	0.7689	0.9398	0.9178	0.7869	0.6789
IDM	255	255	255	255	255
Homogeneity	0.97001	0.9592	0.9625	0.9222	0.96001
Smoothness	1	1	1	1	1

Table 4.1: ACCURACY DATA

Image Type	Success Rate (%)
Brown Blight	90
Algal Leaf Spot	80
Leaf Rust	98
Red Rust	95
Red Spider	97

Table 4.2: THE ACCURACY OF THIS EXPERIMENTAL SYSTEM

4.2.0 Experimental Outcome: This experiment have been performed in MATLAB**.** We added some performance images of different stages while the program is been terminated.

4.2.1 Contrast Enhanced:

Figure 4.1: contrast Enhanced

4.2.2 Image clustering:

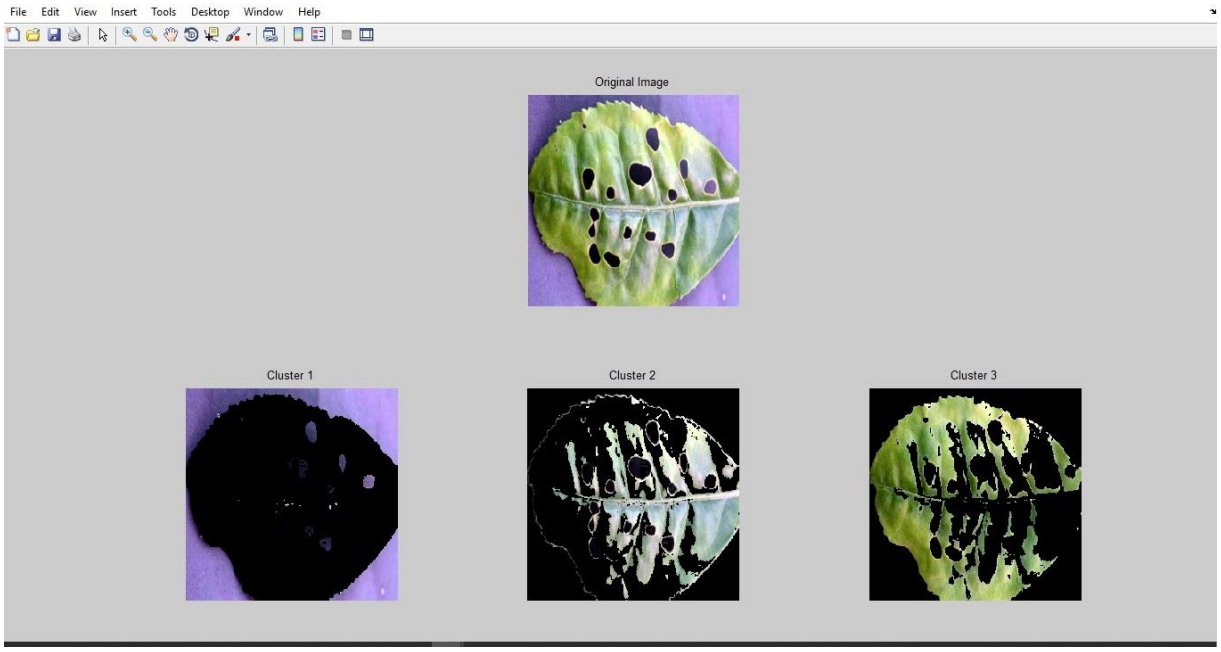

Figure 4.2: clustered Image

4.2.3 Segmented Image:

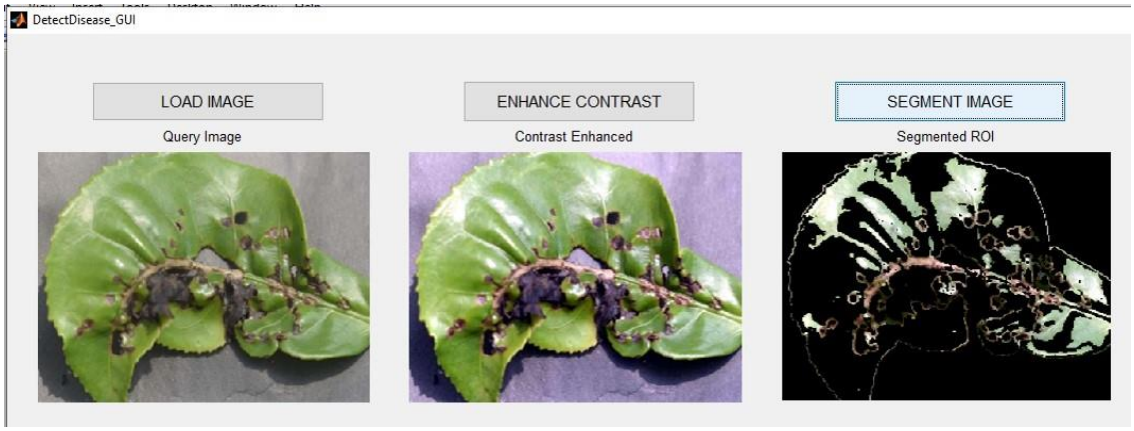

Figure 4.3: Segmented Image

4.2.4 Selection from Cluster:

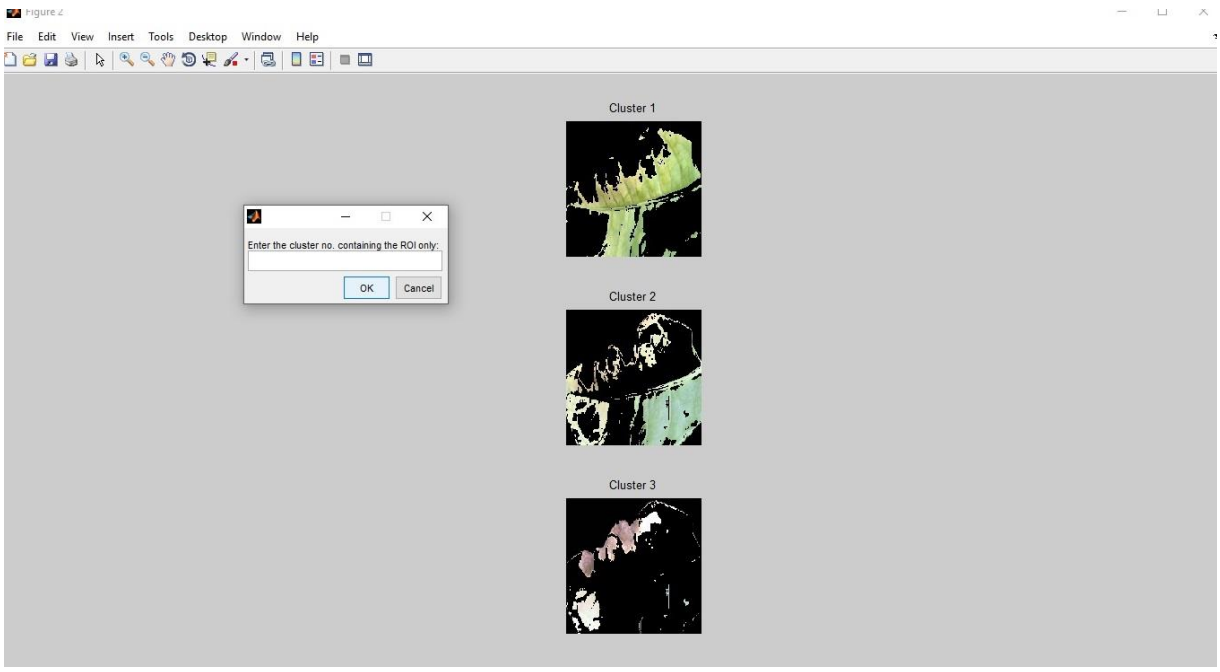

Figure 4.4: Entering the cluster no.

4.2.5Detection of Disease:

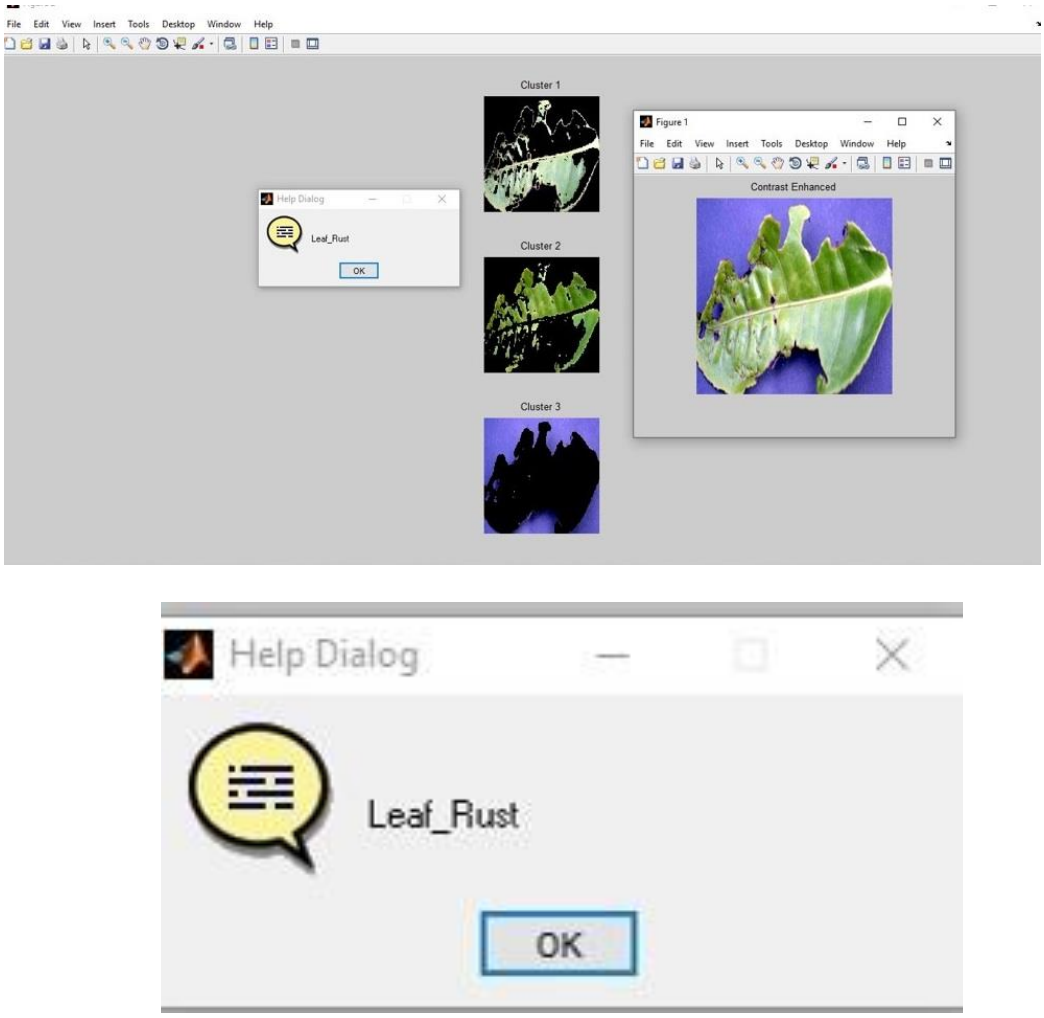

Figure 4.5: Disease Detection GUI

Chapter-5

5. Discussion:

The average features of the system for Algal Leaf Spot, Brown Blight, Grey Blight, Red Rust, Leaf Rust, Red Spider and results after conducting the experiment are shown in Table 4.1. It is seen that there is a significant difference in contrast between healthy and diseased leaf. It also seen that the correlation, standard division, skewness value is greater than healthy leaf value. But the kurtosis feature of healthy leaf value is less than diseased leaf. There is difference between healthy and diseased leaf in RMS value but there is only slight difference between the RMS values of the two types of diseased leafs. So from the table it is seen that when leaf is affected by disease then correlation, standard division, skewness value is increased and there is significant change in contrast. In the experiment the overall accuracy is 93.33% as shown in Table 4.2.

5.1 Conclusion:

The present study deals with automatic disease detection of tea leaf of using image processing techniques. It involves image acquisition, image pre-processing, image segmentation, feature extraction and classification. Development of automatic detection system using advanced computer technology such as image processing help to support the farmers in the identification of diseases at an early or initial stage and provide useful information for its control. We would like to extend our work further on more plant disease detection.

5.2 Future Works:

The future scope of this work is to improve the segmentation process and to classify the disease using different classifier in order to carry out more accurate result.

References

1. H. L. Khatibl, F. Hawels, H. Hamdi, and N. L. Mowelhi."Spectral Characteristics Curves of Rice Plants Infected by Blast," IEEE Geoscience and Remote Sensing Symposium, vol. 2, pp. 526- 528,1993.
2. H. Panda, "The Complete Book on Cultivation and Manufacture of Tea" (2nd Revised Edition) Format: Paperback ISBN: 9788178331683 Code: NI242.
3. Online:https://www.plantvillage.org/en/topics/tea/diseases_and_pests_description_uses_propagation, accessed on September 2016.
4. M. Ahmed and M.S.A. Mamun,"Distrributional Pattern and Seasonal Abundance of Major PestsofTea in Bangladesh", Tea J. Bangladesh, Vol. 41, pp. 1-10, 2012, ISSN: 0253-5483
5. A.F.M.B. Alam, "Profile of tea industry in Bangladesh" pp.1-22 (Jain NK ed. 1999), Global Advances in Tea Science. New Delhi: Aravali Books. 882pp. Chen Z and Chen X. 1989. An analysis of the world tea fauna. J. Tea Sci. Vol. 9,pp. 13–22
6. Elham Omrani, Benyamin Khoshnevisan, Shahaboddin Shamshirband, Hadi Saboohi, Nor Badrul Anuar, Mohd Hairul Nizam Md Nasir, "Potential of radial basis function-based support vector regression for apple disease detection", Measurement, Vol.55, pp.512–519, Sept-2014, (Elsevier).
7. Santanu Phadikar and Jaya Sil, "Rice Disease Identification using Pattern Recognition Techniques", Proceedings of 11th International Conference on Computer and Information Technology (ICCIT 2008), 25-27 December, 2008, Khulna, Bangladesh, pp. 420-423.
8. Haiguang Wang, Guanlin Li, Zhanhong Ma and Xiaolong Li, "Image Recognition of Plant Diseases Based on Principal Component Analysis and Neural Networks", 8th International Conference on Natural Computation (ICNC 2012), pp.246-251.
9. Dheeb Al Bashish, Malik Braik, and Sulieman Bani-Ahmad, "A Framework for Detection and Classification of Plant leaf and Stem Diseases", 2012 International conference on Signal and Image Processing, Chennai, India, pp.113-118
10. International Journal of Computer Applications (0975 – 8887) Volume 114 – No. 17, March 2015
11. Isabelle Guyon and Andr´e Elisseeff , "An Introduction to Feature Extraction", Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.
12. Y. Liu, X.Yao, "Ensemble learning via negative correlation", Neural Networks 12 (1999) 1399–1404
13. Melville and Mooney, "Creating Diverse Ensemble Classifiers to Reduce Supervision", PhD Thesis, Department of Computer Sciences, University of Texas at Austin, November 2005.
14. Hafiz T. Hassan, Muhammad U. Khalid and Kashif Imran, "Intelligent Object and Pattern Recognition using Ensembles in Back Propagation Neural Network", International Journal of Electrical & Computer Sciences (IJECS-IJENS) Vol:10 No: 06.
15. Robi Polikar, "Ensemble based systems in decision making", Article IEEE Circuits and Systems Magazines, 2006.

16. S. Arivazhagan, R. Newlin Shebiah, S. Ananthi, S. Vishnu Varthini, “Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features”, Agric Eng Int: CIGR Journal, Vol.15, pp.211-217, March 2013.
17. V. Cherkassky, M. Yunqian,"Practical selection of SVM parameters and noise estimation for SVM regression," Neural Networks, Volume 17, Issue 1, Pages 113-126, 2004.
18. E. G. Ortiz- García, S. Salcedo-Sanz, Á. M. Pérez-Bellido, J. A. Portilla Figueras, “Improving the training time of support vector regression algorithms through novel hyper-parameters search space reductions,” Neurocomputing, Volume 72, Issues 16–18, Pages 3683-3691, 2009.
19. Y. Yuan, M. Zhang, P. Luo, Z. Ghassemlooy, L. Lang, D. Wang, B. Zhang, D. Han, “SVM-based detection in visible light communications,” Optik - International Journal for Light and Electron Optics, Volume 151,Pages 55-64, 2017.
20. K. Li, L. Wang, J. J. Wu, Q. Zhang, G. Liao, L. Su, “Using GA-SVM for defect inspection of flip chips based on vibration signals,” Microelectronics Reliability, Volume 81, Pages 159-166, 2018.
21. International Journal of Computer Applications (0975 –8887) Volume 114 –No. 17, March 2015

.
.

Appendix 1

Research Schedule:

The table illustrates our research schedule; this Gantt chart helps us to complete our research in time.

	1-30 Mar	1 Apr-20 May	21May-5Jul	6Jul-3Sep	4Sep-13Oct	14Oct-13Nov	14Oov-4Dec	5Dec-8Jan	9Jan-15Feb	16Feb-15Mar
Planning	**30days**									
Literature		**50dsys**								
Matlab			**55days**							
Research				**60days**						
Statistical					**40days**					
Analysis						**30days**				
Assessment							**20days**			
Making								**35days**		
Write up									**37days**	
Completion										**28days**

Table: Research Schedule