International Journal of Engineering Research and Development
ISSN: 2278-067X, Volume 1, Issue 6 (June 2012), PP.35-42
www.ijerd.com

Classification of Biological Species Based on Leaf Architecture

Gurpreet Kaur[1], Gurpinder Kaur [2]
[1]*Student of Dept.of Computer Science Engineering, SVIET-Banur, India.*
[2] *Head of Dept.of Computer Science, SVIET-Banur, India.*

Abstract—Plants play an important role for the development of human society. The urgent situation is that due to environmental degradation, many rare plant species on the earth are still unknown and are at the margin of extinction, so it is necessary to keep record for plant protection. This research focuses on using digital image processing for the purpose of automate classification and recognition of plants based on the images of the leaves. The system consists of 4 main modules, 1) image acquisition, 2) image preprocessing, 3) image recognition and 4) display result. In the image acquisition module leaf image is captured by using digital camera. In the image preprocessing module, various image processing techniques are applied for preparing a leaf image for the features extraction process. In the image recognition module, various features are extracted from the leaf image and recognize it. In the display result module displays the recognition results. 12 kinds of leaves were taken to carry out the experiment. The accuracy of the system is 97.9 percent.

Keywords— Automatic Leaf Recognition, Classification Algorithm, Image Processing, Feature extraction, performance analysis.

I. INTRODUCTION

Plant recognition is beneficial for scientists as well as laymen. Computer aided technologies can make the process of plant recognition much easier; botanists use morphological features of plants to recognize them. These features can also be used as a basis for an automated classification tool. For example, images of leaves of different plants can be studied to determine effective algorithms that could be used in classifying different plants Different shape related features [8] were extracted from these images using image processing algorithms. Depending on these features, a statistical classification of plants was conducted. The classification scheme was then validated using a set of test images. The present paper proposes a scheme to develop the leave images recognition system that helps humans in memorizing and recognizing species of plants. The main improvement is done in the feature extraction and classification phase.

The organization of the paper is as follows: section I includes introduction, section II provides an overview of related work, section III outlines the purposed methodology section IV provides experimental results obtained and section V provides the overall conclusion and section VI provides the scope for future research.

II. PREVIOUS WORK

Studies have been done on the automation of plant classification and recognition. A large percentage of such works were based on the extraction of a single feature from the image of a plant part such as the leaf, or the flower. Some studies extracted multiple features but taking single family of plant. Some studies focused on image-based plant classification, while others worked on image-based plant recognition

Warren [1] introduced an automatic computerized system that used 10 images of each Chrysanthemum species as its input for checking the variation in the images. In this study, features such as shape, size and color of the flower, petal, and leaf were described mathematically. Different rose features were extracted and used in the recognition scheme for pattern recognition. The study, however, was restricted to Chrysanthemum species only. Miao *et al*., [2] developed an evidence-theory-based rose plant classification taking different features of roses.

In another study by Heymans *et al*.,[3], a back-propagating neural network approach was used to distinguish different leaves of Opuntia species. Again, the study was limited only to the variety of the Opuntia family.

X. F. Wang *et al* [4] introduced a MMC hyper sphere classifier for doing plant classification based on certain leaf features. First, image segmentation was done on the leaf images; then eight geometric features such as rectangularity, circularity, eccentricity, and seven moment invariants were extracted for classification. Finally, these shape features was addressed using a hyper sphere classifier. For the experimental results, he successfully classified 20 classes of plant leaves and the resulting recognition rate is up to 92.2 percent.

Nam *et al*., [5] studied a shape-based leaf image retrieval system .He introduced two novel approaches, namely, an improved maximum power point (MPP) algorithm and a revised dynamic matching method. The shape of the leaf is defined using a region-based shape representation technique. Image matching was done on the basis of the distances between the points.

III. METHODOLOGY

This section will explain the major activities involved to achieve the objectives and solve the background problems.

The structure chart of the system

The system structure chart consist four main modules, which are 1) image acquisition module, 2) image preprocessing module, 3) image recognition module, and 4) display result module (as shown in Figure 1).

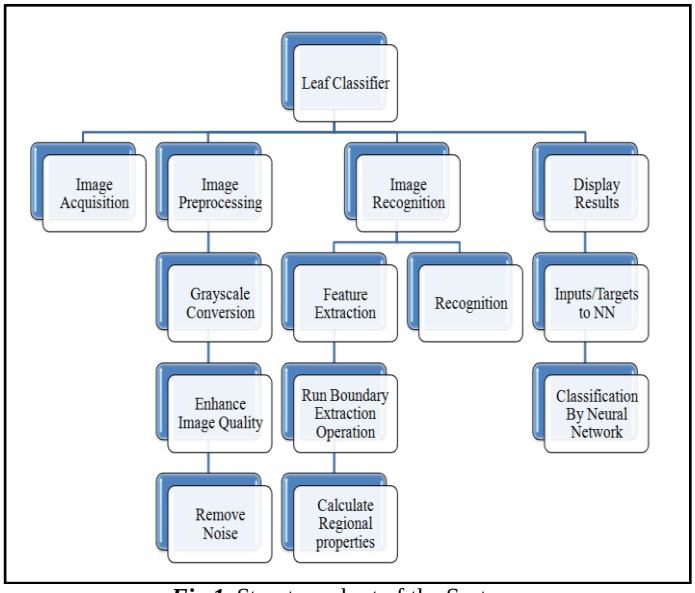


Fig 1. Structure chart of the System.

Each module has the following description.

A .Image Acquisition

Leaf images are collected from variety of plants with a digital camera. In this research, 12 species of different plants are taken. Each species includes 12 samples of leaf images. These leaf images come in with different of size, shapes and class.

B. Image Pre-processing

Pre-processing [7] usually contains a series of sequential operations which includes prescribing the image size, conversion of gray-scale images to binary images (monochrome) file and modifying the scaling and rotation factors of the image. The images pre-processing steps used in this research are as follows:

a) The images are reduced to 200 x 200 pixel dimensions in order to decrease computation load.

b) The grey-scale images are converted to binary image. The binary images are often produced by thresholding a grays-scale or color image. Then, the image is saved using the *.jpg* format. The weighted averaging method is used in order to separate an object in the image from the background. The formula used to convert RGB value of a pixel into its grayscale value:

Gray = 0.2989 *R + 0.5870 * G + 0.1140 * B

Where R, G, B corresponds to the color of the pixel, respectively.

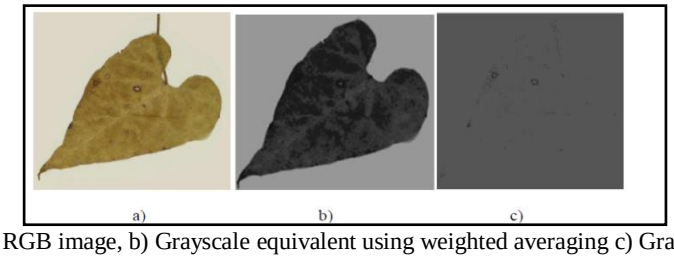


Fig 2: a) An arbitrary RGB image, b) Grayscale equivalent using weighted averaging c) Grayscale equivalent using averaging

C. Image enhancement

The aim of image enhancement [9] is to improve the interpretability or perception of information in images for human viewers, or to provide `better' input for other automated image processing techniques. These enhancement operations are performed in order to modify the image brightness, contrast or the distribution of the grey levels. Examples of enhancement operations:-Removing blurring and noise, increasing contrast, and revealing details.

D. Removing Noise from Images

Digital images are prone to a variety of types of noise. Noise is the result of errors in the image acquisition process that result in pixel values that do not reflect the true intensities of the real scene. This sub-module clears noise in the leaf picture for easily to extract their features. The system uses the erosion and dilation technique [12] to clear the noises.

E. Feature Extraction

The next phase in the plant leaf identification is the feature extraction phase [6]. The main advantage of this stage is that it removes redundancy from the image and the leaf images are represented by a set of numerical features. The classifier used these features to classify the data. The Texture Feature Extraction is one of the main subjects in pattern recognition. We used GLCM for texture feature extraction. The GLCM functions characterize the texture of an image by calculating how pairs of pixel with specific values and in a specified spatial relationship occur in an image.

Pseudo Code for feature extraction species

```
For each Image In Database
Read each Image
Convert each Image in gray
GLCM
Extract texture features
Extract shape features
Save feature textfile.text
End
End.
```

The following features are used for doing leaves classification.

- ***Extraction of Isoperimetric Quotient*:** In order to extract the isoperimetric quotient, extraction of the area and the perimeter of leaves are required. Given a binary leaf image, the area of the given leaf is calculated as the number of pixels in the foreground. Canny edge detector provides the edge image for the foreground region and perimeter is calculated as the number of pixels in the boundary of the foreground region.
- ***Extraction of Eccentricity*:** Region-based method [7] is used to estimate the best fitting ellipse for the extraction of eccentricity. Fig. illustrates a best fitting ellipse to a polygon. It also shows the major axis, *Ma* and the minor axis, *Mi* of the best fitting ellipse.

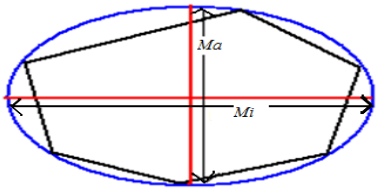


Fig.3: The best fitting ellipse to a polygon with its major and minor axes

- ***Extraction of Aspect Ratio*:** For a rectangular shape, aspect ratio is defined as the ratio of its breadth by length. For leaves having irregular shapes, the aspect ratio [8] is defined as the ratio of the minor axis to the major axis of the best fitting ellipse.
- ***Leaf Area*:** The value of leaf area is evaluated by counting the number of pixels having binary value 1 on smoothed leaf image. It is denoted as A.
- ***Leaf Perimeter*:** is evaluated by counting the number of pixels containing leaf margin
- ***Area ratio of perimeter*:** Ratio of Area to perimeter, It is defined as the ratio of leaf Area A and leaf perimeter P, is calculated by A/P.
- ***Solidity:*** It is defined as the ratio of areole area to convex area.
- ***Length of the major and minor axes*:** Area of the ellipse that just encloses the region.
- ***Upper and Lower triangle area of the leaf.*** The leaf is divided into two half by a horizontal line by the system, and then the system finds an upper- triangle area by dividing the upper leaf area by the upper-half image area. Similarly, a lower triangle area is calculated by dividing the lower leaf area by the lower-half image area.
- ***Boundary feature*:** Sobel edge detection algorithm [9] is applied by the system with threshold values 0.1 and 0.5, to find the leaf boundary ,then the system counts white pixels on each threshold value.

F. Pattern Recognition

Pattern recognition phase perform the validation and evaluation of the performance of the classification scheme.

G. Implementation of Leaf Classifier

The classification is done by using MATLAB software package. The single MLPNN [10] with Levenberg-Marquardt back-propagation algorithm is used to classify the plant species. The single hidden layer is chosen. In hidden

layer and output, the sigmoid activation function is used. The features computed are used for classification. In the present study, we have divided the features into three equal halves for training, validation and testing.

Two algorithms are used for classification and also known as classifiers

1) *Artificial Neural Network* (ANN)*:* Artificial Neural Network (ANN) [10] has been the successfully used classifier in numerous fields. So, it is of interest to use it for leaf analysis. It can be modeled on a human brain. The basic processing unit of brain is neuron which works identically in ANN. The typical structure of neural network is shown in Fig.4 which consists of m input neurons in general and n hidden neurons with single hidden layer. The output layer has only three neurons. The network is called as fully connected network when all the neurons are connected with the adjacent neurons.

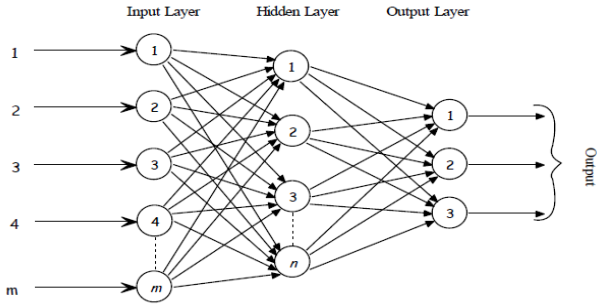


Fig 4: General Structure of a Neural Network

2) *Multilayer Perceptron Neural Network* (MLPNN)*:* The ability of classifying the non-linearly separable classes in a supervised manner enables it as a mostly used neural network for classification purpose. It uses the error correction rule known as back propagation algorithm.

Back Propagation Algorithm

The Back Propagation Algorithm [10] is explained in the following steps:

Step: 1 Initialization

Set all the weights and biases to small real random values.

Step: 2 Presentation of input and desired output

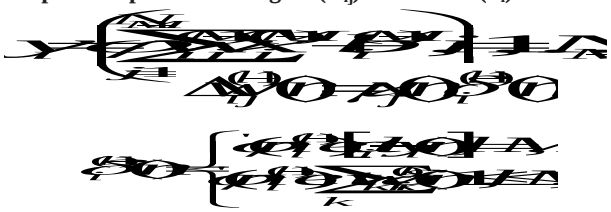
Present the input vector $x(1), x(2), \ldots, x(N)$ and corresponding desired response $d(1), d(2), \ldots, d(N)$, one pair at a time, where N is the number of training patterns.

Step: 3 Calculation of actual outputs

Equation (1) below is used to calculate the output signals $y_1, y_2 \ldots\ldots y_{NM}$

Where w_{ij} are the weights and b_i are the biases.

Step: 4 Adaptation of weights (w_{ij}) and biases (b_i)



in which $x_j(n)$ = output of node j at iteration n, l is layer, k is the number of output nodes of neural network, M is output layer, φ is activation function. The learning rate is represented by μ.

Back propagation Training Algorithm*: Levenberg-Marquardt (trainlm)*

Back propagation algorithm utilizes the Levenberg-Marquardt algorithm [11] for training of the network. The 'trainlm' is a network training function that updates weight and bias values according to Levenberg-Marquardt optimization. The Levenberg-Marquardt consists basically in solving $(H + \lambda I)\,\delta = g$ with different λ values until the sum of squared error decreases. So, each learning iteration (epoch) will consist of the following basic steps:

1. Compute the Jacobian (by using finite differences or the chain rule)
2. Compute the error gradient
   $g = J^tE$
3. Approximate the Hessian using the cross product Jacobian
   $H = J^tJ$
4. Solve $(H + \lambda I)\delta = g$ to find δ
5. Update the network weights w using δ
6. Recalculate the sum of squared errors using the updated weights
7. If the sum of squared errors has not decreased, discard the new weights, increase λ using ν and go to step 4.
8. Else decrease λ using ν and stop.

Variations of the algorithm may include different values for ν, one for decreasing λ and other for increasing it.

IV. RESULTS

This section presents the evaluation of designed classifier by means of different performance indices:

A. Confusion Matrix

The performance of classifier is analyzed using confusion matrix [13] which is also known as table of confusion. It displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The confusion matrix shown in table 1 gives the summary of plant species classification results for the 12 classes.

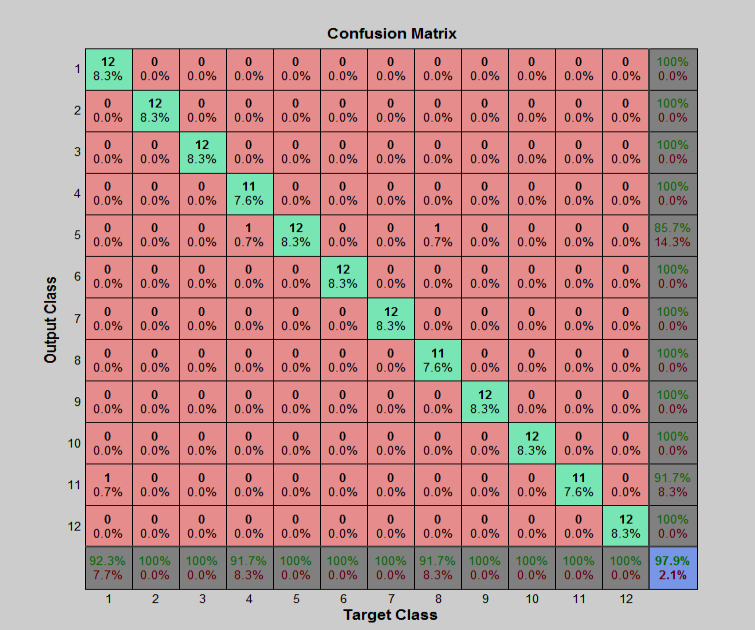


Table 1 Overall Confusion Matrix

In Table 1 we can see that when we consider overall data set, then the accuracy was improved and reaches to 97.9% which is a good amount. In this as we can see classes 1,2,3,4 was correctly classified, class 5 was 1 time misclassified as class 4 by 0.7% and 1 time misclassified as class 8 by 0.7%. Classes 6, 7, 8, 9, 10, 12 were correctly classified. Class 11 is 1 time misclassified as class 1 by 0.7%.

Table 2 Approximate overall Classifier Accuracy

Overall accuracy	97.9%

B. Receiver Operating Characteristics

Receiver Operating Characteristics [15] is a metric used to check the quality of classifiers. For each class of a classifier, threshold values across the interval [0, 1] are applied to outputs. For each threshold, two values are calculated, the True Positive Ratio (the number of outputs greater or equal to the threshold, divided by the number of one targets), and the False Positive Ratio (the number of outputs greater than the threshold, divided by the number of zero targets).Our aim is to minimize false positive rate and increase true positive rate.

ROC graphs for training, validation, testing and overall phases have been shown in Fig 5, Fig 6, Fig 7 and Fig 8. In these graphs we have to see if the data trend analysis (best fit line) is proper or not. It is determined by coefficient of determinant and also called coefficient of goodness (R^2) should be 100 % .

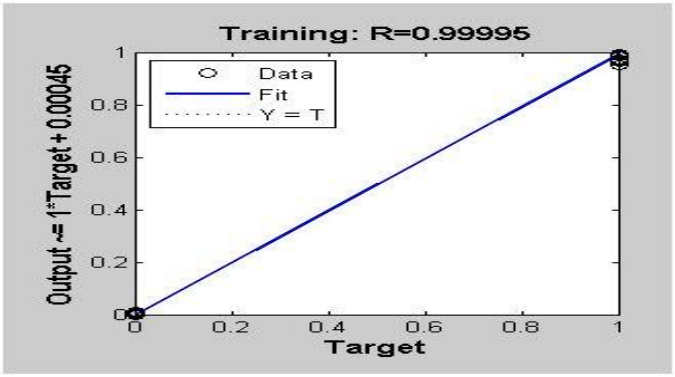


Fig 5: Training Phase ROC curve

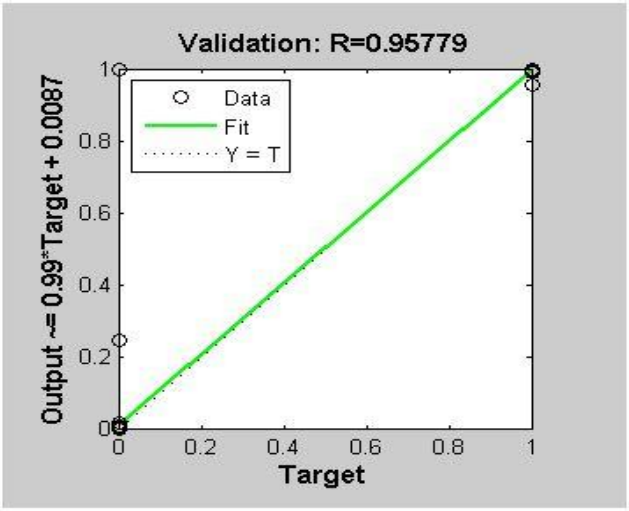


Fig 6: Validation Phase ROC curve

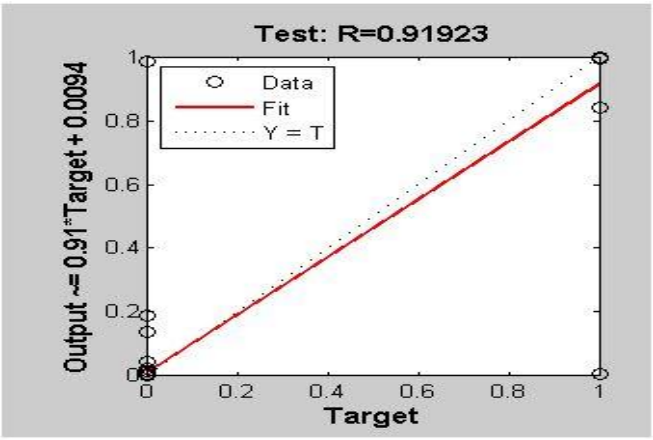


Fig 7: Testing Phase ROC curve

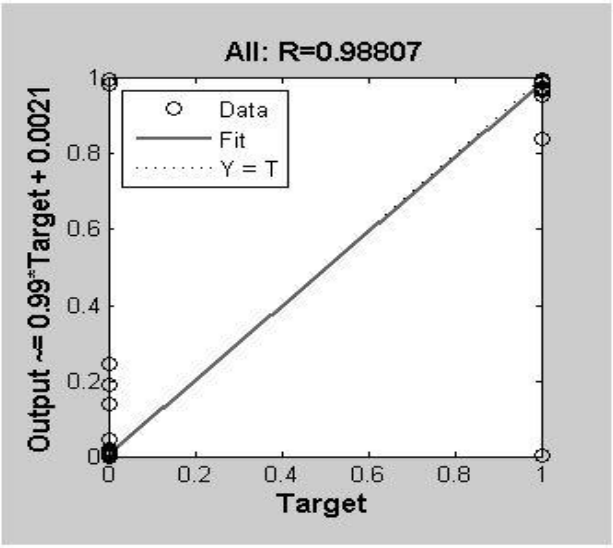


Fig 8: Overall ROC curve

In fig 5, fig 6 we can see the value of R is 0.99995 and 0.95779 which means data has been fitted and classified in a correct manner. We can see that result of validation phase lies on the random guess line (the diagonal line), so the accuracy is 100%.

In Fig 7 we can see that as true positive rate is increasing the false positive rate is also increasing at regular interval so that the coefficient of goodness i.e. R^2 is not 100%.

In Fig 8 it is shown that when all samples are considered after all three training, validation and testing phases, the value of R is 0.98807 which means data has been fitted and classified in a correct manner which is the ideal condition for goodness of classifier

C. Mean Square Error:

MSE is the average squared difference between outputs and targets. Lower values of (MSE) indicate better performance of the network and zero means no error [14].

The performance graph is shown in Fig 9

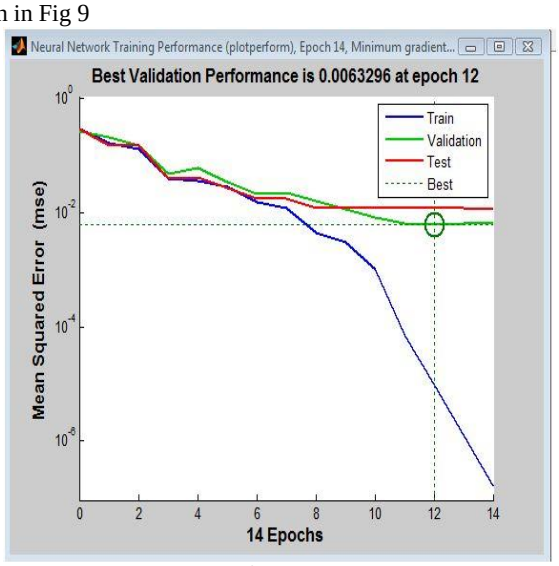


Fig 9 Performance Graph

From performance chart we can see that by simulating the conditions in each epoch we can see that there is not much variation in mean square error. It is more or less steady graph but as the epoch increases in each phase shown by red, green and blue lines the value reaches to best validation value of 0.0063296 which is very close to zero in ideal situation.

V. CONCLUSION

Our objective in research have been to develop a leaf classifier which works on the principle of extracting information based on its architecture using image processing to classify correctly in the plant kingdom. The idea is to develop a classifier which has less computation and complexity and is able to classify and learn the leaf architecture in correct manner.

We have used neural network to solve our classification problem in which we have used LM algorithm for training our classifier, a sampling algorithm called random divide and rule, an algorithm called maxmin for normalizing the inputs and targets they conducted multiple designs to identify best suitable permutation of hidden layer, inputs and output layers. 12 kinds of leaves were taken to carry out the experiment. The accuracy of the system is 97.9 percent.

VI. FUTURE SCOPE

The research can be improved by exploding a larger training set using other leaf shapes. We can explore more algorithms and techniques for the feature extraction and classification of biological species to further improve the accuracy of the identification system. We can further improve the system by reducing the complexity. The main objective could be to find the best algorithms which optimize the performance and complexity. The accuracy of classifier can also be enhanced by using more and equal number of training patterns.

VII. ACKNOWLEDGEMENT

I would like to thanks my parents and my friends for showing trust and giving their support which helps me successful in completing this research process.

REFERENCES

[1]. D. Warren, "Automated leaf shape description for variety testing in chrysanthemums," in *Proc. 6th Int. Conf. Image Process. and Its Applicat.*, Duplin, Ireland, 1997.

[2]. Z. Miao *et al*., "An oopr-based rose variety recognition system," *Engineering Applications of Artificial Intelligence*, vol. 19, issue 5, Amsterdam, Elsevier, 2006, pp. 78-101.

[3]. B. C. Heymans *et al*., "A neural network for Opuntia leaf-form recognition,"*IJCNN*, vol. 3, pp. 2116-2121, 1991

[4]. X. F. Wang et al., "Recognition of leaf images based on shape features using a hypersphere classifier," ICIC, vol. 3644/2005, pp. 87-96, 2005

[5]. Y. Nam *et al*., "Elis: An efficient leaf image retrieval system", in *Proc. Advances in Pattern Recognition Int. Conf.*, Kolkata, India, 2005.
[6]. Jyotismita Chaki, Ranjan Parekh (2011)" Plant Leaf Recognition using Shape based Features and Network classifiers" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2.
[7]. Biva shrestha (2000) " *Classification of plants using images of their leaves*" ICPR, vol. 2, pp. 2507-2510.
[8]. Stephen Gang Wu, Forrest Sheng Bao, Eric You Xu, Yu-Xuan Wang, Yi-Fan Chang and Qiao-Liang Xiang (2007)"*A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network*",arXiv:0707.4289v1 [cs.AI] .
[9]. R. E. Gonzalez and R. E. Woods. Digital Image Processing. Addison-Wesley, 1993.
[10]. N.K Bose.neural network fundamentals with graphs,algorithms and applications, p.Liang, 1994.
[11]. L. Tang, L. Tian, B. L. Steward (2003)"*Classification of broadleaf and grass weeds using Gabor wavelets and an artificial neural network*" Transactions of the ASAE Vol. 46(4): 1247–1254 _ 2003 American Society of Agricultural Engineers ISSN 0001–2351.
[12]. T. Satioh and T. Kaneko (2000), "*Automatic Recognition of Wild Flowers*", ICPR, vol.2, pp. 2507-2510.
[13]. http://en.wikipedia.org/wiki/confusionmatrix
[14]. http://en.wikipedia.org/wiki/MSE
[15]. http://en.wikipedia.org/wiki/ROC

AUTHORS BIOGRAPHY

Ms. Gurpreet kaur was born in the small village of khanna, punjab. After finishing high school in khanna, I moved to BBSBEC fatehgarh sahib, Punjab to pursue a Bachelor's degree in information technology. After graduating with a Bachelor of Science Degree in Fatehgarh sahib from BBSBEC Engineering College-Fatehgarh sahib in 2010,I started my M.tech in computer science and engineering from SVIET,banur(Punjab).