Apple Scab and Marsonina Coronaria Diseases Detection in Apple Leaves Using Machine Learning

1Swati Singh and 2Sheifali Gupta
1Department of Electronics and Electrical Engineering, Chitkara University.
swatii.singh@chitkara.edu.in
2Department of Electronics and Electrical Engineering, Chitkara University.
sheifali.gupta@chitkara.edu.in

Abstract

Agriculture plays an indispensable role in the development of the country especially in the growing country like India where most of the peoples’ revenue is generated from agriculture. Disease affected crops leads to the loss of crop productivity. Therefore, leaf disease prediction in apple cultivation is of considerable importance to overcome these problems. The proposed work intends to predict apple scab and marsonina coronaria apple leave diseases using four different classification algorithms i.e. support vector machine (SVM), K nearest neighbor (KNN), classification decision and regression tree and Naïve Bayes. From the simulation results, it was shown that KNN gives an accuracy of 99.4% in classifying apple scab and marsonina coronaria apple leave diseases as compared to the other classifiers.

Key Words: SVM, KNN, apple disease, marsonina coronaria, apple scab, K means clustering.
1. Introduction

India is ranked number two in the production of fruits. Sixty percent of the population depends on agriculture for employment. The major problem for the farmers is to put an end to the rise of pests that affect crop yields [1]. Pests cause scattered disruption of diseases, which advance to scarcity and food deficit [2]. Moreover the early revelation of pests on leaves is essential as a very small number of leaves with diseases can extend the contagion to the entire plant. The methodology used for fruit diseases recognition is naked eye perception by farmers. Diagnosis of the disease in majority cases is performed visually through naked eye by human [3]. Agriculturists who are skilled may be efficient in diseases recognition, but they also have some disadvantage like agriculturists may lose focus, thus decreasing their accuracy, to maintain quality, training may need to be repeated, agriculturists are expensive. Besides these disadvantages the agriculturist may find it challenging task if the diagnosis is done for extremely large areas.

The agricultural goods’ quality has been hampered because of diseases on plants especially on leaves. [4]. If proper investigation and early solution of the presence of diseases on leaves is not given then the quantity and quality of the crop yield can be reduced [5]. Therefore there is a demand to detect these diseases in initial stage and propose remedies so that intense damage can be bypassed to boost the crop production [6]. By the use of image processing in combination to pattern recognition, many of these issues can be solved. Fig.1 represents an apple leaf affected with disease.

![Figure 1: Apple Leaf Disease](image)

This paper reports on common disorders of apple tree leaves (Apple scab and Marsonina Coronaria), which are classified by training and implementing machine learning algorithms to images of diseased apple tree leaves.

2. Related Work

To help with the detection of diseases in plants, there are a few algorithms implemented although not limited or pertaining to apple leaves. In paper [7] apple rot, apple scab and apple blotch diseases have been taken. K-means Clustering have been used for image segmentation. Features such as color coherence vector (CCV), local binary pattern (LBP), Global color Histogram (GCH) and complete local binary pattern (CLBP) are extracted. In paper [8] different methods like Local Binary Pattern, Color Invariant, Color histogram are used for the feature extraction from images and further classification is carried out with the help of neural network. The paper [9] aimed to detect
healthy sugar beet leaves from diseased ones. Authors used Support Vector Machines which resulted in classification accuracies up to 97%. Notably, to gain the significant feature selection and spectral images this technique requires specialized hardware. In paper [10] five samples are considered. 3 samples of grapes leaves and 2 samples of apples leaves are picked up for detection. As a result 90% accuracy has been achieved using neural network algorithms. Paper [11] tried to categorize six different diseases from leaf images. Features were manually extracted from the image which defined its 10 texture features. SVM was for applied for classification which showed accuracies between 83% to 94%. Paper [12] discusses the different plant diseases detection techniques. A review of all the different classification and detection techniques were done and it was concluded for that classification neural networks and SVM were found competitive with each other and for segmentation the simplest and accurate approach is K means clustering.

3. Apple Leaves Diseases and Symptoms

Proposed work includes number of attacks on apple leaves. Some of the diseased apple leaves and there symptoms as shown below:

**Apple Scab**

This disease is usually seen on leaves and fruits. Apple scab is expanded through fungus producing in former apple leaves on the ground and expands during warm spring weather to contaminate the new year’s growth. The symptoms are velvety olive-green spots with unclear margin that appear on the leaves. Later, the lesions enlarge, and become darker with more distinct margins and heavily infected leaves become distorted and drop early in the summer. The main effects includes mid season defoliation, weakening of the tree, failure of fruit bud formation, reduction in fruit yield, devaluation in fruit quality and increased expense due to fungicide spraying. Figure 2 represents an apple affected by apple scab disease.

![Figure 2](image)

(a) Leaf Effected with Apple Scab Disease (b) Apple Affected with Apple Scab Disease

**Marsonina Coronaria**

The disease symptoms is visible in form of green circular spots that are dark in color and are on upper surface of leaf rendering rise to 5-10 mm brown leaf patches which turn dark brown with time. When it matures, it develops on the
bottom surface of the leaf. Small black patches are seen on the upper side of the leaf. When these are numerous, they combine and produce enormous dark brown blotches and the nearby area becomes yellow. The effects include premature defoliation and reduction in quantity and quality of apples. It also effects flower formation in autumn, leading to depletions in fruit set in the following season. Figure 3 shows an apple and a leaf that has been affected by marsonina coronaria disease.

![Figure 3](image1.png)

(a) Leaf Effected with Marsonina Coronaria Disease  
(b) Apple Effected with Marsonina Coronaria Disease

**Alternaria Leaf Spot/Blight**

Patches on the leaves occur in early summer or late spring. The initial symptoms include spots with 1/8 to 1/4 inch diameter, brown and border. On maturity, the patches turn tan to ash gray. Some of the spots go through more enlargements, and becomes unevenly shaped. Heavily infected apple leaves often fall off, resulting in defoliation. Fruit contamination results in small, dark raised pimple-like lesions. Figure 4 shows leaf and apple fruit that has been affected by alternaria leaf spot disease.

![Figure 4](image2.png)

(a) Apple Leaf Effected with Alternaria Disease  
(b) Apple Affected with Alternaria Disease

**4. Methodology**

This paper aims to classify two apple leaf diseases i.e. apple scab and marsonina coronaria using Matlab. Total of 900 images were taken out of which 350 were of Marsonina, 350 of Apple Scab diseases and 200 of apple leaves without any diseases. The whole dataset have been divided into train data and test data. The
The foremost step was to enhance the image using brightness preserving dynamic fuzzy histogram equalization. This step was followed by extraction of some texture and color features from the enhanced image. Lastly the apple leaf diseases were classified using various classifiers and the accuracies of each of them were tested out of which K Nearest Neighbor proved to be a better approach. Figure 5 illustrates the flow of proposed methodology.

Steps of proposed methodology: The main steps for methodology are as follows:

Capture Apple Leaf Image

Apple leaf images were collected in different regions of Himachal Pradesh and Uttarakhand. Total of 900 images were taken out of which 350 was of Marsonina Coronaria, 350 of Apple Scab diseases and 200 of apple leaves without any diseases. Figure 6 shows different apple leaves with diseases that were captured using a digital camera.
Enhancement Using Brightness Preserving Dynamic Fuzzy Histogram Equalization (BPDFHE)

The image quality has been enhanced using BPDFHE. Here the fuzzy histogram is computed with a suitable membership function. Then the dynamic equalization is done on the histogram partitioned image. After that image brightness normalization is done to preserve its brightness. The detail of the method is given in detail in paper [13]. Fig 7 shows the input image and the output of the enhanced image using BPDFHE.

![Image](https://via.placeholder.com/150)

(a) Input Image, (b) Image obtained after BPDFHE Enhancement

Excess Green and Background Part Removal

Any leaf image is having dominating part of green pixels, so before ROI extraction the background with green area should be removed. For this three components of the enhanced RGB image (I) are extracted i.e. red green and blue part with the help of following algorithm. Figure 8 shows algorithm for Red, Green and Blue component extraction of the enhanced image.

![Algorithm](https://via.placeholder.com/150)

Figure 8: Algorithm for Separating the Red, Green and Blue Part of the Enhanced Image

For green background removal, the implemented algorithm is shown in figure 9:

![Algorithm](https://via.placeholder.com/150)

Figure 9: Algorithm for the Removal of Excess Green Part

For white or black background removal, the algorithm is shown in figure 10:
Here rows and columns are the size of original gray level image. The threshold value used here is 10 which is chosen by hit and trial method. Using these two algorithms, green, white and black background of the image is removed so that results of K means clustering can be more favorable. Figure 11(a) shows the three types of images having green, white and black background which is removed by using the proposed algorithm as shown in figure 11(b).

**ROI Extraction Using k Means Clustering**

After removal of background, k-means clustering algorithm is implemented on the background removed images further to classify it into different clusters. The proposed algorithm gives the better results which was unable to compute using k-means algorithm alone. For this, background removed RGB image is firstly converted to \(L^a*b^c\) space image, where \(L\) is a luminous layer, \(a\) is a chromatic layer and \(b\) is another chromatic layer which will help to separate different parts of image. Now k means clustering is applied to cluster the different parts of the image into two different clusters, one being the diseased part and other the non diseased part.

![Figure 10: Algorithm for the Removal of Black or White Background](image)

![Figure 11: (a) Enhanced Image, (b) Excess Green, Black and White Background Removed Image](image)
Figure 12 (a) shows the original enhanced images, Figure 12(b) shows the cluster obtained after using K means clustering algorithm when used alone i.e. excess background is not removed before applying the algorithm for segmentation whereas figure 12(c) shows the desired cluster i.e. segmented region of interest (ROI) using the proposed algorithm. From figure 12(b) it is clear that for images with black and white background, k means clustering gives the desired cluster but for images having green background, it does not segment the ROI properly. Therefore, only K means clustering algorithm cannot be used for segmentation for live images, the excess green part and the background needs to be removed first. Hence the proposed algorithm for segmentation gives better results as compared to using only K means clustering.

Figure 12: (a) Original Images (b) Segmented Cluster by Using K Means Clustering Alone (b) Desired Cluster Using Proposed Algorithm

**Feature Extraction**

For the feature extraction, region of interest (ROI) will be considered for further processing. After the segmentation step, the features will be extracted from ROI to elaborate the infected region. Many techniques and algorithms have been proposed for defining texture of the image [14]. The methods of texture analysis are categorized into four classes: structural, statistical, transform-based, and model-based. Definitions and details of such techniques can be found in [15]. In the present paper, texture features i.e. gray level co-occurrence matrix (GLCM) features and statistical features are taken into consideration. GLCM features take spatial relationship of pixels into consideration. It calculates such pairs of
pixels having a specified relationship between the pixels. In this, features like contrast, co-relation, energy, entropy, smoothness and homogeneity were calculated. Gray level co-occurrence matrix (GLCM) uses image texture to derive acquires features that represent the image [16]. In statistical features, skewness, standard deviation, kurtosis and variance values are computed. Table 1 shows the values of extracted GLCM and statistical features of figure 12(c).

Table 1: Feature Extraction Results

<table>
<thead>
<tr>
<th>Features</th>
<th>GLCM</th>
<th>Statistical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Figure 12(c)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Image 1</td>
<td>Image 2</td>
</tr>
<tr>
<td>GLCM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrast</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>Co-relation</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>Energy</td>
<td>0.99</td>
<td>0.86</td>
</tr>
<tr>
<td>Entropy</td>
<td>0.07</td>
<td>0.86</td>
</tr>
<tr>
<td>Smoothness</td>
<td>0.99</td>
<td>1</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Statistical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>265.10</td>
<td>13.71</td>
</tr>
<tr>
<td>Variance</td>
<td>0.32</td>
<td>2.80</td>
</tr>
</tbody>
</table>

5. Classification Using Different Machine Learning Algorithms

Classification is the most predominant task for disease prediction, data classification and many more. It is the process which is implemented on features extracted using any feature extraction method. With those features, machine is trained and model gets ready to perform real time prediction. Classification can be performed using four different methods i.e. Support Vector Machine, K nearest neighbors, Naives Bayes, classification decision tree etc. However decision tree can be used to predict both classification and regression which is called classification and regression tree (CART). The following section discusses the different classifiers that have been implemented to identify disease. Following are the different classifiers used:

**Support Vector Machine (SVM)**

Support vector machine is a new promising, non-parametric and non-linear statistical classification technique. It finds its wide range of application in medical diagnostic, optical character recognition, disease prediction, disease detection and many more. During training of SVM model, kernel function is passed as a parameter which helps to do certain calculation faster and decreases time. If one has few points in a high dimensional space, linear SVM is a better option [17]. Classification of Support vector machinery (SVM) results in statistical learning theories.
K Nearest Neighbor

It is one of the simplest classification algorithms. KNN has some of the features which make it easy to interpret output and calculation is less. It is a non-parametric learning algorithm i.e. it does not use any generalization and its training phase is very fast. It makes use of distance of new data to every cluster and depending upon distance it classifies new data into a particular cluster. The basic convention behind KNN classifier is to gather the K previously defined number of training samples that are nearest in the distance to new point and estimate the label from these. The commonly considered distance measure is Euclidean distance [18]. It is also acknowledged as an instance based learning algorithm as it saves the records of training data in multidimensional space. For a particular value of K and each new data point, Euclidean distances are recalculated and the target class is predicted [19][20].

KNN in Matlab was implemented using different distance calculating methods like Minkowsk, Euclidean, Mahalanobis. It is very important in KNN to have a good choice of number of clusters. In the proposed work two clusters were chosen for classification.[21][22]

Naive Bayes

These classify data based on probability and Bayes theorem and are also called probabilistic classifier. One of the wide applications is automatic medical diagnosis. It uses the following formula:

\[ P(x|y) = \frac{1}{2\pi \sigma^2} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right) \]  

Where \( P(x|y) \) represents the probability, Naive Bayes approach provides probability of a real time data pertaining to a particular class. The highest probability class is considered as the most probable class. It is scalable and fast algorithm used for both multiclass and binary classification which can be trained on a small dataset.[23][24]

Decision Tree

This approach is used in statistics, data mining and machine learning. Decision tree categorize data points. Data is passed to different nodes of the tree based on some conditions and data is assigned to respective nodes. Top node is called root whereas nodes which cannot be classified are called leaf nodes. Further at each node probability of each data point is calculated and it gives an impression of its occurrence at a particular node.

Decision tree has many advantages and some of them are shared below. Decision tree helps in selecting features automatically as top few nodes are essential main features. Even missing values will not prevent splitting data for building trees. The tree performance is also not affected by the non-linear relationships and is very easy to interpret and explain.[25]
6. Results and Discussions

The proposed background removal algorithm combined with K means clustering for apple leaf disease detection approach is simulated on MATLAB 2016 running on Windows 7. The datasets used for investigation were created based on real sample images of diseased apple leaves. This dataset is collected from different farms in Himachal Pradesh and Uttarakhand region in May and June. These images were captured using camera with 12 mega pixel resolution. The storage format of the images is jpeg format. A Graphical User Interface (GUI) is also designed so that the detection of apple leaf diseases can be made user friendly.

Figure 13 depicts the graphical user interface (GUI) that was designed to detect the apple leaf diseases. In this, seven push buttons are designed which are performing the functions of ‘load image’, ‘enhance contrast’, ‘segment ROI’, ‘feature extraction’, ‘classification result’, ‘graphs for comparison’ and ‘exit’. By the name of the button it is very clear which type of function it is performing. Further it is showing the values of all the features extracted used for classification purpose. It is also showing the graphical comparison of all the classifier used. At the end it showing the name of the diseases occurring in the apple leaves.

This section discusses the experimental results for the four classifiers i.e. SVM, KNN, Naive Bayes and Decision Tree approach. During experimental evaluation, two performance metrics were measured i.e. True positive rate and false positive rate for all the algorithms. True positive rate and false positive rate plays a very important role in evaluating many machine learning algorithms and holds a great significance. True positive rate measures the proportion of positives that are correctly identified and false positive rate measures the proportion of positives that are not correctly identified and can be mathematically shown in equation (ii) and (iii).
Using true positive rate and false positive rate, Receiver operating characteristics (ROC) is drawn for four classes i.e. KNN, SVM, Naive Bayes, Decision tree shown in figure 14. ROC is a technique to measure the accuracy of the classifiers and is a graph between the true positive rate and the false positive rate. ROC curve gives the ability to access the performance of classification. For this most widely used parameter is AUC (area under curve). AUC is used to compare the performance of different classifiers. From the figure 14 it is clear that AUC is maximum for KNN and minimum for Regression decision tree. Values of AUC for different classifiers are given in Table 2.

Table 2: AUC Values for Different Classifiers

<table>
<thead>
<tr>
<th>Classifier</th>
<th>AUC Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>0.994</td>
</tr>
<tr>
<td>SVM</td>
<td>0.972</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.931</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.625</td>
</tr>
</tbody>
</table>

Figure 14: ROC Curve for Four Classifiers

7. Conclusion

As it is an era of digital technology, therefore one should make the maximum use of it in all the sectors. Agriculture covers major area of India’s geography. More than 70% people depend directly on agriculture. Various crops and fruits suffer from different disease and this produces fewer yields. So investing huge
amount of money in cultivating a crop selected by just intuition or judgment is not a wise option today. Therefore, machine learning came into play. In the proposed work, the prime focus was on apple diseases apple scab and marsonina coronaria and produce different results on disease prediction using artificial intelligence approach. Further as the part of experiment the priority of KNN over other machine learning algorithms in terms of accuracy was also proved. The graphs have been plotted in Matlab showing the variation of true positive v/s false positive rate and using these values accuracy is calculated. From the graphs it is clear that KNN is using maximum area under curve and accuracy of more than 99% can be achieved.

References


