INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An effective approach to feature extraction for classification of plant diseases using machine learning

S Jeyalakshmi[1]⃰, R Radha[2]

1 Research Scholar, Research Department of Computer Science, SDNB Vaishnav College for Women, Chromepet, Chennai, 600044, Tamil Nadu, India
2 Associate Professor, Research Department of Computer Science, SDNB Vaishnav College for Women, Chromepet, Chennai, 600044, Tamil Nadu, India

Abstract

Objectives: To make automatic classification of diseased potato and grape leaf from normal potato and grape leaf. **Methods:** Experimental sample size of 3000 and 4270 Potato and Grape leaf images were used respectively. The diseased and healthy leaf image samples were taken from PlantVillage dataset. The color features viz., average Red, Green, Blue and Hue intensities of Lesion region were calculated. Features namely Contrast, Dissimilarity, Homogeneity, Energy, Correlation, ASM, and Entropy were extracted from hue lesion region. Also, histogram features such as mean and standard deviation were extracted from hue infected region. Then, data normalization was done on feature set to bring all features into a common scale. Finally, Naïve Bayes, K Nearest Neighbor and Support Vector Machine Classifiers were applied on the above said feature sets. **Findings:** The Dataset was split in the ratio of 80% and 20% for training and test sets. The classifiers NB, KNN and SVM classified Potato leaves with an accuracy of 88.67%, 94.00% and 96.83% respectively and Grape leaves with an accuracy of 81.87%, 93.10% and 96.02% respectively. For both the species, SVM classifier gave the highest accuracy. Also, it was found that the proposed method performs well as compared with the related works in the literature. **Novelty/Applications:** An effective feature extraction method to classify grape and potato diseases was proposed in this research work. Also, it was found that the proposed method performs well as compared with the related works in the literature.

Keywords: RGB color space; HSV color space; histogram; color features; grey-level co-occurrence matrix; texture features

1 Introduction

Early detection and management of these diseases are essential to prevent plants from being infected in large numbers, thereby avoiding yield loss and economic loss. In[1] had proposed a method to classify grape plant diseases such as Black Rot, Esca, Leaf Blight and healthy leaves. In this work, texture features such as Energy,

Homogeneity, Contrast, Dissimilarity, Correlation and Angular Second Moment were extracted from infected region of leaf image. Random Forest Tree, AdaBoost and SVM classifiers were applied and it was determined that SVM classifier gave a highest accuracy of 93.04% among the three classifiers. In[2] had extracted features namely Contrast, Correlation, Energy, Homogeneity, Mean, Standard Deviation, Entropy, Variance, Cluster Shade, Kurtosis, Skewness, Cluster Prominance. They had classified grape leaf diseases such as Black Rot, Downey Mildew, Powdery Mildew, Leaf Roll and Healthy leaves from leaf images using SVM classifier and obtained an accuracy of 94%. In[3] had developed Kohonen's Neural Network for classification of Grape diseases Powdery Mildew Downey Mildew, Black Rot and Normal leaves with features namely Energy, Entropy and Correlation. An accuracy of 93.44% was achieved by this method. In[4] had suggested Back Propagation Neural Network for predicting grape diseases Downey Mildew, Powdery Mildew, Black Rot, Leaf Roll And Normal Leaf. They had fed features such as Energy, Entropy, Correlation, Cluster Prominence and Cluster Shade into the neural network and had observed an accuracy of 92.94%.

In[5] had developed a Back Propagation Neural Network model to classify healthy and diseased potato leaves viz., Early Blight, Insect damage and Roll Viral disease. They had extracted color features such as Mean, Median, Variance of R, G, B, H, S and I components of RGB and HSI color spaces respectively. Also, shape feature area and texture features like Energy, Entropy, Contrast, Homogeneity and co-relation were also used in classifying the diseases. In[6] classified potato diseases late blight, early blight and healthy leaves by extracting texture features like Contrast, Correlation, Energy and Homogeneity; Histogram features viz. Mean, Standard deviation, entropy, skew and energy of color planes using SVM classifiers and got an accuracy of 95%.

In[7] proposed a back propagation network to recognize the grape diseases powdery mildew and downy mildew with a dataset size of 85 diseased leaf images and obtained an accuracy of 91%. In[8] suggest grape disease detection system using SVM classifier. They have achieved an accuracy of 88.89% with a dataset size of 137 images. In[9] proposed a vine leaf classification system to classify as healthy, downy mildew, powdery mildew and black rot through one class classifiers. The conflicts among classifiers were solved by the nearest support vector strategy. The conflicts occurred and resolve in more than 50% of the test image. The conflict resolution should occur less frequently. The model has been trained with 8 images per class achieved an accuracy of 95% on 46 test images. In[10], Tiwari and Tarun suggest Support Vector Machines–Cuckoo Search optimization algorithm for classification of plant leaf diseases Alternaria Alternata, Cercospora Leaf Spot, Anthracnose and Bacterial Blight. They had experimented the study with a sample size of 150 diseased leaf images and obtained an accuracy of 96.5% to 98.5%.

In[11] have reviewed automated disease recognition and Cite[6] and commented that it would be better to provide the performance of the system on a larger dataset. The same comment can be applied to[7–10]. Comparatively a larger dataset size was considered in this study.

Thus, from the previous reports, there is no mention on the kind of gray scale image used for extracting texture features of the image. Hence, it can be taken that the authors have done RGB image converted into gray scale. The disease symptoms can be distinguished from one another from variations in color and shades of color and these color information can be acquired from Hue value. The present work extracts for Histogram based features and texture features from Hue component of the lesion region to get the information most significant in distinguishing the diseases.

This study attempt to distinguish healthy Potato leaves from the diseased ones. Figure 1 (a) and Figure 1(b) shows healthy Grape and Potato leaves in which entire leaf area is in green color. The causes and symptoms of the diseases with a sample leaf are tabulated in Table 1.
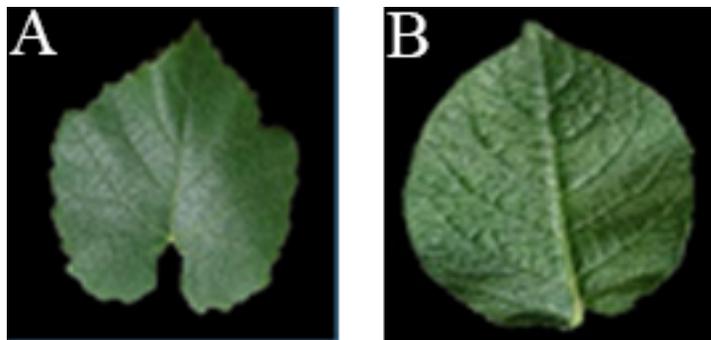

Fig 1. Healthy leaf image (a) Grape(b) Potato

Table 1. Causes and symptoms of diseases in potato and grape species

Specie	Disease	Cause	Symptom	LDS
Potato	Early Blight	Fungus; *Alternaria solani* [12]	Early symptoms: small irregular to circular dark brown spots restricted by leaf veins. On severely infected leaves: small lesions coalesce and cover large areas of the leaf [12]	Refer fig. A
	Late Blight	Fungus; *Phytophthora infestans* [13]	circular to irregular–shaped dark brown or black lesion	Refer fig. B
Grape	Black Rot	Fungus *Guignardia bidwellii* [14]	Reddish brown and circular-to- angular spots that merge into irregular blotches	Refer fig. C
	Esca (Black Measles)	Fungi *P. aleophilum* and *Phaeomoniella chlamy-dospora* [15]	Interveinal (in between veins) striping starts out as dark red and become necrotic (premature death of cells)	Refer fig. D
	Leaf Blight	Fungus [16]	Lesions are dull red to brown in color turn black later. If disease is severe this lesions may coalesce.	Refer fig. E

LDS: Leaf with diseased spot

A   B   C

D   E
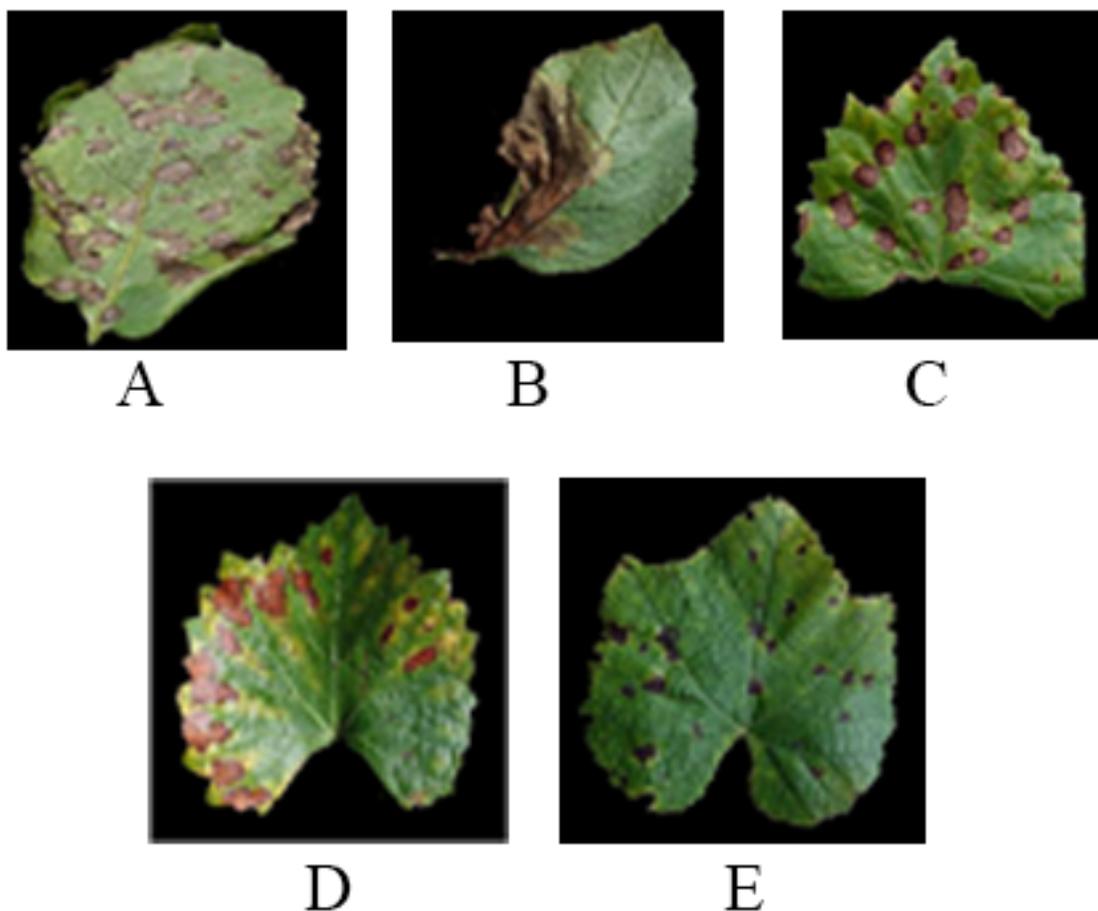
2 Materials and Methods

The experiments done in this study were carried out on the Plant Village Dataset[17]. A data set size of 2000 diseased potato leaf images and 3270 diseased grape leaf and 1000 healthy leaf images of both the species were used in implementing this research work. All the images considered are of size 256 X 256.

The Plant Village dataset is a collection of 54,306 images of healthy and diseased plant leaves 14 plant species and 26 diseases.

2.1 Color features

Average intensity values of Red, Green and Blue components of RGB color space and Hue component of HSV color space are calculated as color features. It is calculated by finding the average pixel values of the Grey scale image and is given by

$$AVG_i = \sum \frac{f_i(x,y)}{N} \tag{1}$$

where $f_i(x,y)$ is the intensity value of pixel in component i, is the total number of pixels in the image, and represents the color components Red, Green Blue and Hue.

2.2 Histogram features

Histogram plots the frequency of occurrence of each intensity value in an image. The formula for calculating Weighted Mean and Weighted Standard deviations[18] from histogram are given below:

$$Weighted\ Mean\ \mu^* = \frac{\sum_{i=1}^{n}(x_i * w_i)}{\sum_{i=1}^{n} w_i} \tag{2}$$

$$Weighted\ Standard\ Deviation\ \sigma_w = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_i - \mu^*)^2}{\frac{\left(N'-1\right)\sum_{i=1}^{N} w_i}{N'}}} \tag{3}$$

where w_i is the weight of the i$^{\text{th}}$ observation, N' is the number of non-zero weights, μ^* is the weighted mean.

2.3 Grey level co-occurrence matrix (GLCM) and Image Texture

An image texture is a spatial arrangement of intensities or Grey Levels in an image or selected region of an image. GLCM is tabulation of how frequent different combination of Grey levels occurs in an image. In this work, Hue component of the HSV color space was used as gray image as variations in color value determines the disease.

Any gray scale image has 256 gray levels ranging from 0 to 255 and hence the size of GLCM will be 256 X 256. In this research work, the size of GLCM matrix generated was 32X32 as the number of gray levels was reduced to 32.

In this study, the leaf image was partitioned into 16 X 16 patches and patches having more than 10% of information were considered for processing. Patches having less than 10% information were discarded since for most of the seed pixels the gray level intensity values of their neighboring pixels be 0. And these pixels may not contribute to textural information and may lead to misclassification. The texture features such as Contrast, Dissimilarity, Homogeneity, Angular Second Moment, Energy, Correlation and Entropy of useful patches were taken as the texture features for the leaf.

2.4 Feature normalization

Feature Normalization converts the feature values to a common range of values. Normalization of feature values is required when the features have different range of values. It is an important pre-processing step required for applying classification algorithms, like K Nearest Neighbors[18] and Support Vector Machines[19], which computes distance measure. SVM assumes that the data are in the range 0 to 1 or -1 to 1[20]. But, for certain algorithms, like Naïve Bayes, feature normalization may not have much difference[20].

Min-Max Scaling normalizes each feature to a given range of values using[18]

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{12}$$

Where max and min are feature range[21].

2.5 Classifiers

Classification in machine learning is a supervised learning method in which models are trained to learn the mapping function, from input X to output Y, Y= f(X). Here, X is the feature set and Y is the set of Categories or Classes. Then, this mapping function is used in predicting the classes of new observations.

2.5.1 Naïve Bayes classifier

Naive Baysian classifier is a probabilistic supervised learning algorithm. The algorithm is used to predict the class Y, for the feature set X by applying Bayes rule. The Bayes rule uses conditional probability P(X|Y], which can be calculated from the training dataset, to find P[Y|X]. Bayes rule is given by

$$P(Y \mid X) = \frac{P(X \mid Y) * P(Y)}{P(X)} \tag{13}$$

Naïve Bayes algorithm can be applied when there are multiple features and all of them are independent of each other and is given by

$$P(Y = k \mid X) = \frac{P(X \mid Y = k) * P(Y = k)}{P(X)} \tag{14}$$

2.5.2 KNN classifier

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. Figure 2 illustrates K Nearest Neighbor Classification.
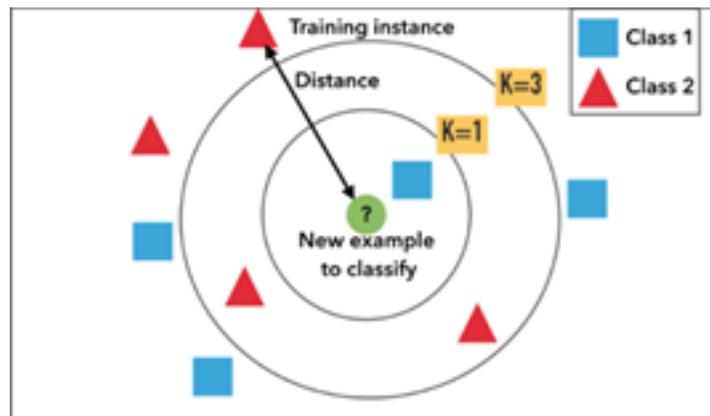
Fig 2. KNN classifier [22]

2.5.3 SVM classifier

Support Vector Machine is a supervised machine learning algorithm which is commonly be used in classification problems. This method plots data in n-dimensional space with n feature values as co-ordinate positions. The algorithm outputs an optimal hyper-plane that clearly classifies data points and the samples on the margin are support vectors. The dimension of the hyper-plane is determined by the number of features. Further, this hyper-plane is used in predicting new examples.

The advantage of using SVM classifier is it tries to achieve a maximum margin. Figure 3 shows a maximum-margin hyper-plane. A margin is a split-up of line to the closest class point. A good margin is the one in which the data point of one class does not cross the other classes.
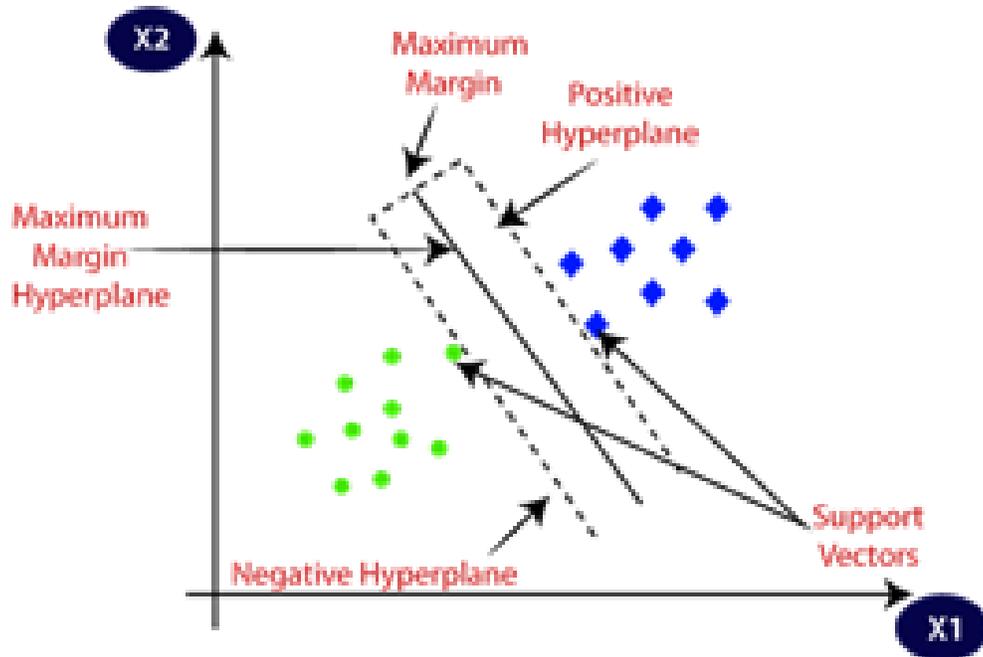
Fig 3. Maximum-margin hyper-plane

2.6 K-Fold cross validation

K-Fold Cross Validation is a technique used in estimating the performance of a machine learning models. In this, the dataset is split into k-parts, called folds. In the first iteration, the first fold is used to test the model and the remaining k-1 folds are used to fit the model. In the second iteration, the second fold is used as test set and the remaining k-1 folds are used as training set. The process is repeated until every fold is given a chance to be the held out test set. Two important sources of errors bias and variance can be obtained from K-Fold Cross Validation. A High bias and a low variance indicate underfitting that is the model does not fit the data well. The model is said to overfit, when the model learns the data excessively well such that it also fit noise present in the data. This is the situation where the model performs extremely well on training data but performs poorly for test data. In order to neither overfit nor underfit the model needs to be a generalized one. A generalized model fits to the data set such that it performs equally well on both training and test set. This is a result of low bias and low variance or trade-off between bias and variance.

3 Proposed methodology

The overall workflow of this research work is shown in the Figure 4. In the proposed work the background, from RGB leaf image, was removed using automatic enhanced GrabCut algorithm [23]. Figure 5 shows the input leaf image and leaf image obtained after applying enhanced GrabCut algorithm. Then the lesion (infected) region was segmented from the leaf. Further, Red, Green and Blue color features were extracted from segmented RGB image, Hue color feature, histogram and texture features of Hue component of lesion region of HSV color space were extracted. Naïve Bayes, K Nearest Neighbor and Support Vector Machine classifiers were applied on the above said feature sets to classify the diseased leaf image.
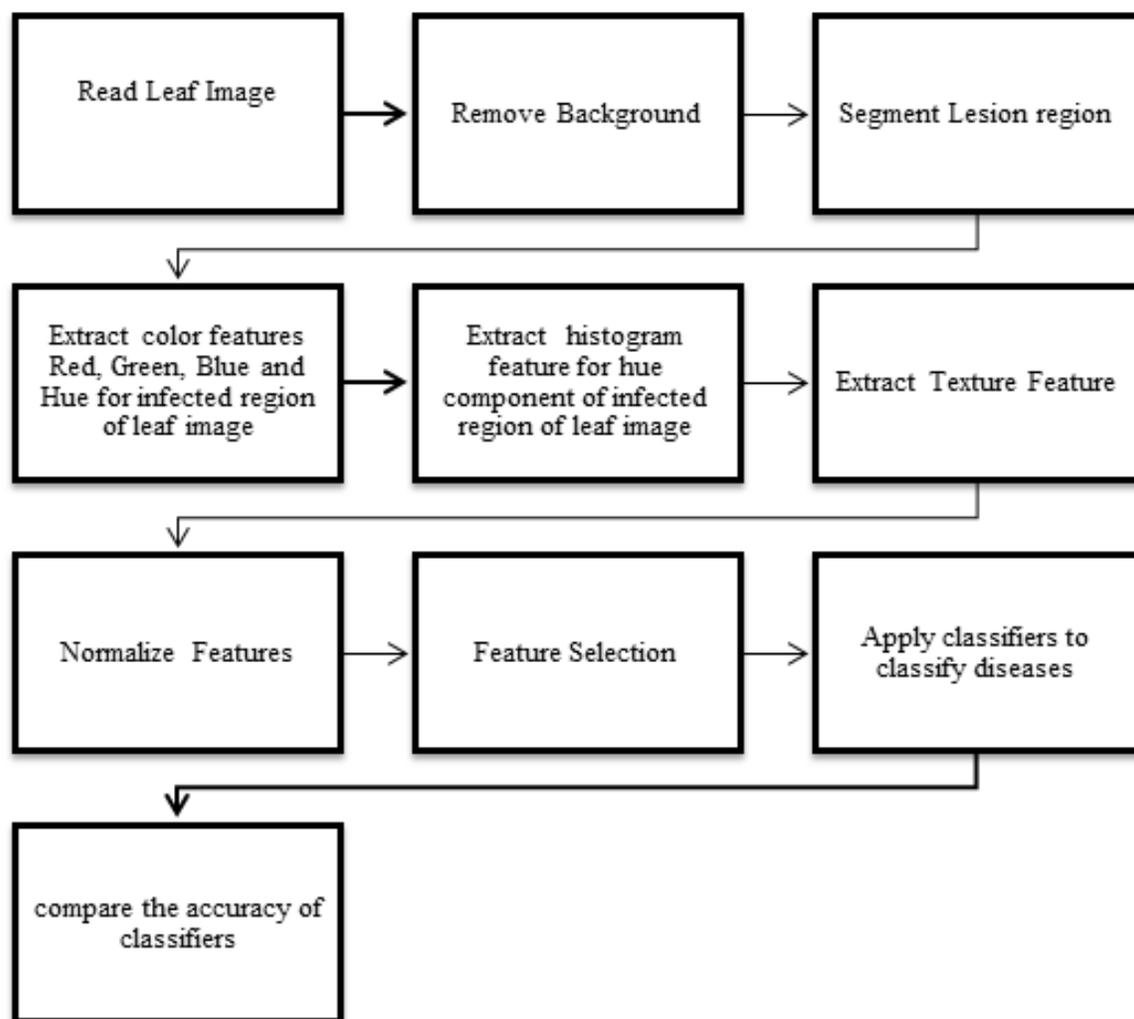
Fig 4. Overall workflow
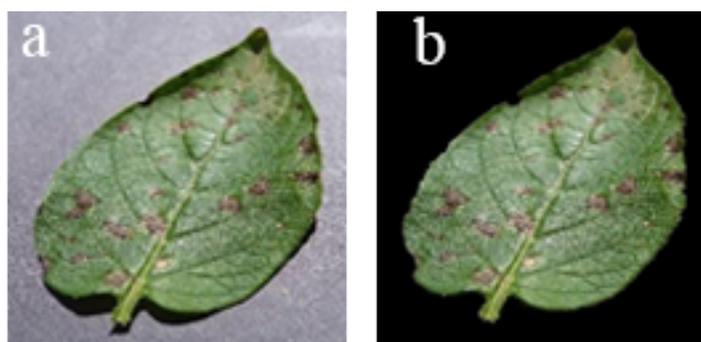
Fig 5. Background removal (a) Input Leaf image (b) Leaf image after removing background

3.1 Segmentation of lesion region

Let img_{leaf} be the RGB leaf image. img_{leaf} was converted into HSV color space, img_{hsv}. The hue component, img_{hue}, was extracted from img_{hsv}. A mask corresponding to Green region was created by thresholding hue values between 36 and 104 on img_{hue}. The mask was applied on img_{hue} for segmenting Lesion region, img_{lesion}. The results obtained by segmenting infected region from leaf image are tabulated in Figure 6.

Species	Disease	Infected Leaf Image img_{leaf}	Uninfected region mask	Region of Interes(Infected region) img_{lesion}
Potato	Early Blight			
	Late Blight			
	Healthy			
Grape	Black rot			
	Leaf Blight (Isariopsis)			
	Healthy			
	Esca (Black Measles)			
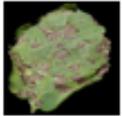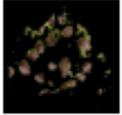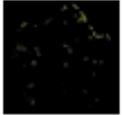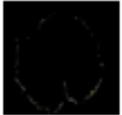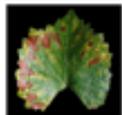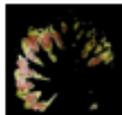
Fig 6. Lesion region segmentation
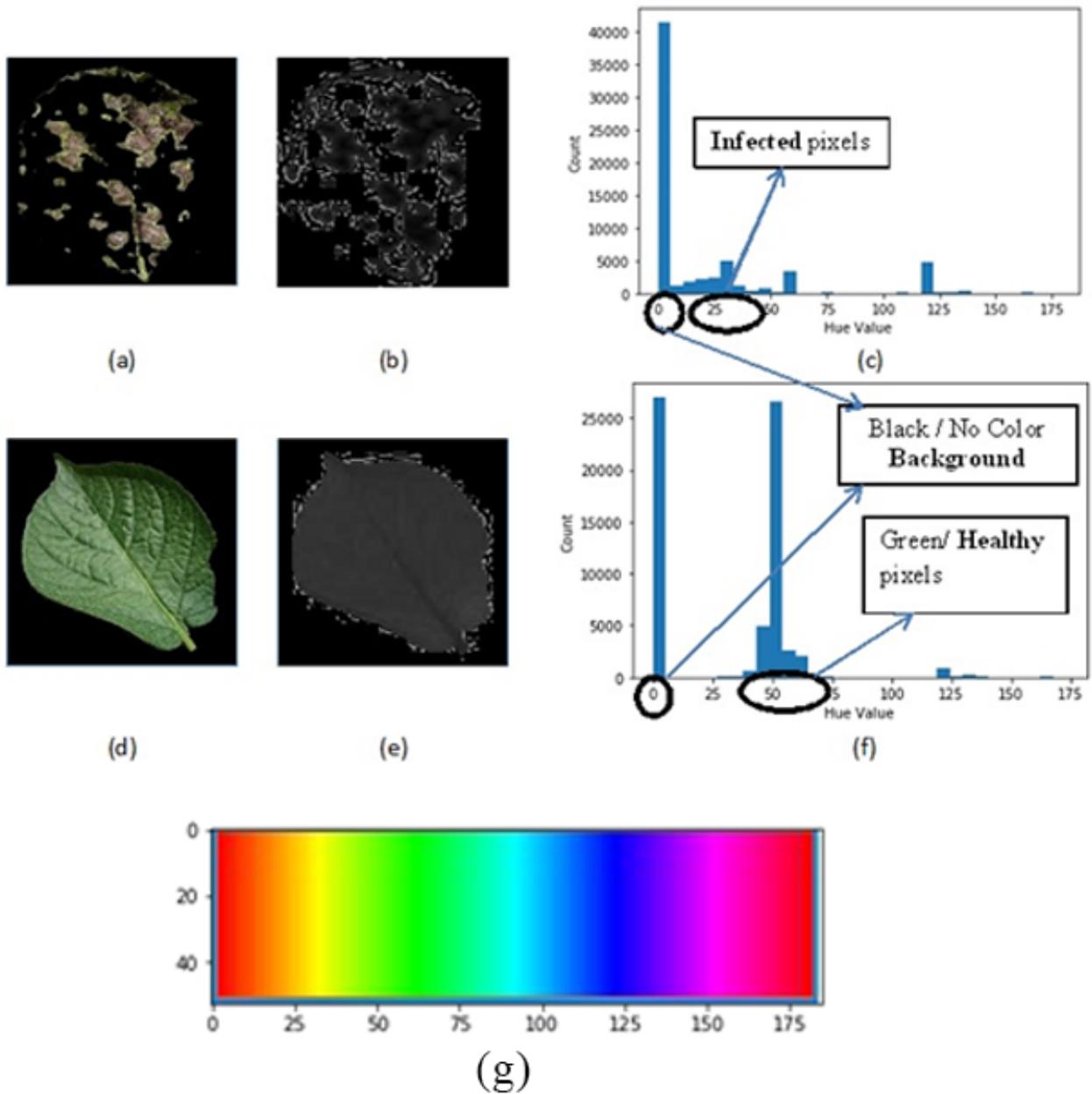
Fig 7. (a) infected region (b) hue component (c) 3Histogram of hue component of lesion region (d) Healthy leaf (e) hue component (f) Histogram of hue component of healthy region (g) Hue values corresponding totrue colors in degrees

3.2 Feature extraction

This section discusses the method of extraction of various features used in this research work in classifying the plant diseases.

3.2.1 Color features

Algorithm 1 extracts average color features of Red, Green and Blue components of RGB image and Hue component of HSV color space. The average color features extracted from algorithm 1.1 is given in Table 2 .

Table 2. Color features for lesion region

Species	Class/Disease	Hue Intensity	Red Intensity	Green Intensity	Blue Intensity
		avg_{hue}	avg_{red}'	avg_{green}	avg_{blue}
Potato	0: Early Blight	25.40	118.63	114.33	92.83
	1: Healthy	31.45	67.00	68.06	50.94
	2: Late Blight	22.09	95.91	87.82	63.95
Grape	0: Black Rot	29.87	90.47	88.06	55.79
	1:Esca (Black Measles)	23.46	97.26	71.67	53.68
	2:Healthy	31.37	133.83	135.80	97.40
	3: Leaf Blight	31.91	123.95	126.76	60.99

Algorithm 1: Extraction of Color Feature

Procedure ExtractColorFeature

Input: img_{lesion}

Output: color features avg_{red}, avg_{green}, avg_{blue}, avg_{hue}

$\qquad \text{img}_{red}$' img_{green}, img_{blue} = split(img_{lesion})

$\qquad \text{avg}_{red}$ =AverageIntensity(img_{red})

$\qquad \text{avg}_{green}$ =AverageIntensity(img_{green})

$\qquad \text{avg}_{blue}$ =AverageIntensity(img_{blue})

$\qquad \text{img}_{hsv}$ = rgbtohsv(img_{lesion}) img_{hue}, img_{sat},

$\qquad \text{img}_{hue}$' = split(img_{hsv})

$\qquad \text{avg}_{hue}$ = AverageIntensity(img_{hue})

End Procedure

Algorithm 1.1: AverageIntensity

Input $\text{img}_{leafcolor}$

Output avg

Procedure AverageIntensity($\text{img}_{leafcolor}$)

$\qquad$ m, n = size($\text{img}_{leafcolor}$)

$\qquad$ sum=0

$\qquad$ For i = 1 to m do

$\qquad\qquad$ For j= 1 to n do

$\qquad\qquad\qquad$ sum =sum + $\text{img}_{leafcolor}$(i,j)

$\qquad$ End For

$\qquad$ avg = sum / (m*n)

$\qquad$ Return avg

End Procedure

3.2.2 Histogram features

Histogram for img_{hue} plotted with bins representing Hue value on x axis and the number of occurrence of the hue color on the y axis, was plotted as shown in Figure 7 (c) and Figure 7(f). The hue values corresponding to true colors are depicted in Figure 7(g). The mean and the standard deviation were calculated from histogram. The results obtained are given in Table 3 . Histogram features mean and standard deviation were extracted from hue value of lesion region of one leaf image. The infected regions show variation in color. There will be huge variations in color i.e., hue value when there is a change in color. Hence, the standard deviation takes a higher value than mean value for lesion region of infected leaves.

Table 3. Histogram feature

Species	Class/Disease	Mean	SD
Potato	0: Early Blight	27.67	36.66
	1: Healthy	7.45	2.10
	2: Late Blight	18.14	26.36
Grape	0: Black Rot	20.12	26.09
	1:Esca (Black Measles)	23.46	31.22
	2:Healthy	11.44	7.43
	3: Leaf Blight	43.37	53.79

3.2.3 Texture features
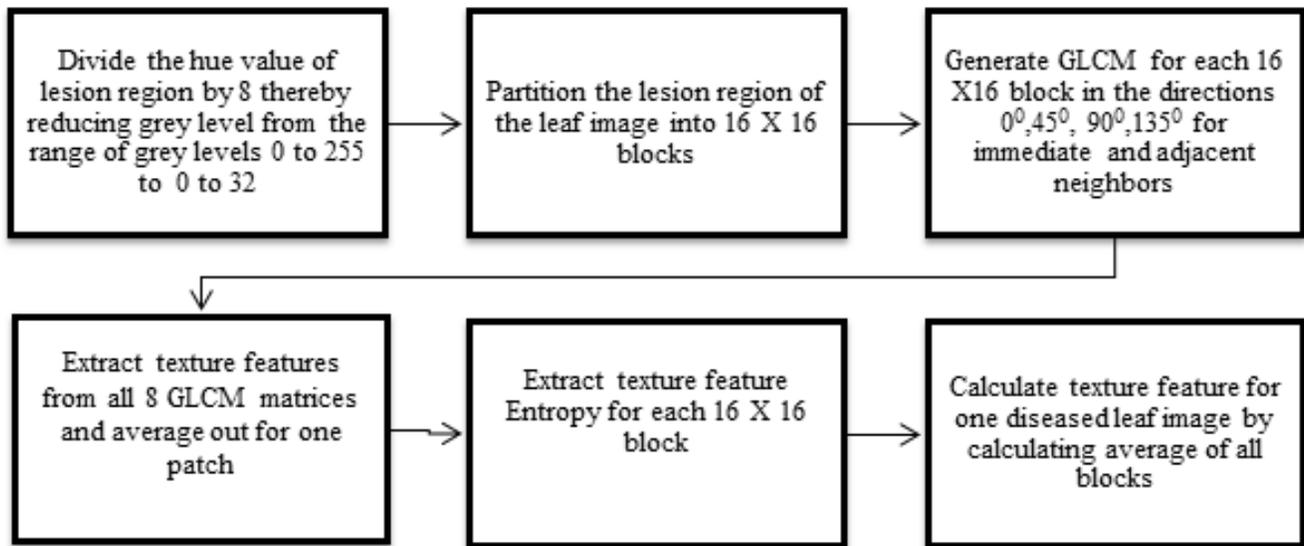
Fig 8. Workflow for texture feature extraction

The infected leaf tissues (lesion region) are rough in nature whereas the normal leaf tissues are smooth. With the help of texture features, it can be determined that whether a region of a leaf image is a rough one or a smooth one. For a rough region, the difference between neighboring grey pixel values will be very large whereas for a smooth region, the neighboring pixels will have the same or closer grey values. Figure 8 demonstrates the sequence of steps performed in this research work to extract texture feature and Algorithm 2 implements the same.

Algorithm 2: Extract texture features

Input: img_{lesion}

Output: Texture features Contrast, Dissimilary, Homogeneity, Energy, Correlation, ASM, Entropy

Procedure ExtractTextureFeature

      img_{hsv} = rgbtohsv(img_{lesion})

      img_{hue}, img_{sat}, img_{value} = split(img_{hsv})

      greygroup = 8

      greylevel=(256/greygroup)

      img_{hue} = img_{hue}/ greygroup

      m, n = size(img_{hue})

      [no_of_patches , patch]=Partition img_{hue} into 16 X 16 patches

      For i= 1 to no_of_patches do

            // If the patch has 10% or more useful pixels then extract texture features for this patch

            neighbors = [1,2] // immediate neighbor and alternate pixels

            direction=[0^0, 45^0,90^0,135^0]

            g= GLCM(patch[i],neighbors, direction, greylevel)

            // 8 values corresponding to different neighbors & directions will be returned

            contrast = sum(g.contrast)/8

            dissimilary = sum(g.dissimilarity)/8

            homogeneity = sum(g.homogeneity)/8

            energy = sum(g.energy)/8

            correlation = sum(g.correlation)/8

            ASM = sum(g.ASM)/8

            entropy = g.entropy

      End For

End Procedure
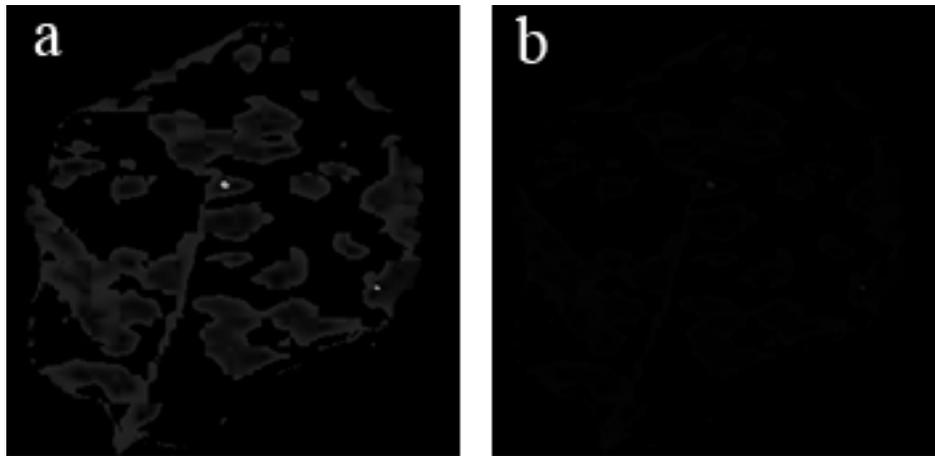
Fig 9. Hue component of Lesion Region (a) Hue component of Lesion Region (b) Image obtained by dividing Hue by 8 (Not visible to naked eyes)

The input images, used in this research work, are of size 256 X256. Generating a Grey Level Co-occurrence matrix of size 256 X 256 for each image is a complex task. Thus, the Grey levels in hue component of lesion region were divided by 8 to reduce the Grey levels into 32(0 to 31). This will reduce the size of GLCM from 256 X 256 into 32 X 32 thereby increasing the calculation speed and decreasing the complexity. Figure 9 (a) shows the hue component of lesion region. Figure 9(b) shows the image obtained from dividing hue value by 8 which is not visible to the naked eyes.

It can be observed from the Table 5 (a) that the intensity levels of neighboring pixel values are either very close or vary with huge difference.

Close neighboring hue values indicate neighboring pixels are of same color or of same shades of color. When the pixel values are divided by 8, they will be grouped into same gray level or into adjacent grey levels as shown in Table 5(b). A huge variation

Table 5. Hue values and Hue/8values for a 16 X 16 block (a) Sample Hue values of lesion region (b) Hue value/8

(a)

4	175	174	177	1	5	4	4	6	7	8	8	6	4	3	4
4	172	170	174	4	8	6	7	8	8	7	7	5	5	4	5
10	178	174	179	7	9	10	9	10	10	9	8	6	5	4	6
13	3	0	2	9	11	11	11	11	10	8	7	5	3	4	6
17	9	2	4	10	13	12	12	12	11	9	6	4	1	2	6
18	10	4	5	11	12	13	12	11	11	8	5	1	0	1	6
18	10	4	4	10	11	11	11	11	8	6	3	179	177	179	5
18	9	2	3	8	11	11	10	8	8	5	1	178	177	178	4
17	5	178	179	6	9	8	9	7	6	3	2	179	177	179	7
17	4	176	175	4	8	8	8	8	6	4	5	2	4	7	12
17	4	174	176	6	8	6	5	7	7	10	10	10	12	14	19
18	6	177	179	7	10	7	7	10	12	16	18	18	20	22	25
18	10	2	4	11	13	11	11	15	18	21	23	25	28	29	31
19	12	6	9	16	15	13	15	18	22	26	29	29	32	33	33
19	15	11	12	16	15	17	19	24	28	29	32	32	35	34	35
21	15	18	18	22	25	26	28	31	32	33	34	34	35	35	35

(b)

|															
1	22	22	22	0	1	1	1	1	1	1	1	1	1	0	1
1	22	21	22	1	1	1	1	1	1	1	1	1	1	1	1
1	22	22	22	1	1	1	1	1	1	1	1	1	1	1	1
2	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1
2	1	0	1	1	2	2	2	2	1	1	1	1	0	0	1
2	1	1	1	1	2	2	2	1	1	1	1	0	0	0	1
2	1	1	1	1	1	1	1	1	1	1	0	22	22	22	1
2	1	0	0	1	1	1	1	1	1	1	0	22	22	22	1
2	1	22	22	1	1	1	1	1	1	0	0	22	22	22	1
2	1	22	22	1	1	1	1	1	1	1	1	0	1	1	2
2	1	22	22	1	1	1	1	1	1	1	1	1	2	2	2
2	1	22	22	1	1	1	1	1	2	2	2	2	3	3	3
2	1	0	1	1	2	1	1	2	2	3	3	3	4	4	4
2	2	1	1	2	2	2	2	2	3	3	4	4	4	4	4
2	2	1	2	2	2	2	2	3	4	4	4	4	4	4	4
3	2	2	2	3	3	3	4	4	4	4	4	4	4	4	4

in hue colors indicate that the neighboring pixels are of different colors and division by 8 results in corresponding pixel values with different grey levels. Hence, dividing the hue value by 8 thereby reducing the grey levels will not affect the performance of the classification model as color difference was an important parameter that was taken into consideration for texture analysis.

It is meaningless to generate GLCM elements for image regions where not enough or no information is present. It can be seen from Figure 10 that the infected leaf area occupies less than 50%, of the overall image area not including background pixel area, uninfected leaf area. This is the regions of interest to be taken into consideration for generating GLCM. Rest of the image area can be discarded. To take into account only the infected regions, the leaf image Grey8Imglesion is partitioned into 16 X 16 blocks. The entire lesion region as a whole needs to be considered when there is a definite shape and shape considered as a feature in classifying the disease. The plant disease symptoms considered in this work doesn't have a definite shape. Though initial symptoms show definite shape, as the disease develops further the lesion regions merge together and become irregular in shape. Hence, partitioning the image into 16 X 16 patches will not affect the result.

Rejected

patch with less than 10% information

Processed further

patch with greater than 10% information
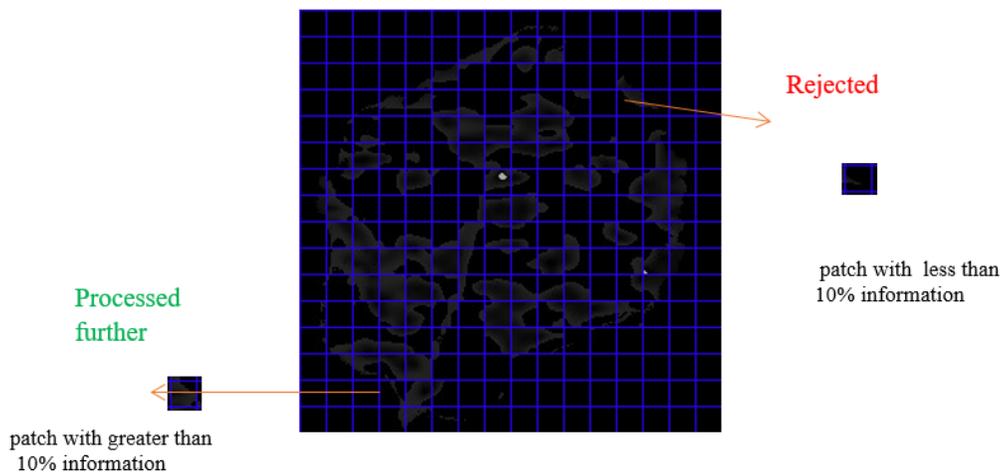
Fig 10. 16 X16 Lesion segment (Shown with Hue component)

Each patch with less than 10% information is rejected and not considered for further processing. From experiments, it was found that the patch with less than 10% of information does not affect the result.
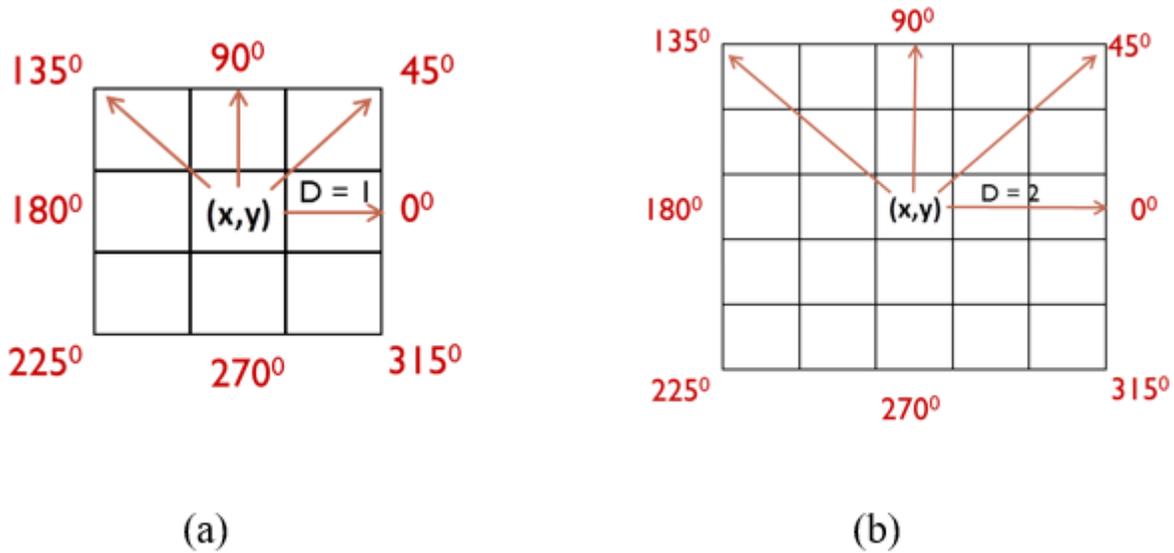
Fig 11. Distances with angles 0°, 45°, 90°,135° (a) Immediate neighbour (Distance D=1) (b) Alternate neighbour (Distance D=2)

For each of the useful patch the Grey Level co-occurrence matrices, for immediate neighboring pixels and alternate neighboring pixels in the directions with angles 0^0, 45^0,90^0 and 135^0 degrees([0, pi/4, pi/2, 3*pi/4]), were generated as illustrated in Figure 11. As the GLCM matrix is a symmetric matrix, the upper and lower triangular elements are same and these duplicate values were omitted and only the feature values for the angles 0^0, 45^0, 90^0 135^0 were calculated. This result in 8 set of values, two for each direction and each distance, for the feature sets contrast, dissimilarity, energy, correlation and ASM etc. Averages of features were calculated to obtain one set of features for each patch. Table 6 shows the GLCM features for one patch.

Table 6. LCM texture features for one patch

Direction	Distance =1 (Immediate Neighbor)				Distance =2 (Alternate Neighbor)			
	0°	45°	90°	135^0	0^0	45^0	90^0	135^0
Contrast	0.20	0.50	0.33	0.14	0.43	0.50	0.64	0.14
Dissimilarity	0.05	0.12	0.08	0.04	0.11	0.12	0.16	0.04
Homogeneity	0.99	0.97	0.98	0.99	0.97	0.97	0.96	0.99
Energy	0.95	0.95	0.95	0.96	0.95	0.95	0.95	0.96
Correlation	0.85	0.57	0.73	0.87	0.65	0.57	0.38	0.87
ASM	0.90	0.90	0.90	0.92	0.90	0.90	0.90	0.92

GLCM features give the degree of correlation between pairs of pixels with gray level values. These gray level values represent hue value or color information in this study. The changes in hue value indicate a change in color or shades of color. Here, Inter-pixel correlation between adjacent pixels were measured by taking distance =1. Degree of inter-pixel correlation between alternate pixels at distance =2 to get better information about details of texture. This helps in discriminating various diseases based on hue values. Thus, GLCM features were calculated for all useful patches and averaged out to get one feature set for one leaf image.

The average features of all useful patches were obtained and considered as the GLCM texture feature for the infected region. For all useful patches, Entropy was obtained for the entire patch and averaged out for calculating the Entropy for one leaf image. Table 7 shows the GLCM texture feature set obtained for one leaf image.

Table 7. GLCM Texture feature values for one leaf

Species	Class/Disease	Contrast	Dissimilarity	Homogeneity	Energy	Correlation	ASM	Entropy
Potato	0: Early Blight	1.85	0.44	0.86	0.63	0.66	0.42	1.10
	1: Healthy	0.73	0.19	0.95	0.95	0.13	0.90	0.01
	2: Late Blight	0.89	0.32	0.89	0.63	0.68	0.42	0.78
Grape	0: Black Rot	5.37	0.64	0.84	0.64	0.53	0.44	1.05
	1: Esca (Black Measles)	4.12	0.60	0.83	0.63	0.51	0.42	0.94
	2:Healthy	1.24	0.41	0.87	0.75	0.63	0.56	0.07
	3: Leaf Blight	3.31	0.54	0.86	0.60	0.64	0.38	1.49

3.2.4 Univariate Analysis

Univariate data analysis takes single data, summarizes it and finds the patterns in the data[24]. Here, univariate analysis was done to determine the range of feature values for each class. Figures 12 and 13 show the univariate analysis corresponding to Potato and Grape diseases.
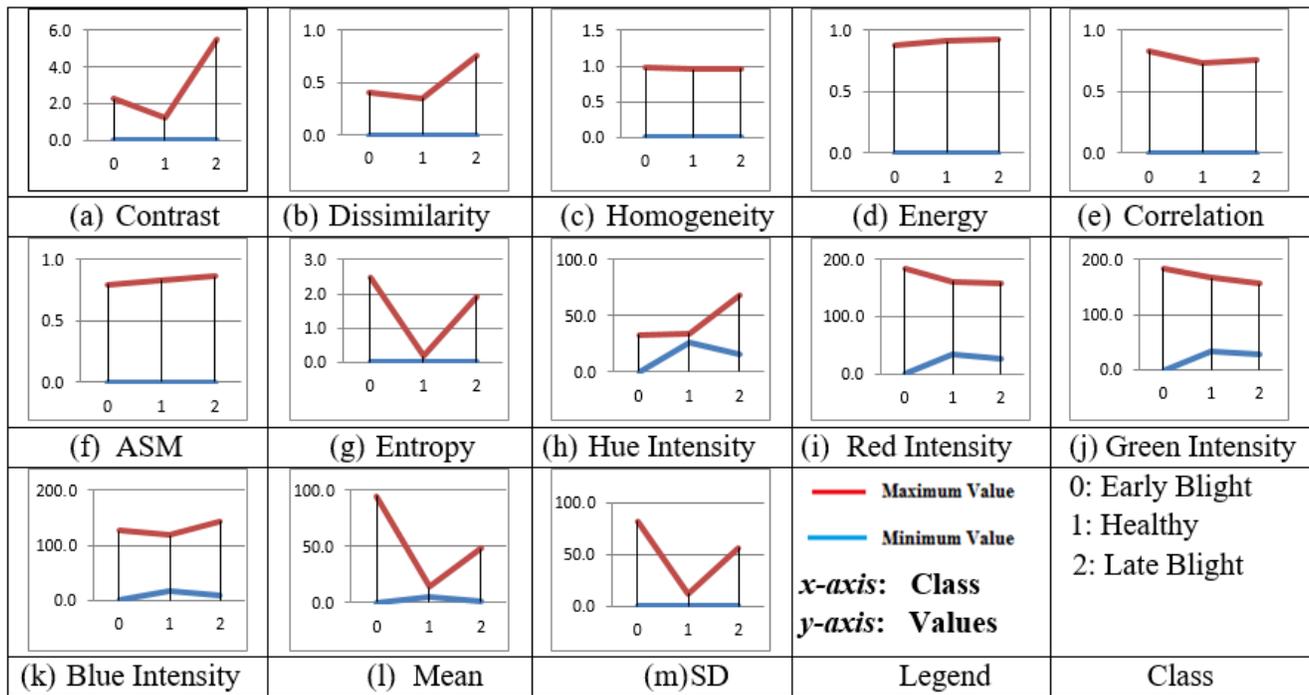
Fig 12. Range of feature values corresponding to potato diseases
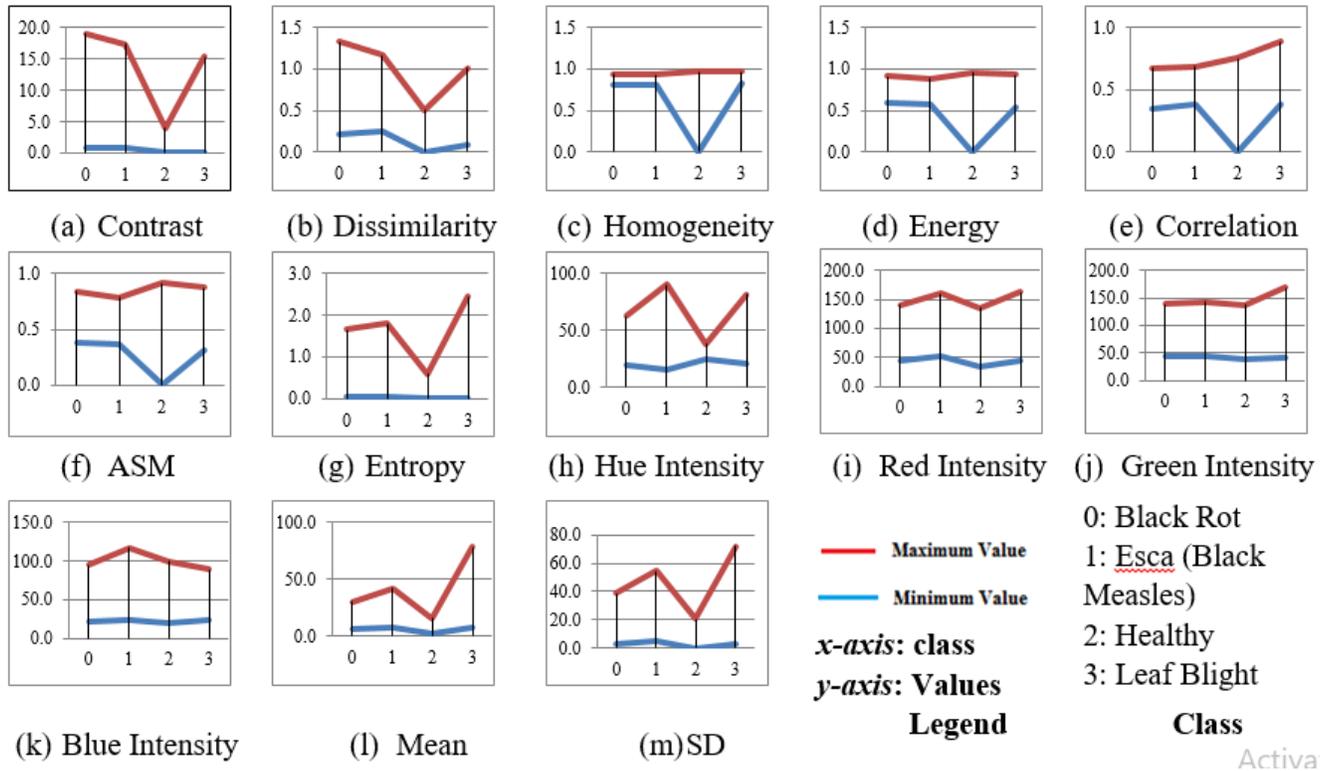
Fig 13. Range of feature values corresponding to grape diseases

3.2.5 Feature Normalization

Features like Hue Intensity, Red Intensity, Green Intensity, Blue Intensity, Mean of Hue intensity, SD of Hue Intensity, Contrast, Dissimilarity and Entropy values are not in normalized form. They are normalized, using Min-Max Scaling, between 0 and 1. Table 8 shows the normalized feature values. Note that the features such as Homogeneity, Energy, Correlation, and ASM are already in normalized form.

Table 8. Normalized features

Species	Class	Contrast	Dissimilarity	Entropy	Hue Intensity	Red Intensity	Green Intensity	Blue Intensity	Mean	SD
	0: Early Blight	0.07	0.21	0.54	0.20	0.46	0.43	0.38	0.30	0.50
Potato	1: Healthy	0.14	0.19	0.01	0.32	0.04	0.05	0.20	0.05	0.02
	2: Late Blight	1.00	1.00	0.01	0.83	0.32	0.30	0.51	0.02	0.05
	0: Black Rot	0.26	0.42	0.17	0.28	0.27	0.23	0.16	0.10	0.19
Grape	1:Esca (Black Measles)	0.09	0.33	0.44	0.05	0.53	0.31	0.34	0.27	0.45
	2:Healthy	0.14	0.29	0.04	0.25	0.41	0.37	0.52	0.10	0.09
	3: Leaf Blight	0.09	0.31	0.50	0.23	0.71	0.74	0.37	0.49	0.71

3.2.6 Feature selection techniques

To improve the speed of the machine learning algorithms, feature selection techniques are applied. Chi^2 Statistical test and ANOVA test have been performed on the normalized feature set and target variable. The results obtained are tabulated in Table 9.

Table 10 shows the classification accuracy obtained by keeping top significant features and removing irrelevant features.

Table 9. Feature significance in predicting plant diseases

Feature	Chi Square Score	Feature	ANOVA Score
Mean	74.43	Green Intensity	1912.37
SD	69.95	Red Intensity	1644.39
Green Intensity	59.63	SD	884.82
Entropy	54.20	Blue Intensity	786.07
Red Intensity	44.37	Mean	654.42
Blue Intensity	19.64	Entropy	387.83
Hue Intensity	3.71	Hue Intensity	169.30
ASM	0.23	ASM	30.57
Contrast	0.21	Energy	21.78
Dissimilarity	0.08	Correlation	11.82
Energy	0.05	Contrast	5.21
Correlation	0.04	Dissimilarity	4.28
Homogeneity	0.00	Homogeneity	3.64
(a) Chi^2 Test		(b) ANOVA Test	

Table 10. Classification accuracy obtained by feature selection techniques

No. of Features	Potato			Grapes		
	NB	KNN	SVM	NB	KNN	SVM
13	88.67%	**94.00%**	**96.83%**	81.87%	93.10%	**96.02%**
12	89.95%	91.46%	89.95%	**88.51%**	93.59%	94.72%
11	87.69%	91.46%	87.69%	88.00%	92.93%	94.39%
10	**90.20%**	92.78%	90.20%	87.91%	**94.32%**	95.90%
9	89.45%	90.70%	89.45%	87.85%	93.66%	94.52%
8	90.20%	91.21%	90.20%	88.24%	92.21%	93.53%
7	83.92%	84.67%	83.92%	87.05%	92.73%	93.13%

Feature selection techniques are used to select features that are useful for classification. The prediction accuracy obtained from the classifiers, tabulated in Table 10, clearly shows that classification accuracy increases when all the 13 features are taken into account. Here, the fact that removing features reduces accuracy and adding all features improves accuracy indicates that all the extracted features were significant and hence all the 13 features, namely, Contrast, Dissimilarity, Homogeneity, Energy, Correlation, Angular Second Moment (ASM), Entropy, Hue Intensity, Red Intensity, Green Intensity, Blue Intensity, Mean and Standard Deviation (SD), were used in classifying the plant diseases.

3.2.7 Data set size
The size of the dataset used in this research work is 3000 potato leaf images and 4270 grape leaf images. The dataset was split into 80% and 20% for training and test sets. Table 11 presents the training and test size.

Table 11. Training and test set size

Species	Class Label/ Disease	Training Set 80%	Test Set 20%	Total
Potato	0: Early Blight	820	180	1000
	1: Healthy	810	190	1000
	2: Late Blight	770	230	1000
	Total	**2400**	**600**	**3000**
Grape	0:Black Rot	820	202	1022
	1:Esca (Black Measles)	978	246	1224
	2:Healthy	792	208	1000
	3:Leaf Blight	825	199	1024
	Total	**3415**	**855**	**4270**

4 Results and Discussion

Table 12 shows the confusion matrices for NB, KNN and SVM classifiers. Out of 600 Early Blight test set samples 532, 564 and 581 diseased leaf images had been correctly classified by NB, KNN and SVM respectively. 700, 796 and 821 diseased leaf images out of 855 test samples had been classified correctly by NB, KNN and SVM respectively.

Specie	Naïve Bayes Classifier	KNN Classifier	SVM Classifier

Potato — Naïve Bayes Classifier

True Label \ Predicted Label	0	1	2
0	147	0	33
1	2	171	17
2	13	3	214

Potato — KNN Classifier

True Label \ Predicted Label	0	1	2
0	174	0	6
1	0	186	4
2	22	4	204

Potato — SVM Classifier

True Label \ Predicted Label	0	1	2
0	176	0	4
1	0	188	2
2	11	2	217

Grape — Naïve Bayes Classifier

True Label \ Predicted Label	0	1	2	3
0	161	33	7	1
1	46	196	0	4
2	35	0	173	0
3	21	7	1	170

Grape — KNN Classifier

True Label \ Predicted Label	0	1	2	3
0	178	17	3	4
1	17	225	1	3
2	2	0	205	1
3	9	1	1	188

Grape — SVM Classifier

True Label \ Predicted Label	0	1	2	3
0	188	11	2	1
1	11	234	0	1
2	3	0	204	1
3	3	0	1	195

Table 12. Confusion matrix

Table 13 shows the various performance metrics such as precision, recall, F1-score, sensitivity, specificity accuracy and Kappa score. From the table it can be observed that SVM classifier gives a maximum accuracy of 96.83% and a kappa score of 0.91 for diseased Potato leaf images.

Table 13. Performance metrics (a) Potato disease classification (b) Grape disease classification

Classifier	Class	Precision	Recall	F1-Score	Sensitivity	Specificity	Support	Accuracy	Kappa
NB	0: Early Blight	90.74%	81.67%	85.96%	81.67%	96.43%	180		
	1: Healthy	98.28%	90.00%	93.96%	90.00%	99.27%	190	88.67%	0.77
	2: Late Blight	81.06%	93.04%	86.64%	93.04%	86.49%	230		
KNN	0: Early Blight	88.78%	96.67%	92.55%	96.67%	94.76%	180		
	1: Healthy	97.89%	97.89%	97.89%	97.89%	99.02%	190	94.00%	0.90
	2: Late Blight	95.33%	88.70%	91.89%	88.70%	97.30%	230		
SVM	0: Early Blight	94.12%	97.78%	95.91%	97.78%	97.38%	180		
	1: Healthy	98.95%	98.95%	98.95%	98.95%	99.51%	190	96.83%	0.91
	2: Late Blight	97.31%	94.35%	95.81%	94.35%	98.38%	230		
(a)									
Classifier	Class	Precision	Recall	F1-Score	Sensitivity	Specificity	Support	Accuracy	Kappa
NB	0: Black Rot	61.22%	79.70%	69.25%	79.70%	84.38%	202		
	1: Esca (Black Measles)	83.05%	79.67%	81.33%	79.67%	93.43%	246	81.87%	0.76
	2: Healthy	95.58%	83.17%	88.95%	83.17%	98.76%	208		
	3: Leaf Blight	97.14%	85.43%	90.91%	85.43%	99.24%	199		
KNN	0: Black Rot	86.41%	88.12%	87.25%	88.12%	95.71%	202	93.10%	0.88
	1: Esca (Black Measles)	92.59%	91.46%	92.02%	91.46%	97.04%	246		

Continued on next page

|	*Table 13 continued*								
	2: Healthy	97.62%	98.56%	98.09%	98.56%	99.23%	208		
	3: Leaf Blight	95.92%	94.47%	95.19%	94.47%	98.78%	199		
SVM	0: Black Rot	91.71%	93.07%	92.38%	93.07%	97.40%	202		
	1: Esca (Black Measles)	95.51%	95.12%	95.32%	95.12%	98.19%	246	96.02%	0.94
	2: Healthy	98.55%	98.08%	98.31%	98.08%	99.54%	208		
	3: Leaf Blight	98.48%	97.99%	98.24%	97.99%	99.54%	199		
(b)									

For diseased Grape leaf images and healthy leaf images, an accuracy of 96.02% and a Kappa score of 0.94 (almost perfect agreement) were obtained. 10 -Fold Cross Validation was performed to evaluate the machine learning models Naïve Bayes, K-Nearest Neighbor and Support Vector Machines. Bias and Variance measures obtained are tabulated in Table 14 and it can be concluded from the results that SVM classifier best suits the data as both bias and variance have low values.

Table 14. Evaluation of machine learning models

Model	Potato		Grape	
	Bias	Variance	Bias	Variance
NB	0.450323	0.000015	0.573610	0.000007
KNN	0.205609	0.000028	0.219493	0.000029
SVM	0.122151	0.000009	0.140370	0.000015

The results obtained from previous works in the literature compared with the proposed work. It was found that the proposed work gives the highest classification accuracy of 96.02% for grape diseases and 96.83% for potato diseases as shown in Figure 14.
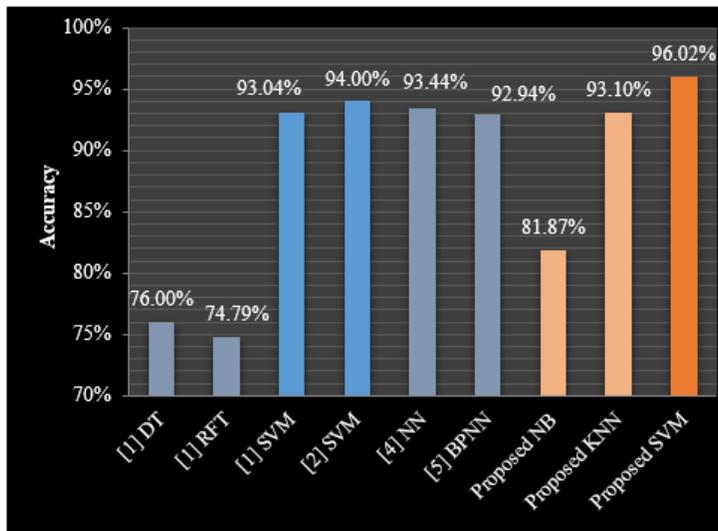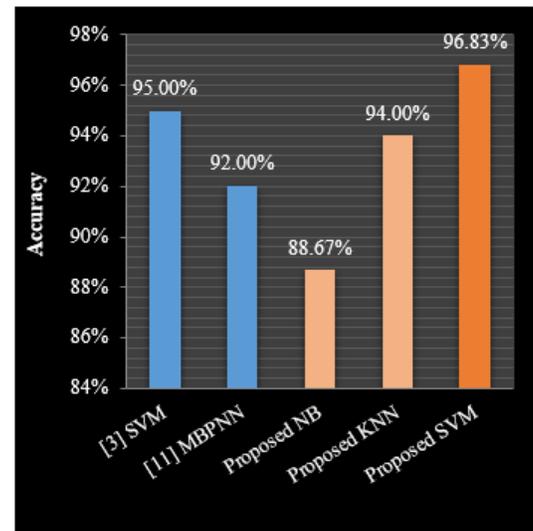
Fig 14. Comparative study of classification results (a) Grape diseases (b) Potato diseases

5 Conclusion

This study is focused on segmentation of lesion region and classifying plant diseases from plant leaf image using color, texture and histogram features. Naïve Bayes, KNN and SVM classifiers were tested and with regard to classification of diseases, SVM classifier gave the highest accuracy of 96.83% for Potato leaf images and 96.02% for Grape leaf images. A Kappa value of 0.91 and 0.94 for Potato and Grape species respectively indicates that there is a perfect agreement with the ground truth and predicted values.

For Potato plant, out of 180 Early Blight infected leaf images 147, 174 and 176 were classified correctly by Naïve Bayes, K Nearest Neighbor and Support Vector Machines respectively; 171, 186 and 188 Late Blight diseased leaves were predicted correctly out of 190 test samples by Naïve Bayes, K Nearest Neighbor and Support Vector Machines respectively; out of 230 healthy leaf images 214,204 and 217 were correctly predicted by Naïve Bayes, K Nearest Neighbor and Support Vector Machines respectively.

For Grape plant, out of 202 Black Rot infected leaf images 161, 178 and 188 were classified correctly by Naïve Bayes, K Nearest Neighbor and Support Vector Machines respectively; 196, 225 and 234 Esca (Black Measles) diseased leaves were predicted correctly out of 246 test samples by Naïve Bayes, K Nearest Neighbor and Support Vector Machines respectively; 173, 205 and 204 Leaf Blight diseased leaves were predicted correctly out of 208 test samples by Naïve Bayes, K Nearest Neighbor and Support Vector Machines respectively; out of 199 healthy leaf images 170,188 and 195 were correctly predicted by Naïve Bayes, K Nearest Neighbor and Support Vector Machines respectively.

The research work can be augmented by classifiers like Decision Trees and Neural Networks can be modelled and the results obtained can be compared with the proposed method. Future research directions could include classification of other species such as Tomato, Corn, Soyabean, Orange etc.

References

1) Jaisakthi SM, Mirunalini P, Thenmozhi D, Vatsala. Grape Leaf Disease Identification using Machine Learning Techniques. *Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019)*. 2019;p. 1–6. Available from: https://doi.org/10.1109/ICCIDS.2019.8862084.

2) Tarannum Z, Sankha BS, Nayak N, Smitha N, Rao A. Classification of Diseases in Grape Plants Using Multiclass Support Vector Machine. *International Journal of Emerging Research in Management & Technology*. 2017;6(5):250–254.

3) Kharde KP, Kulkarni HH. Grape Leaf Disease Detection using Embedded Processor". *International Research Journal of Engineering and Technology (IRJET)*. 2016;3(7):1078–1082.

4) Kakade RN, Ahire DD. Real Time Grape Leaf Disease Detection. *International Journal of Advance Research and Innovative Ideas in Education*. 2015;1(4):598–610.

5) Athanikar G, Bardar P. Potato Leaf Diseases Detection and Classification System". *International Journal of Computer Science and Mobile Computing*. 2016;5(2):76–88.

6) Islam M, Dinh A, Wahid K, Bhowmik P. Detection of Potato Diseases Using Image Segmentation and Multiclass Support Vector Machine. In: IEEE 30th Canadian Conference on Electrical and Computer Engineering(CCECE). 2017;p. 1–4. Available from: https://doi.org/10.1109/CCECE.2017.7946594.

7) Wang H, Li G, Ma Z, Li X. Image recognition of plant diseases based on backpropagation networks. In: and others, editor. 5th International Congress on Image and Signal Processing, Chongqing. 2012;p. 894–900. Available from: https://doi.org/10.1109/CISP.2012.6469998.

8) Padol PB, Yadav AA. SVM classifier based grape leaf disease detection. In: Conference on Advances in Signal Processing (CASP), Pune. 2016;p. 175–179. Available from: https://doi.org/10.1109/CASP.2016.7746160.

9) Pantazi XE, Moshou D, Tamouridou AA. Automated leaf disease detection in different crop species through image features analysis and One Class Classifiers. *Computers and Electronics in Agriculture*. 2019;156:96–104. Available from: https://dx.doi.org/10.1016/j.compag.2018.11.005.

10) Tiwari VM, Tarun G. Plant leaf disease analysis using image processing technique with modified SVM-CS classifier". *International Journal of Engineering & Management Technology*. 2017;5:11–17.

11) Ngugi CL, Abelwahab M, Abo-Zahhad M. Recent advances in image processing techniques for automated leaf pest and disease recognition - A review. *Information Processing in Agriculture*. 2020. Available from: https://doi.org/10.1016/j.inpa.2020.04.004.

12) Potato Disease - Early Blight. 2019. Available from: https://www.canr.msu.edu/resources/potato_diseases_early_blight_e2991.

13) Nolte P, Secor AG, Gudmestad CN. Wound healing, decay and chemical treatment of cut potato tuber tissue. *American Potato Journal*. 1987;64:1–9. Available from: https://dx.doi.org/10.1007/bf02853223.

14) Integrated Pest Management in Crop Production. 2020. Available from: http://vikaspedia.in/agriculture/crop-production/integrated-pest-managment/ipm-for-fruit-crops/ipm-strategies-for-grapes/grapes-diseases-and-symptoms.

15) Heckert NA, Filliben JJ, National Institute of Standards and Technology Handbook Series. Let Subcommands and Library Functions. In: and others, editor. NIST Handbook 148: Dataplot Reference Manual ;vol. 2. 2003.

16) Grape. . Available from: https://plantvillage.psu.edu/topics/grape/infos.

17) Plant Village dataset . . Available from: https://github.com/spMohanty/PlantVillage-Dataset.

18) Asaithambi S. How and When to Scale your Features. . Available from: https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e.

19) Chang CC, Lin CJ. LIBSVM. *ACM Transactions on Intelligent Systems and Technology*. 2011;2:1–27. Available from: https://dx.doi.org/10.1145/1961189.1961199.

20) Kumar N. Using Support Vector Machines Effectively. . Available from: https://neerajkumar.org/writings/svm.

21) . . Available from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

22) Bronshtein A. A Quick Introduction to K-Nearest Neighbors Algorithm. . Available from: https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7.

23) Jeyalakshmi S, Radha R. A novel approach to segment leaf region from plant leaf image using automatic enhanced grabcut algorithm. *An international Journal of Advanced Computer Technology*;8(11):3485–3493.

24) . . Available from: https://www.statisticshowto.datasciencecentral.com/univariate/.