Can Bipartite Graph Be Effective Tool for Clustering: A Self Organizing Map-based Automated Plant Disease Detection

Aditi Ghosh
University of Burdwan
Parthajit Roy
roy.parthajit@gmail.com

University of Burdwan

Research Article

Keywords: Singular Value Decomposition, Self Organizing Map, Bipartite Graph, Image Segmentation, Ray Tracing.

Posted Date: September 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3328263/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at International Journal of Wavelets, Multiresolution and Information Processing on February 1st, 2025. See the published version at https://doi.org/10.1142/S0219691325300014.

Can Bipartite Graph Be Effective Tool for Clustering: A Self Organizing Map-based Automated Plant Disease Detection

Aditi Ghosh[1] and Parthajit Roy[1*]

[1]The Department of Computer Science, The University of Burdwan, Purba Bardhaman, 713104, West Bengal, India.

*Corresponding author(s). E-mail(s): roy.parthajit@gmail.com;
Contributing authors: aditi_ghosh9@yahoo.com;

Abstract

An innovative approach for automated plant disease identification has been proposed in this study. The main contribution of this study is the introduction of a bipartite graph-based clustering technique that has been used for image segmentation, a feature extraction methodology using Self Organizing Map (SOM), and a ray tracing method. Numerous research works have already been done in this area with their respective merits and demerits. But this bipartite graph-based clustering for image segmentation, feature extraction using SOM and ray tracing technique has not been used in any of these studies as far as we are aware. The core idea behind this clustering technique is to represent similar spatial data points using a bipartite graph and then Singular Value Decomposition has been used on that graph for clustering. It is common to use SOM for clustering. However, in this study, SOM has been used for feature extraction. First, a spatial dependency matrix based on the pixel value of the gray image has been constructed using SOM. Then some statistical features have been computed from this matrix. Using the ray tracing method, the length of the most extended cluster i.e. the length of the most extended disease-affected patch, and the distribution of the clusters in the image have been computed. The accuracy of our model has been greatly enhanced using these features only. It has been experimented on disease-affected Grape leaf images taken from the Plant Village Dataset. This model outperforms state-of-the-art models which is shown in the result section. Not only that, the proposed

features produce better accuracy rather than using some existing features. This comparison has also been shown in the result section. The result has been validated using K-fold cross-validation. Last, but not the least, these features produce good accuracy using different classifiers also.

Keywords: Singular Value Decomposition; Self Organizing Map; Bipartite Graph; Image Segmentation; Ray Tracing.

1 Introduction

Normal growth, development, and productivity of a plant can be impeded due to various diseases. Certain conditions that are responsible for these diseases to occur are host plants, pathogens, and some supportive environmental conditions. Pathogens [1] are the microorganisms that can cause diseases. Plant pathogens can be Viruses, Bacteria, Fungi, Nematodes, etc. In this study, the species grape and its diseases Esca, Isariopsis, and Black Rot have been considered for classification.

Esca [2], a type of Grapevine Trunk Disease (GTD) is brought about by the fungi Phaeoacremonium aleophilum and Phaeomoniella chlamydospora. Interveinal striping is the main symptom of this disease. The lesions look like tiger strip patterns. The affected regions become dark red in color. Affected leaves become completely dry and drop in premature condition.

Isariopsis [3], a type of leaf spot, is brought about by the fungi Pseudocercospora vitis. Initially, these spots are tiny and appear on both surfaces of the leaf as irregular to angular chlorotic patches. Later on, these spots become purplish brown to black. The leaves which appear near the ground are affected severely by this disease.

During the hot and humid weather, a fungal disease called Black Rot [4] attacks grape vines severely. This disease is brought about by an ascomycetous fungus, Guignardia bidwellii. The lesions appear on the upper leaf surface with reddish-brown circular spots. Later on, these lesions become dark brown enclosed by a black margin. All these diseases and healthy leaves have been shown in Fig.1.

Before these diseases become severe, it is of utmost importance to identify them properly to prevent their spread into new regions. Generally, some laboratory-based techniques are used for this. Enzyme-Linked Immunosorbent Assay (ELISA), Immunofluorescent Staining, and Lateral Flow Immunoassays (LFIA) are the most common among them. However, these methods are too expensive because of the generation of monoclonal antibodies. Another drawback of these methods is low specificity. There are some DNA-based methods such as Polymerase Chain Reaction (PCR), nested PCR, etc. to identify pathogens. These methods are faster as well as cost-effective but need domain experts and specific equipment. Due to these limitations of using laboratory-based methods, the automated image processing-based method is the best

(a) Healthy leaf (b) Black Measles (c) Isariopsis (d) Black Rot

Fig. 1: (a) Healthy grape leaf. (b) Black Measles containing tiger strip pattern. (c) Red to brown spot in Leaf Blight. (d) Black Rot containing dark brown spots.

choice. It is cost effective but its accuracy is not very authentic as compared to the laboratory-based methods. Several research works are going on to improve accuracy. The Present research work proposes an image processing-based automated model for doing the same. The main focus of this study is the introduction of an innovative bipartite graph-based image segmentation technique and a feature extraction methodology using Self Organizing Map (SOM) and ray tracing method. Using these approaches, the accuracy of the model has been improved significantly.

2 Related Works

Jaisakthi et al. [5] worked on identifying grape diseases. They identified the same diseases that we have taken for classification. Different statistical features obtained from the Gray Level Co-occurrence Matrix (GLCM) were used as features. Discrete Wavelet Transform (DWT) was also used for feature extraction. They used different classifiers to measure the model's accuracy. The best result they obtained was from the Support Vector Machine (SVM) classifier with 93% test accuracy. In the research work of Hassan et al. [6], the same types of diseases were identified with 91.37% efficiency. Convolutional Neural Network (CNN) was used for the training and testing. Identification of the same diseases of grapes with the same dataset was done by Lin et al. [7]. Different CNN models were applied to the dataset. The best result they obtained was 86.29%. Ghosh et al. [8] proposed an automated model for the same using two species of potato and grape on the same dataset. A CNN architecture was built for disease classification. For potatoes, the model's efficiency was 87.47% and it was 91.96% for the grape. Ghosh et al. [9] also proposed an automated model for identifying apple leaf diseases. GLCM, Gabor Filter, and Local Binary Pattern (LBP) were used as features. Using ANN, they achieved 94.79% accuracy. Yu et al. [10] introduced an automated model for apple leaf diseases. Most relevant features from the disease-affected regions were extracted after performing image segmentation. Their model's efficiency was 89.4%. An automated model to recognize apple leaf diseases was also developed by Zhong et al. [11]. They

used multi-label classification, regression, and focus loss function and achieved 93.31%, 93.51%, and 93.71% efficiency respectively.

From the above discussion of the related works, it has been seen that each study has its respective merits and demerits. Some study uses high dimensional feature vector whereas in some research work the model's accuracy is not good. In Table 1, the details of the research works done by Jaisakthi et al. [5], Hassan et al. [6], Lin et al. [7], and Ghosh et al. [8] have been given with their drawbacks. So there is still a scope for improvements in this domain. This research work proposes such an automated model which uses less number of features while maintaining good accuracy. The main contributions of our study have been given in Table 2.

The remaining sections of this paper have been organized as follows. The details of the proposed methodology have been discussed in section 3. Section 4 describes the experimental setup and the experimental findings have been analyzed in section 5. The conclusion has been drawn in section 6.

3 Proposed Methodology

The proposed approach consists of the following steps. Each step has been discussed below in details.

- ***Data Augmentation***
- ***Background Elimination***
- ***Image Enhancement***
- ***Data Augmentation***
- ***Proposed Bipartite graph based clustering***
- ***Proposed Ray Tracing based method for feature extraction***
- ***Proposed SOM based feature extraction***
- ***Classification***

3.1 Data Augmentation

The dataset on which we are working contains only 423 healthy grape leaf images whereas the number of disease-affected leaves in each category is above 1000. Hence data augmentation has been done for better performance. It is a process of generating artificial images through various operations performed on the original images. In this study, some geometric transformations like flip, crop, rotate, and scaling have been applied to the existing images of the healthy leaves. After data augmentation, the total number of healthy leaves became 1692. The outcome of this step has been given in Fig. 2.

3.2 Background Elimination

Images that have been taken from the Plant Village Dataset [12], contain color background. For further processing, we need to eliminate this background to convert the images into scan-like images. Grab cut [13] algorithm has been applied to the images for background elimination. It was proposed by Rother

Table 1: Details of some existing research studies including the name of the dataset, disease types, features and classifiers. It also includes the accuracy and the drawbacks of those models.

Author	**Dataset Used**	**Disease Types**	**Features Used**	**Classifier & Accuracy**	**Drawbacks**
Jaisakthi et al. [5]	Plant Village Dataset.	Esca, Isariopsis & Black Rot of Grape.	GLCM texture features and DWT features.	SVM, 93%	1. Total 72 features were used. So the dimension of the feature vector is too high which leads to longer execution time. 2. The model's accuracy was not high.
Hassan et al. [6]	Plant Village Dataset.	Esca, Isariopsis & Black Rot of Grape.	Does not require any features.	CNN, 91.37%	1. In spite of using CNN, the model's accuracy was not adequate. 2. Resource requirement for CNN architecture is also a constraint.
Lin et al. [7]	Plant Village Dataset.	Esca, Isariopsis & Black Rot of Grape.	Does not require any features.	CNN, 86.29%	1. CNN requires more resources which is a barrier for implementing the model in mobile devices. 2. In-spite of using CNN architecture, the model's accuracy was poor.
Ghosh et al. [8]	Plant Village Dataset.	Esca, Isariopsis & Black Rot of Grape.	Does not require any features.	CNN, 91.96%	1. Resource constraint for CNN is the main issue. 2. Model's efficiency was not satisfactory.

Table 2: Details of the key contributions of the present study. One is for image segmentation and the other two are for feature extraction.

1. Introduction of a new bipartite graph-based clustering. It has been used to separate the disease-affected and unaffected regions from the leaf image.
2. Introduction of a feature extraction methodology using SOM. This method has been used to extract some important features from the images.
3. Introduction of a feature extraction technique using a method called Ray Tracing. Using this method, the longest length from the disease-affected regions and their distribution has been computed and the same is taken as a feature.

Fig. 2: Outcome of the Data Augmentation step. Original image and its augmented images using scaling, flip, rotation etc.

et al. in 2004. This algorithm is based on Graph cuts. To get the regional terms, the Gaussian Mixture Model (GMM) is developed from the color space of the foreground and background rectangular region. Euclidean distance between neighborhood pixels has been used to obtain the boundary term. These two terms have been used to form the energy function.

Let, the input image consist of n pixels where P denotes the set of all pixels on the image. Then, $p, q (p \in P, q \in P)$, the set of unordered pairs of neighboring elements in P is represented as N. Let B be a binary vector and $B = B_1, B_2, B_3, ...B_p, ...B_n$, where B_p represents the assignment of pixel p in set P and 0, 1 corresponds to the background and foreground respectively. The energy function of the graph is defined in equation 1 [13].

$$\alpha(B, c, \phi, P) = \beta(B, c, \phi, P) + \gamma(B, P) \tag{1}$$

where β and γ represent the regional and boundary term respectively. The regional term β is defined in equation 2 [13].

$$\beta(B, c, \phi, P) = \sum_{p \in P} \delta(B_p, c_p, \phi, P) \tag{2}$$

where δ and ϕ is defined in equation 3 and 4 [13].

$$\delta(B_p, c_p, \phi, P) = -\log \pi(B_p, c_p) + \frac{1}{2} \log \det \sum (B_p, c_p)$$

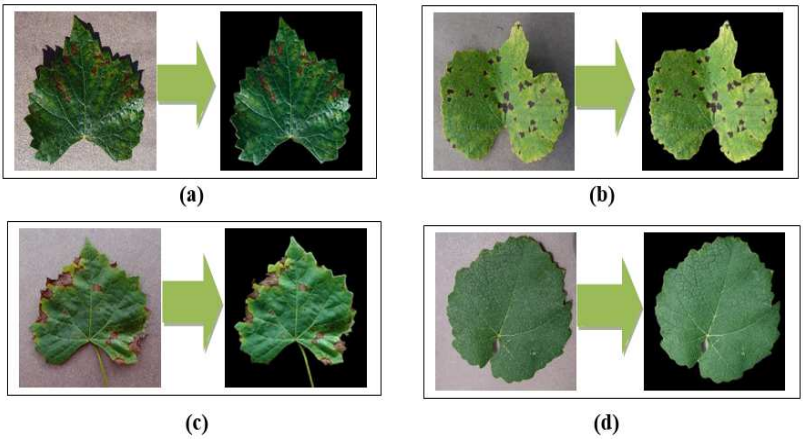


Fig. 3: Outcome of the Background Elimination step. The original images in each class and the corresponding images after Background Elimination using Grab Cut algorithm.

$$+\frac{1}{2}(I_p-\mu(B_p,c_p))^T\sum(B_p,c_p)^{-1}(I_p-\mu(B_p,c_p)) \tag{3}$$

$$\phi=\{\pi(B,c),\mu(B,c),\sigma(B,c)\},B=0,1;\ c=1,2,...,C. \tag{4}$$

where the weight, mean vector and covariance matrix of each Gaussian component is represented by π, μ and σ respectively. I_p is the pixel of point p. c_p represents the Gaussian component of pixel p. The boundary term γ is defined in equation 5 [13].

$$\gamma(B,P)=\zeta\sum_{\{p,q\}\in N}[B_p\neq B_q]\exp(-\eta\|I_p-I_q\|^2) \tag{5}$$

where η determines the contrast of the image and ζ is a constant which is generally taken as 50. The outcome of this procedure has been shown in Fig. 3.

3.3 Image Enhancement

To increase the performance of image analysis, contrast enhancement is used in image processing. The dataset that has been used in this study contains RGB color images with some background. After removing it from these original images it is required to enhance the contrast of these images for further processing. There are various techniques for image contrast enhancement. Out of these, the Contrast Limited Adaptive Histogram Equalization (CLAHE) [14] and Look Up Table (LUT) methods have been applied to the images for contrast improvement. Then Peak Signal to Noise Ratio (PSNR) was computed between the original and enhanced image for both methods. For every image,

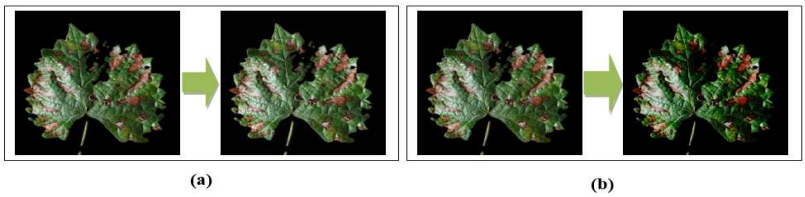


Fig. 4: Outcome of the image enhancement step. (a) Original image and enhanced image using CLAHE. The PSNR between the two is 33.7694 (b) Original image and enhanced image using LUT. The PSNR between the two is 22.66055.

both the methods have been applied and whichever of the methods is producing better PSNR has been considered. A higher PSNR value has been achieved for CLAHE rather than using LUT for a sample image shown in 4. Hence, the CLAHE method has been chosen for this particular image to enhance the image contrast. The outcome of applying CLAHE and LUT has been shown in Fig. 4(a) and (b) respectively.

CLAHE is a local contrast enhancement technique that relies on histogram modification. CLAHE is a form of Adaptive Histogram Equalization (AHE) that works locally and overcomes the limitations of global approaches. Instead of considering the whole image, a small portion of the image called tiles has been used. To obtain better results, the color space has been changed to LAB. Then the three components have been separated from each other and CLAHE has been applied to the Luminance channel as it outperforms equalizing all the three channels of an RGB image. Contrast-limited clipping has been applied on the histogram to prevent over-amplification of noise. In our study, after applying CLAHE, the PSNR between the original and the enhanced image shown in 4 is 33.7694 with a grid size of (8,8) and a clip limit of 0.1. PSNR between the same using LUT is 22.66055. So, for this particular image, CLAHE-based image enhancement has been taken. For other images whichever of the method is giving better PSNR has been chosen for that image.

3.4 Proposed Bipartite Graph based clustering

It is important to separate the disease-affected and unaffected regions from the leaves to obtain some of the proposed feature values. In this proposed approach we are going to propose a bipartite graph-based clustering technique that gives better results than K-Means. This clustering technique works on weighted undirected bipartite graphs using Singular Value Decomposition (SVD). It is based on a cutting-edge matching methodology.

The first step of our algorithm is to create a bipartite graph from the images. All the RGB pixel values have been extracted from the images. These are the real RGB pixel values with 3 dimensions. Now some random RGB pixel values of 3 dimensions have been generated between 0 to 255. These are the randomly generated RGB pixel values. Now the distance from each RGB

pixel value to each randomly generated RGB value has been computed using Euclidean distance. This is the concept of the bipartite graph as there is no edge connectivity inside these two sets. A pictorial representation of this concept has been given in Fig. 5. The RGB values are denoted as a_1, a_2, a_3,.....,a_n whereas the randomly generated RGB values are denoted as b_1, b_2, b_3,.....,b_m. Now a matrix A of dimension $m \times n$ has been created and the entries of this matrix are the corresponding weights obtained from the computed distances between two sets. Being a bipartite graph of (n+m) pixel values, the size of the adjacency matrix of the graph is also (n+m). So the adjacency matrix M can be defined by equation 6.

$$M = \begin{pmatrix} \Theta_{n\times n} & A^T_{m\times n} \\ A_{n\times m} & \Theta_{m\times m} \end{pmatrix} \tag{6}$$

In equation 6, $\Theta_{n\times n}$ and $\Theta_{m\times m}$ are null matrix of order $n \times n$ and $m \times m$ respectively. Further computation will be carried out on matrix A.

SVD [15] can be useful as it reduces the higher dimension of the data by finding the most important patterns in it. Hence it is also useful in feature extraction as it keeps only the most important singular values and the associated singular vectors. In this study, the SVD of matrix A has been computed. SVD decomposes this matrix into 3 matrices given in equation 7. From this U matrix k number of vectors have been taken. Then K-Means clustering [16] has been applied to these vectors. This approach is basically a kind of transformation. Applying this methodology before doing K-Means clustering gives better results. Fig. 6 depicts the outcome of Mean Shift segmentation, K-Means segmentation, and the proposed bipartite graph-based segmentation of two sample images. From this figure, it can be noted that the outcome of the proposed methodology produces better segmentation. This procedure has been described in algorithm 1.

$$A = U \times D \times V^T \tag{7}$$

Algorithm 1 Bipartite graph based image segmentation algorithm

Input: Color image I of size $n \times n$.

1: Take a set P_m number of RGB values of dimension 3.
2: Create a Bipartite Graph g consisting of m + n vertices and $m \times n$ edges.
3: The Adjacency matrix $M = \begin{pmatrix} \Theta_{n\times n} & A^T_{n\times m} \\ A_{m\times n} & \Theta_{m\times m} \end{pmatrix}$
4: Consider B from the adjacency matrix for SVD based computation.
5: $k \leftarrow numberofclusters$
6: Compute SVD of A $= U \times D \times V^T$
7: Take first k number of vectors from U and cluster them using k-means clustering.

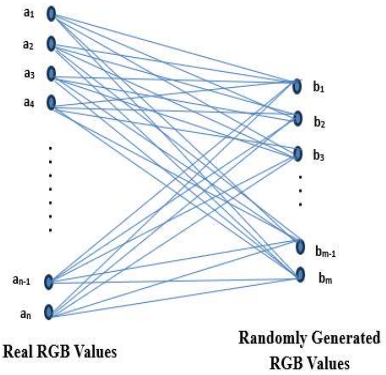


Fig. 5: Formation of the bipartite graph from the image. The set a_1, a_2, a_3,.....,a_n corresponds to the real RGB values of the image whereas the set b_1, b_2, b_3,.....,b_m corresponds to the randomly generated RGB values.

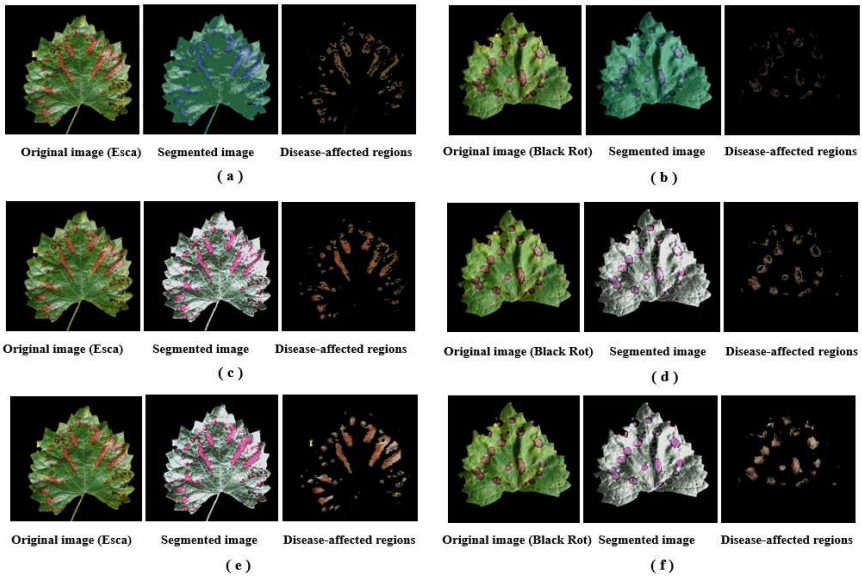


Fig. 6: Outcome of the image segmentation using Mean Shift, K-Means, and the proposed bipartite graph based clustering. (a)-(b) Original image, Segmented image and the disease affected regions of two sample images using Mean Shift. (c)-(d) Original image, Segmented image and the disease affected regions of two sample images using K-Means. (e)-(f) Original image, Segmented image and the disease affected regions of two sample images using proposed bipartite graph based clustering.

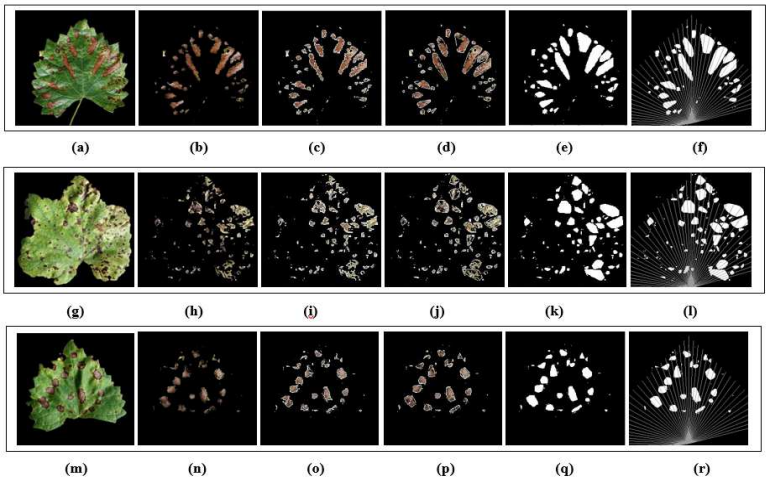


Fig. 7: Outcome of each step for computing the length of the longest cluster. Figure from (a)-(f), (g)-(l) and (m)-(r) represent original image, segmented image, image contour, convex hull, image filling and line drawing on esca, isariopsis and black rot respectively.

3.5 Proposed Ray Tracing based method for Feature Extraction

We have also proposed an innovative technique to compute the length of the longest disease-affected cluster and the distribution of the clusters has been proposed. The accuracy of the model using these features has been discussed in the result section. In addition to these features with some existing features, accuracy has been increased significantly.

It has been seen from our observation that the shapes of the affected regions of various types of diseases are different. In the disease type Esca, the shape of the disease-affected regions is like a wheel spoke and they are centered on the leaf. For the other types, the shape is not like that. Therefore, it can be a unique feature if it is computed properly. In this study, the longest length of the disease-affected region has been computed. In order to do this convex hull of the segmented images has been computed. Then all the segmented regions have been filled with white pixel values. From the original image centre of the leaf has been found using PCA. From this coordinate position, a line has been drawn using the Bresenham line drawing algorithm up to the leaf margin. All the foreground pixels that fall on the line have been counted. This value gives the length of that cluster. However, we are trying to find the length of the longest cluster. So, the line has been drawn again from the center position with a 5^0 interval and the foreground pixels on the line have been counted again. This process has been repeated with 5^0 interval. In this way, the length of the longest cluster has been computed easily.

Now, it is important to know the distribution of the clusters as it also gives some patterns. From this pattern, it can be understood how the clusters have been distributed in the leaf. Whether they are scattered throughout the image or grouped in some portion of the image can be known from this pattern. In order to keep this information as a feature we have divided the entire image into four zones. Then from the centre of the leaf that has already been computed using PCA, the number of clusters that fall in each zone has been counted. To do this we need to know the coordinate positions of these clusters. So each cluster has been wrapped within a rectangular box and their coordinate position has been obtained. These coordinate positions have been compared with the center position of the leaf obtained from PCA. So each cluster belongs to a zone and is counted as a feature. Using this approach total of five features have been obtained. One feature is the length of the longest cluster and the remaining four features are the number of clusters belonging to the four zones. While implementing this approach, some outcomes have been generated in the intermediate stage that has been given in Fig. 7.

3.6 Proposed SOM based method for feature extraction

A kind of ANN, called SOM, was developed by Kohonen [17]. The unsupervised learning approach is generally followed in SOM. Generally, it is used for clustering and dimensionality reduction. However, in this study, SOM is used to take out the most representative features from images. In order to do this first of all we need to create the input for SOM. It has been found that a pairwise count of pixel values can produce important information about the image.

Now, these values have been given as input to the SOM. SOM consists of two layers. One is the input layer and the other is the output layer. To train SOM, the size of the input should be (t,f), where t is the number of training samples and f represents the number of features. The weights of size (f,c) have been initialized. The number of clusters is represented by c. The winning vector will be updated after iterating over all the input. The weight update rule for SOM has been given in equation 8.

$$\psi_{i,j} = \psi_{i,j}(old) + \phi(t) \times (x_i^n - \psi_{i,j}(old)) \tag{8}$$

In this equation, ϕ represents the learning rate, i and j represent the training examples of i^{th} feature and the winning vector respectively and n represents the n^{th} training example from the input data.

In the dataset, all the images are of size 256×256. Therefore, to obtain pairwise pixel values, the number of rows and columns will be 255×255 and 2 respectively. After receiving these inputs, SOM will classify these values into their corresponding clusters. In our case, the number of clusters will be 255×255. The same pixel pair values will fall in the same cluster while different pixel pair values will fall into different clusters. From this clustering, the number of pixel pairs that fall in each cluster has been counted. These values have been stored inside a matrix S of size equal to the grid size of the

SOM. The size of this matrix is $D \times D$ where D is the number of gray levels. The entries in this matrix are denoted by $S(\alpha, \beta)$ which means the number of times the gray level α occurs with gray level β. From this matrix, various statistical features that contain relevant information have been computed. All of them have been discussed below.

Entropy Randomness among the pixel intensity distribution is known as the entropy. Maximum entropy can be obtained when all the entries in the S matrix are of the same magnitude. On the other hand, this value can be minimal when all the entries in the S matrix are different. To compute this value equation 9 has been used.

$$E = \sum_{\alpha=0}^{S_D-1} \sum_{\beta=0}^{S_D-1} -S(\alpha, \beta) \times \log S(\alpha, \beta) \tag{9}$$

Homogeneity The local homogeneity of the image can be obtained from the S matrix. The closeness of the elements of the S matrix to the diagonal entries is the measure of homogeneity. It can be computed from the equation 10.

$$H = \sum_{\alpha=1}^{S_D} \sum_{\beta=1}^{S_D} \frac{1}{1 + \|\alpha - \beta\|^2} S(\alpha, \beta) \tag{10}$$

Contrast The local variations present in an image are known as the contrast of the image. When the entries of the S matrix are concentrated away from the leading diagonal, there is a large variation present in the image. In this case, the contrast will be high. It can be computed from the equation 11.

$$C = \sum_{\alpha=0}^{S_D-1} \sum_{\beta=0}^{S_D-1} (\alpha - \beta)^2 S(\alpha, \beta) \tag{11}$$

Uniformity Uniformity among the gray level distribution in an image is an important feature. From equation 12 this value can be computed. If this value is 1, it means that the image is a constant image.

$$U = \sum_{\alpha=0}^{S_D-1} \sum_{\beta=0}^{S_D-1} S^2(\alpha, \beta) \tag{12}$$

Correlation The linear dependency in terms of the gray level between neighboring pixels is termed a correlation.

To extract some more features from the matrix S, SVD has been performed on this matrix. It decomposes the S matrix into three components as $S = U \times D \times V^T$. From this U matrix, the energy and entropy have been computed.

3.7 Classification

Now the feature vector has been set up. The next step is to classify them accordingly. At first, the ANN has been used for classification. We have taken only 13 features which is the number of neurons in the input layer. Due to the four categories, the number of neurons in the output layer is only 4. Using the trial and error method, the number of hidden layers and the number of neurons in each layer is 2 and 23 respectively. The learning rate of the ANN has been set up as 0.01 and the use of the sigmoid activation function has produced the highest accuracy using ANN. To prove the consistency of our feature vector different classifiers like SVM, K Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), Random Forest (RF), and Decision Tree (DT) [18] have been used to measure the efficiency. All outcome has been analyzed in the result section.

4 Experimental Setup

Before analyzing the findings of our study, it is of utmost importance to have a look at the dataset, parameters for measuring accuracy, use of existing features, and classification models. These details have been discussed in this section.

4.1 Dataset

Any automated model that is built using ML needs data to train and test the model. Plant Village Dataset [12] contains varieties of plant species images that are affected by different diseases. For example, it contains various disease-affected images of tomatoes, potatoes, grapes, apples, corns etc. In this study, grape diseases have been chosen for classification. All these images are RGB color images of size 256×256. Presentation of the dataset has been given in table 3. Total sample size and sample size for training and testing have been given. The train test ratio is 80:20.

4.2 Accuracy Measurement Parameters

The result of classification has to be measured in terms of some accuracy measurement metrics. It is a multi-class classification problem. A total of six accuracy measurement parameters have been taken to judge the efficiency of our model. These are precision, recall, f1-score, Cohen Kappa score, Matthew's correlation coefficient (MCC), and Jaccard score. Precision is the ratio between True Selection and all the selected instances. Recall can be stated as the ability of the model to find all the relevant cases in the dataset. The score ranges from 0 to 1 and is computed from the harmonic mean of precision and recall. These three have been defined in equation 13,14 and 15 [19].

$$Precision = \frac{TS}{(TS + FS)} \tag{13}$$

Table 3: Presentation of the data set. Total number of healthy and unhealthy grape leaf was 5332.

Disease name	Image samples for each class	Sample size	Sample size for training	Sample size for testing
Isariopsis		1076	743	333
Black Rot		1180	970	210
Esca		1384	1149	235
Grape Healthy		1692	1403	289

$$Recall = \frac{TS}{(TS + FR)} \tag{14}$$

$$F1\_Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{15}$$

where TS, FS and FR are the True Selection [19], False Selection [19] and False Rejection [19] respectively.

Cohen Kappa [20], a statistical parameter is used to measure the inter-rater as well as intra-rater reliability for qualitative data. It is denoted by κ and is more robust compared to others. This is defined in equation 16.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{16}$$

where p_0 represents the observed proportional agreement between actual and predicted values. This can be generated from the confusion matrix by taking the ratio between the sum of the diagonal and non-diagonal entries. The expected agreement is represented by p_e.

MCC [21] is one of the best performance measurement metrics that establishes the accuracy of a classifier. MCC for multi-class classification is defined in equation 17 [22].

$$MCC = \frac{p \times n - \sum_c^C b_c \times a_c}{\sqrt{(n^2 - \sum_c^C {b_c}^2) \times (n^2 - \sum_c^C {a_c}^2)}} \tag{17}$$

where,
c represents the class from 1 to C.
n denotes the number of samples.
p denotes the number of correctly predicted samples.

a_c represents the number of times class c truly occurred.
b_c represents the number of times class c was predicted.

Jaccard score [23] is also an important performance measurement metric. In the case of multi-class classification, this score is computed for each class using micro, macro, and weighted averages. The micro average computes this score globally by taking into consideration the total TP, FN, and FP. In the case of the macro average, the unweighted mean of the score has been used after computing the Jaccard score separately for each class. For the weighted average, it gives the weighted mean of the score. The weights are the number of samples in each class.

4.3 Classification Models

To test the effectiveness of our model, we have taken some existing standard classifiers. These are DT, GNB, KNN, SVM, RF, and ANN [18]. Existing studies [18] also used these classifiers. Accuracy obtained from all these classifiers has been given in the result section. It can be seen that for all the classifiers the accuracy is significant and close to each other. This happens because of our accurate feature set.

4.4 K Fold Cross Validation

The result has been validated using K fold cross validation [24]. Here, the value of K has been set as 5. The results obtained from different folds using different classifiers have been given in table 5 and the mean score obtained from different classifiers has been shown in Fig. 14.

5 Experimental Findings

In this section, the importance of our features has been justified using the results obtained from the experimental study. The justification has been explained from three perspectives. Firstly, the model has been compared with the state-of-the-art models. Secondly, a comparison has also been made between our feature set and other existing features. Thirdly, it has also been shown that these features produce good results irrespective of the classifiers. Before discussing all these points, we need to analyze the result that has been obtained using our feature set. Using the features obtained from the ray tracing method, the accuracy of the model reaches up to 72.78%. From the SOM extracted features, the efficiency is 82.56%. Now, combining these two feature sets the accuracy is increased and reaches up to 94.11%. Finally, the SVD extracted features are added to this feature set which enhances the efficiency up to 95.22%. A pictorial representation has been given in Fig 8.

Now, let us analyze the efficiency in terms of precision, recall, and f1-score given in table 4. It has been seen from the table that all three values are 100% in the case of Isariopsis. It signifies that all the affected leaves of this category have been identified from others without any error. For Black Rot also, these

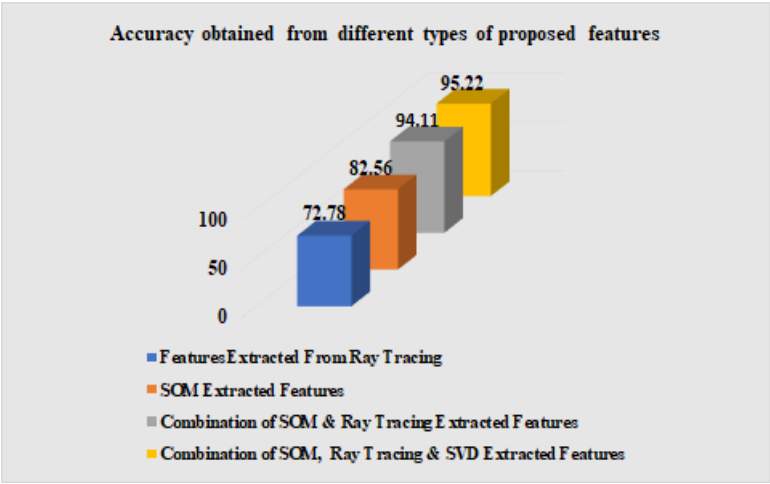


Fig. 8: Efficiency obtained using different feature combinations from the proposed features.

three values are greater than 96% approximately. For the other two categories, these values are also good. This can also be verified from the confusion matrix shown in Fig. 9. Both the FP and FN of Isariopsis are 0. FP and FN are also low for Black Rot. It signifies that these two categories have been classified properly compared to the others.

Table 4: Details of three performance measurement metrics Precision, Recall and F1-Score in each category.

Disease name	Image sample for each class	Precision	Recall	F1 score
Esca (Black Measles)		0.92	0.89	0.91
Black Rot		0.96	0.97	0.97
Isariopsis (Leaf Blight)		1.00	1.00	1.00
Healthy leaf		0.91	0.93	0.92

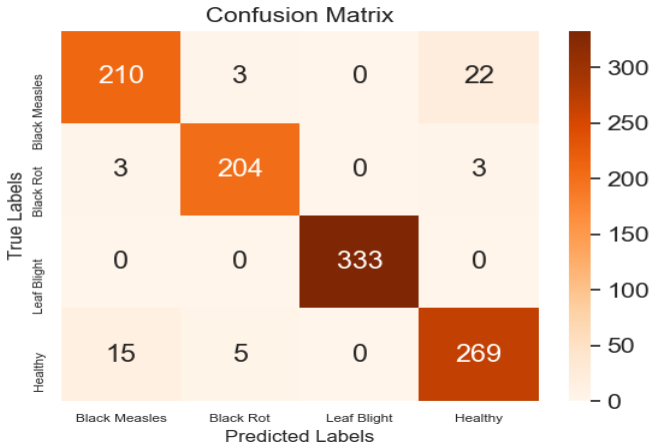


Fig. 9: Confusion matrix of our model corresponds to the Black Measles, Black Rot, Leaf Blight and Healthy leaf.

5.1 Comparison with other progressive models

Now we will establish the importance of the proposed feature set from the first perspective i.e. comparison with state-of-the-art models. In Fig. 10, our model has been compared with four progressive models. From the pictorial representation of Fig.10, it has been seen that there are three models developed by Lin et al., Hassan et al., and Ghosh et al. using CNN with efficiency 86.29%, 91.37%, and 91.96% respectively and one model developed by Jaisakthi et al. got the efficiency 93%. The efficiency of our proposed model is 95.22% and it outperforms among these four models.

5.2 Comparison with other feature set

To show the relevance of our feature set from the second perspective, let us compare it with other features using different classifiers. The features that have been taken for comparison are SURF, HOG, and LBP [25]. All these three are feature descriptors with high dimensions. Using only SURF, the accuracy is just 60.77% obtained from SVM. Similarly, using HOG and LBP [25], the accuracy obtained from SVM are 75.67% and 86.16% respectively. A line diagram of it has been shown in Fig. 11(a).

The proposed features have been compared with SURF, HOG, and LBP using ANN. The accuracy obtained from these features is 61.65%, 70.86%, and 85.78% respectively while the accuracy of the proposed features is 94.65%.

Using KNN, the accuracy produced from these three features is 55.04%, 72.54%, and 83.11% respectively. But for the proposed features, the efficiency becomes 92.03%.

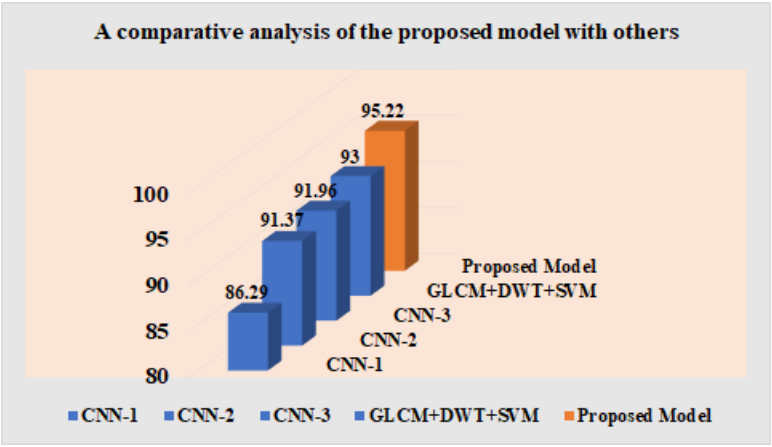


Fig. 10: A comparative study of our model with CNN-1 [7], CNN-2 [6], CNN-3 [8], and GLCM+DWT+SVM [5]. All these studies used the same kind of grape diseases from the Plant Village Dataset.

From RF, the efficiency achieved from the three feature descriptors is 53.82%, 70.59%, and 81.67% respectively. But using the proposed features, the efficiency becomes 92.59%.

Using GNB and DT, the accuracy has been measured for these three features and proposed features. For GNB, these three feature descriptors produce 52.87%, 69.74%, and 79.88% accuracy whereas for proposed features, it is 87.44%. Finally, from DT, the accuracy becomes 51.66%, 68.98%, and 78.09% for the three feature descriptors. The proposed features produce 88.65% accuracy using DT. All these comparisons have been shown using a line diagram in Fig. 11 (b), (c), (d), (e), and (f). It has been seen from these figures that the proposed feature set outperforms all three feature descriptors for all the classifiers used. Not only that, the dimension of the proposed feature set is very low compared to the others.

5.3 Comparison of different accuracy measurement metrics using different classifiers

The third perspective to show the efficiency of the proposed feature set is the usage of different classifiers. A graphical representation to show the overall accuracy obtained from the different classifiers has been given in Fig. 13. It has been seen that the highest accuracy has been achieved from SVM. In this case, the overall accuracy is 95.22%. More than 92% overall accuracy has been achieved using ANN, KNN, and RF. More than 87% accuracy has been achieved using GNB and DT.

The graphical representation of other accuracy measurement metrics using different classifiers has also been given in Fig. 12 (a), (b), (c), (d), (e), and (f). Let us analyze the Cohen-Kappa value obtained from the different classifiers.

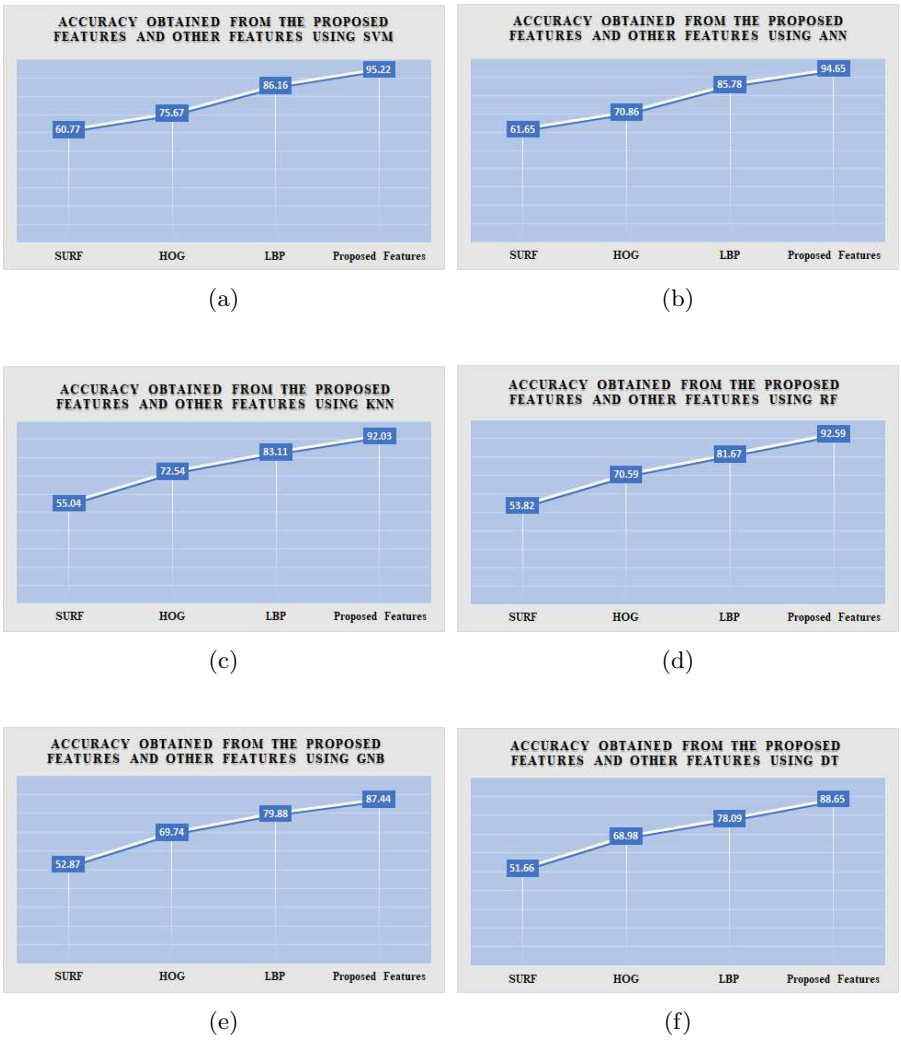
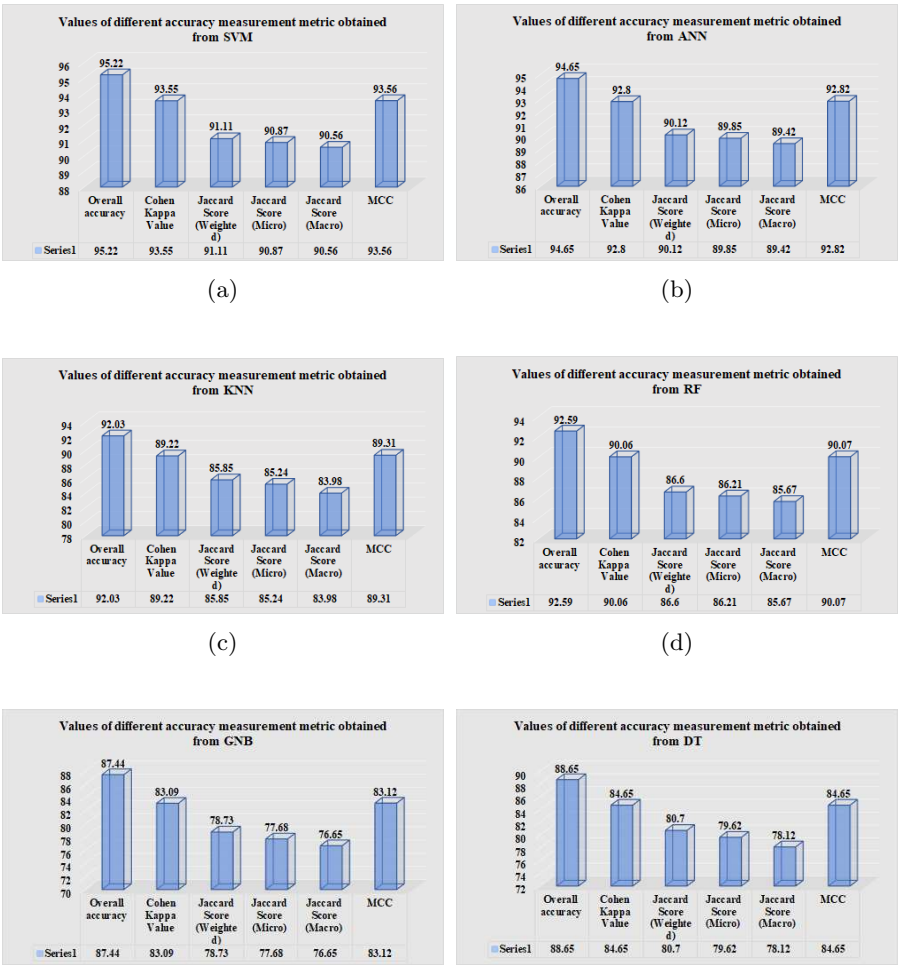


Fig. 11: A comparative study of the proposed features with SURF, HOG and LBP using (a) SVM, (b) ANN, (c) KNN, (d) RF, (e) GNB, and (f) DT.

Generally, a Cohen Kappa value greater than 81% is considered to be an almost perfect agreement between two raters. In our study, the Kappa value for ANN, SVM, KNN, RF, GNB and DT are 92.8%, 93.55%, 89.22%, 90.06%, 83.09% and 84.65% respectively which is given in Fig.12 (a), (b), (c), (d), (e), and (f). So, the Cohen-Kappa value of our model is also good for different classifiers.

The Jaccard score is also good for our model. The three Jaccard scores weighted, micro, and macro for the SVM classifier are 91.11%, 90.87%, and 90.56% respectively. For all the classifiers, these three values have been shown



(a) (b)

(c) (d)

(e) (f)

Fig. 12: Values of different performance measurement metric i.e. MCC, Jaccard Score(Macro), Jaccard Score (Micro), Jaccard Score (Weighted), Cohen Kappa, and Overall accuracy using (a) SVM, (b) ANN, (c) KNN, (d) RF, (e) GNB, and (f) DT.

in Fig.12 (a), (b), (c), (d), (e), and (f). MCC using ANN and SVM are 92.82% and 93.56% respectively. For all other classifiers also, it gives good results. All the values have been shown in Fig.12 (a), (b), (c), (d), (e), and (f). So, for all classifiers, all the performance measurement metrics produce significant accuracy, which shows the importance of the proposed feature set. This thing establishes that our computed features are classifier-independent. That means

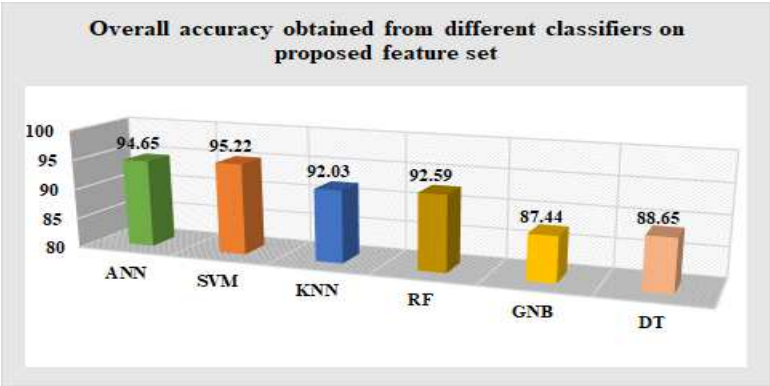


Fig. 13: A comparative analysis of accuracy using the proposed feature set obtained from ANN, SVM, KNN, RF, GNB, and DT. Highest accuracy obtained is 95.22% using SVM.

the accuracy we have achieved that is not for the classifier but for our computed features.

We have established the relevance of our feature set from all three perspectives i.e. in comparison with other models, comparison with other stand-alone features, and irrespective of the classifier. Not only that, good performance using different parameters has also been achieved using a very small number of features. It also eliminates the drawback of using feature descriptors i.e. the curse of dimensionality.

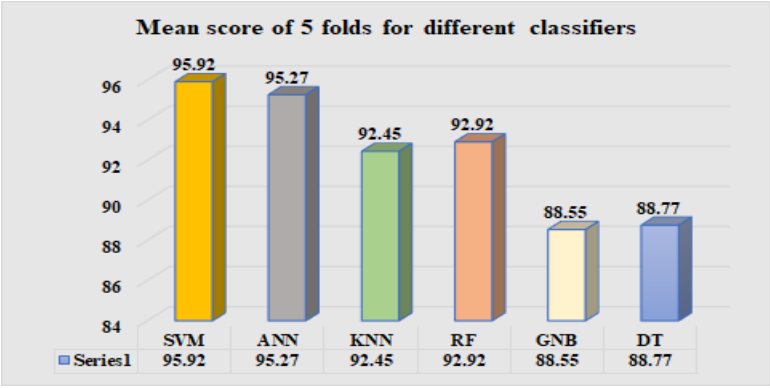


Fig. 14: The mean score of 5 folds obtained from K fold cross validation for different classifiers.

Table 5: Values of cross validation scores obtained from different folds of K fold cross validation for different classifiers. Here the value of K has been set as 5.

Classifiers	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
SVM	95.23%	96.68%	94.71%	95.38%	97.59%
ANN	94.33%	95.47%	96.82%	95.56%	94.18%
KNN	92.16%	93.25%	91.08%	92.51%	93.27%
RF	93.71%	92.61%	92.19%	93.43%	92.64%
GNB	88.98%	87.47%	89.72%	90.33%	86.26%
DT	89.81%	87.91%	88.78%	86.69%	90.67%

5.4 K fold cross validation

To test the effectiveness of any ML model, it is of utmost importance to perform data validation. K fold cross validation [24] technique has been used here to prove the robustness of the proposed model. The outcome obtained from different folds of the K fold cross-validation for different classifiers has been given in table 5. The mean score obtained from 5 different folds for different classifiers has been shown in Fig. 14.

5.5 Analysis of the proposed feature set using pair plot, train-test accuracy curve, and train-test loss curve

Now, let us analyze the proposed feature set by drawing a pair plot of the data. A pair plot has been generated on our feature set shown in Fig. 15. We can visualize the clusters from this pair plot. At least two separate clusters have been visualized from each pair. The formation of three clusters has also been visualized in some pairs. As LBP is a good texture feature, we have also generated a pair plot to visualize the relationship among these data. It is given in Fig. 16. We can see that in each pair, the formation of at least two clusters is not visible always. If we compare these two pair plots, it can be stated that the formation of clusters is obviously better in our feature set rather than LBP. Therefore we are getting better results using our features than LBP.

Fig. 17 depicts the curve of train-test accuracy with the number of epochs. It has been seen that there is no problem of overfitting. A curve of train-test loss with the number of epochs has also been shown in Fig. 18.

6 Conclusion

This study deals with an approach for building a classification model to recognize plant leaf diseases. Firstly, a bipartite graph-based clustering approach to separate disease-affected and unaffected parts has been proposed. In the feature extraction step, this study proposes two approaches to computing relevant features. These are using SOM and the ray tracing method. The results obtained using these approaches give better results compared to the state-of-the-art models. In comparison with some stand-alone features, the proposed feature set produces better results. Not only that, using different classifiers

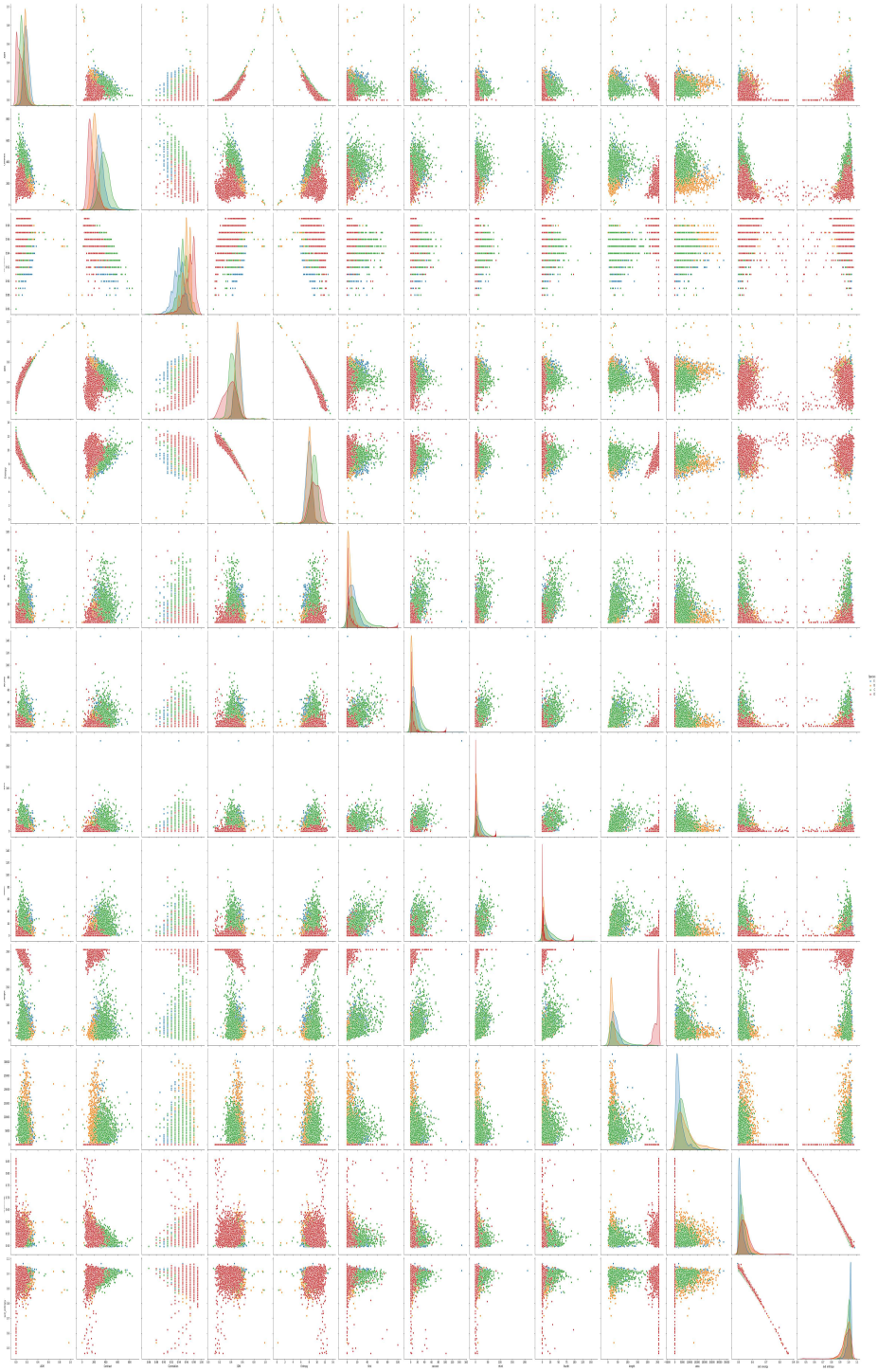


Fig. 15: Pair plot of the proposed features computed from the Plant Village dataset on grape disease. Formation of the clusters from the computed data points are visible better than LBP.

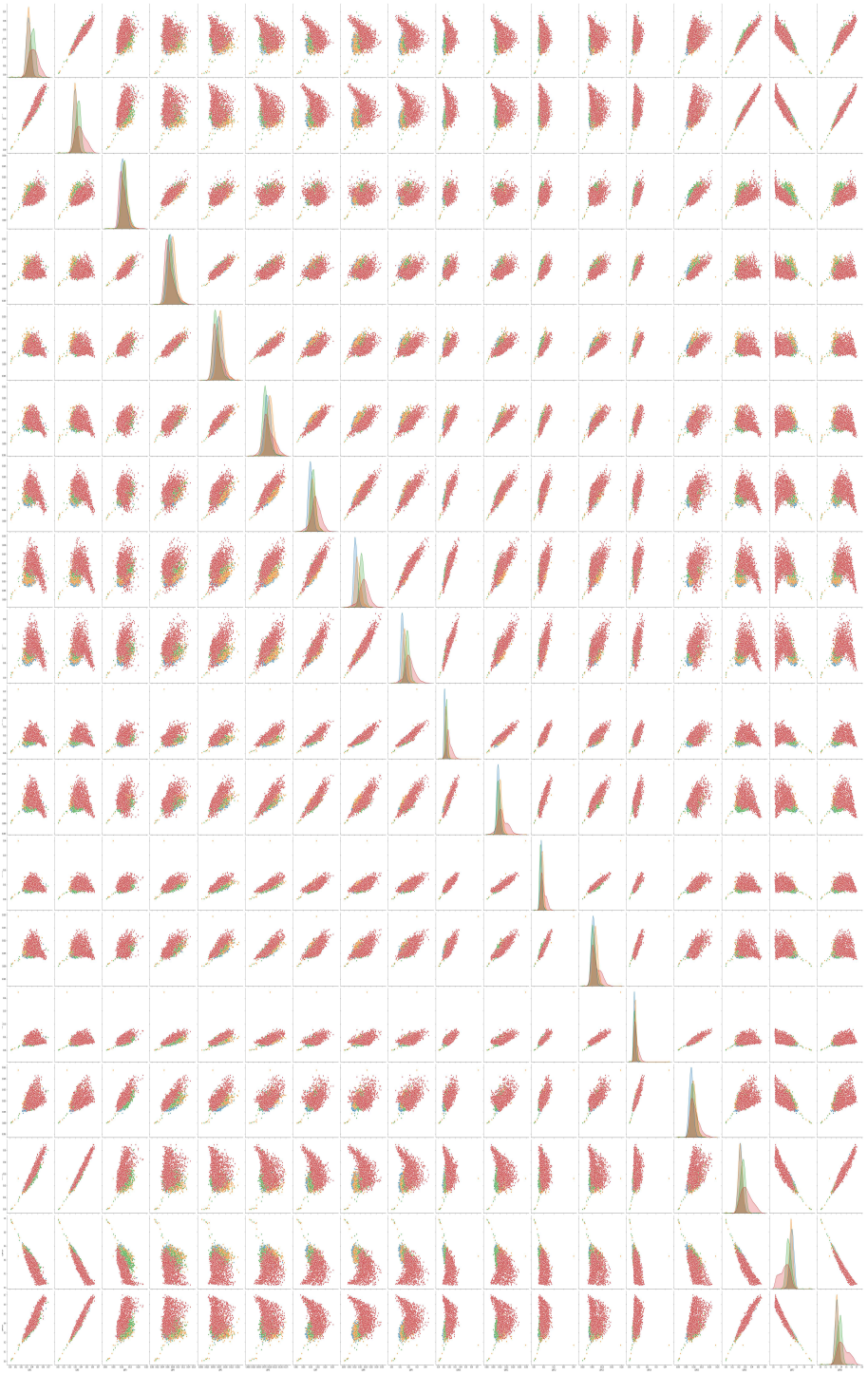


Fig. 16: Pair plot of LBP features computed from the Plant Village dataset on grape disease. But the formation of different clusters from the data points are not visible properly.

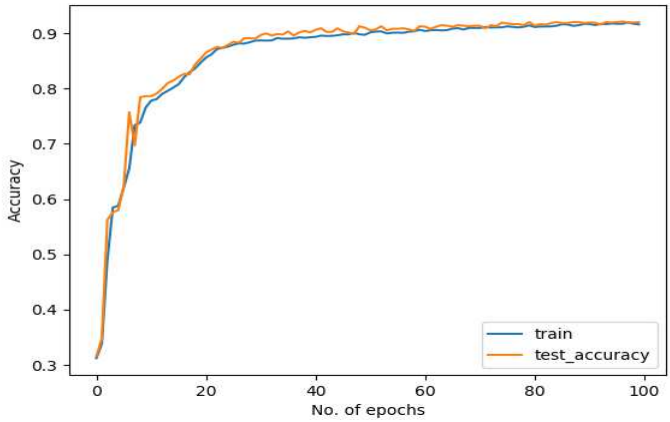


Fig. 17: The curve represents the train test accuracy. The train test accuracy is increased simultaneously with number of epochs. There is no problem of over fitting.

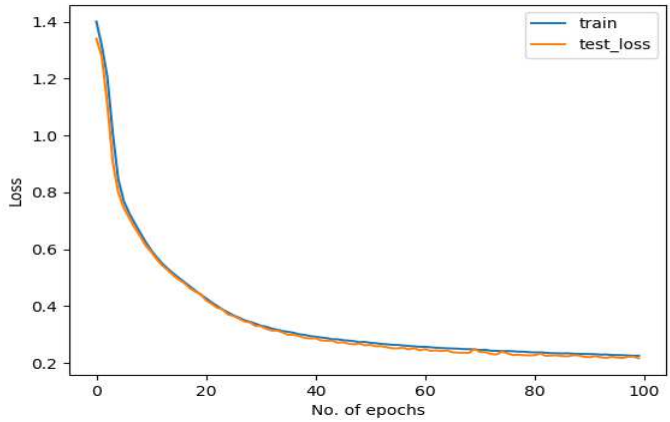


Fig. 18: The curve represents the train test loss. The train test loss is decreased simultaneously with number of epochs.

also the suggested approach produces a consistent accuracy. All these results have been analyzed in the result section. One of the biggest advantages of the model is that by using a very small number of features, significant accuracy has been achieved. All these things together establish the relevance of our model. The future scope of improvement of our study is to increase the accuracy more with some other set of features.

Author Contributions

Both the implementation and writing have been done by the first author. The conceptual part of the study has been contributed by the second author.

Funding

The present study does not have any kind of funding.

Data Availability

We don't have any available data for this research study.

Declarations

Conflict of interest

This research work is free from any conflict of interest for both authors.

Ethical approval

This research work is free from involving any studies with human participants or animals.

References

[1] Sakshi Tewari and Shilpi Sharma. Chapter 27 - molecular techniques for diagnosis of bacterial plant pathogens. In Surajit Das and Hirak Ranjan Dash, editors, *Microbial Diversity in the Genomic Era*, pages 481–497. Academic Press, 2021,http://dx.doi.org/https://doi.org/10.1016/B978-0-12-814849-5.00027-7.

[2] Giovanni Bortolami, Elena Farolfi, Eric Badel, Regis Burlett, Herve Cochard, Nathalie Ferrer, Andrew King, Laurent J Lamarque, Pascal Lecomte, Marie Marchesseau-Marchal, Jerome Pouzoulet, Jose M Torres-Ruiz, Santiago Trueba, Sylvain Delzon, Gregory A Gambetta, and Chloe E L Delmas. Seasonal and long-term consequences of esca grapevine disease on stem xylem integrity. *Journal of Experimental Botany*, 72(10):3914–3928, 03 2023,http://dx.doi.org/https://doi.org/10.1093/jxb/erab117.

[3] F. Mehmet. Occurrence of isariopsis leaf spot or blight of vitis rupestris caused by pseudocercospora vitis in turkey. 2017.

[4] Márton Szabó, Anna Csikasz-Krizsics, Terézia Dula, Eszter Farkas, Dóra Roznik, Pál Kozma, and Tamás Deák. Black rot of grapes (guignardia bidwellii)—a comprehensive overview. *Horticulturae*, 9:130, 01 2023,http://dx.doi.org/https://doi.org/10.3390/horticulturae9020130.

[5] S.M. Jaisakthi, P. Mirunalini, D. Thenmozhi, and Vatsala. Grape leaf disease identification using machine learning techniques. In *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pages 1–6, http://dx.doi.org/10.1109/ICCIDS.2019.8862084.

[6] Moh. Arie Hasan, Dwiza Riana, Sigit Swasono, Ade Priyatna, Eni Pudjiarti, and Lusa Indah Prahartiwi. Identification of grape leaf diseases using convolutional neural network. *Journal of Physics: Conference Series*, 1641(1):012007, nov 2020,http://dx.doi.org/10.1088/1742-6596/1641/1/012007.

[7] Jianwu Lin, Xiaoyulong Chen, Renyong Pan, Tengbao Cao, Jitong Cai, Yang Chen, Xishun Peng, Tomislav Cernava, and Xin Zhang. Grapenet: A lightweight convolutional neural network model for identification of grape leaf diseases. *Agriculture*, 12(6), 2022,http://dx.doi.org/10.3390/agriculture12060887.

[8] Aditi Ghosh and Parthajit Roy. *AI Based Automated Model for Plant Disease Detection, a Deep Learning Approach*, pages 199–213. 05 2021,http://dx.doi.org/10.1007/978-3-030-75529-4_16.

[9] Aditi Ghosh and Parthajit Roy. A leaf image-based automated disease detection model. In Mukesh Saraswat, Harish Sharma, K. Balachandran, Joong Hoon Kim, and Jagdish Chand Bansal, editors, *Congress on Intelligent Systems*, pages 863–875, Singapore, 2022,http://dx.doi.org/10.1007/978-981-16-9416-5_63. Springer Nature Singapore.

[10] Hee-Jin Yu and Chang-Hwan Son. Leaf spot attention network for apple leaf disease identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020,http://dx.doi.org/https://doi.org/10.1109/CVPRW50498.2020.00034.

[11] Yong Zhong and Ming Zhao. Research on deep learning in apple leaf disease recognition. *Computers and Electronics in Agriculture*, 168:105146, 2020,http://dx.doi.org/10.1016/j.compag.2019.105146.

[12] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016,http://dx.doi.org/10.3389/fpls.2016.01419.

[13] Zhaobin Wang, Yongke Lv, Runliang Wu, and Yaonan Zhang. Review of grabcut in image processing. *Mathematics*, 11:1965, 04 2023,http://dx.doi.org/https://doi.org/10.3390/math11081965.

[14] R.A. Manju, G. Koshy, and Philomina Simon. Improved method for enhancing dark images based on clahe and morphological reconstruction. *Procedia Computer Science*, 165:391–398, 2019, http://dx.doi.org/https://doi.org/10.1016/j.procs.2020.01.033. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

[15] Steven Brunton and J. Kutz. *Singular Value Decomposition (SVD)*, pages 3–46. 02 2019,http://dx.doi.org/https://doi.org/10.1017/9781108380690.002.

[16] Sukhvir Kaur, Shreelekha Pandey, and Shivani Goel. Plants disease identification and classification through leaf images: A survey. *Archives of Computational Methods in Engineering*, 26, 01 2018,http://dx.doi.org/10.1007/s11831-018-9255-6.

[17] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990,http://dx.doi.org/https://doi.org/10.1109/5.58325.

[18] Kaur Sukhvir, Pandey Shreelekha, and Goel Shivani. *Plants Disease Identification and Classification Through Leaf Images: A Survey*, pages 3–46. 2019,http://dx.doi.org/https://doi.org/10.1007/s11831-018-9255-6.

[19] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006,http://dx.doi.org/10.1016/j.patrec.2005.10.010. ROC Analysis in Pattern Recognition.

[20] A.S. Kolesnyk and N.F. Khairova. Justification for the use of cohen's kappa statistic in experimental studies of nlp and text mining. *Cybernetics and Systems Analysis*, 2022,http://dx.doi.org/https://doi.org/10.1007/s10559-022-00460-3.

[21] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14, 02 2021,http://dx.doi.org/https://doi.org/10.1186/s13040-021-00244-z.

[22] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multiclass classification: an overview, 08 2020.

[23] Sam Fletcher and Md Islam. Comparing sets of patterns with the jaccard index. *Australasian Journal of Information Systems*, 22, 03 2018,http://dx.doi.org/https://doi.org/10.3127/ajis.v22i0.1538.

[24] Kaushika Pal and Biraj. V. Patel. Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 83–87, 2020,http://dx.doi.org/https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00016.

[25] Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Local binary patterns and its application to facial image analysis: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 41:765–781, 11 http://dx.doi.org/10.1109/TSMCC.2011.2118750.