Multi-Class Crops Plant Leafs Classification Using Machine Learning Techniques

Hanif Khan Pathan[#1], DrDhanrajVerma[*2]

[#1]CSE Department,Dr APJ Abdul Kalam University

Indore Dewas bypass road Indore,India

1 hanifkp20@gmail.com

2 dhanrajmtech@gmail.com

#2 CSE department,Dr APJ Abdul Kalam University

Indore Dewas bypass road Indore, India

Abstract – in this paper, we introducing a plant disease detection system. The aim is to use image classification techniques based on Machine Learning (ML). The proposed plant disease detection technique has applied in three phases, feature selection, learning, and testing. In feature selection we use color, texture and shape features of image, which are used in various Image Information Retrieval (IR) technique. In order to recover shape features two segmentation techniques K-Means and Fuzzy c Means (FCM) algorithm is used. Next for the texture analysis LBP (Local Binary pattern) has used. Further for color features, the Color Grid Movement (CGM) has considered. In addition, to prepare effective features, different combinations of features have also been prepared. These combinations are used for image classification. The Support Vector Machine (SVM) and Artificial Neural Network (ANN) has considered for classification task. The experimentations with a popular image dataset Plant Village has been used for simulating the effectiveness of the proposed model. However, in this work have used limited number of crops namely corn, grapes and apple. After implementation of the crops disease detection model the experimental performance of all the combinations of features and classifiers has been measured. The Accuracy, error rate, memory and time consumption has considered as criteria to be select most appropriate model for plant disease detection model. The results show the combination of FCM, LBP and CGM outperform as compared to other combinations. Additionally, when this combination is used with the SVM and ANN, than ANN demonstrate higher accuracy in detection of disease in plant. Finally, the future extension of the proposed work has also been discussed.

Keywords: Image classification, Multi-class classification, image feature selection, crop images, leaf images, performance analysis*;*

I. INTRODUCTION

In India, agriculture activities act as the main source of livelihood for more than 74% of the population. That offers employment to 52% of the labors. It contributes 14 -15% in the GDP [1]. The production and income from the crops is significantly affecting the sustainability of farmers. In recent years a significant development and transformation in agriculture has been observed. The scientific and technical techniques helps in improving the productivity of crops and reduce the cost of farming by controlling the wastages, and preserving from the infection and disease. Such kind of farming is call smart farming, where the machine learning and sensor technologies are employed in farms to improve productivity, quality and income from the agriculture business.

Among various modules of smart farming the disease detection and prevention has a significant place. A number of techniques and methods are contributed by the researchers and engineers for effective disease detection in plant leaf images [2]. The leaf based analysis is useful in early stage disease detection. Therefore, a number of contributions we have found based on Convolutional Neural Networks (CNN) [3], Support Vector Machine (SVM) [4] and image segmentation [5]. By motivation of these articles in this paper, an efficient and accurate ML methods for plant disease detection using leaf image classification has proposed. Thus, first we involve the investigation of different features selection technique used for image classification.

In this context, some essential techniques are identified additionally two classification techniques are considered for exploring most optimal solution. Therefore, for shape features k-means and FCM clustering techniques are used. Additionally, for texture analysis, the LBP has implemented. Finally, the color features using CGM (color grid movement) is used as feature extraction. This section provides the overview of the application area and direction of the investigation. Additionally, an overview of the work is also provided. In next section the overview of the proposed disease detection model has been discussed. Additionally, explain the different components of the model. Next the experimental results are discussed and, the future work has proposed.

II. PROPOSED MODEL

In this paper we investigate effect of different feature selection techniques in classification accuracy. The aim is to identify the suitable combination of feature selection and ML algorithms to prepare the required plant disease detection model, by analyzing the plant image leaf classification. An overview of the required ML model for this task is demonstrated in figure 1.

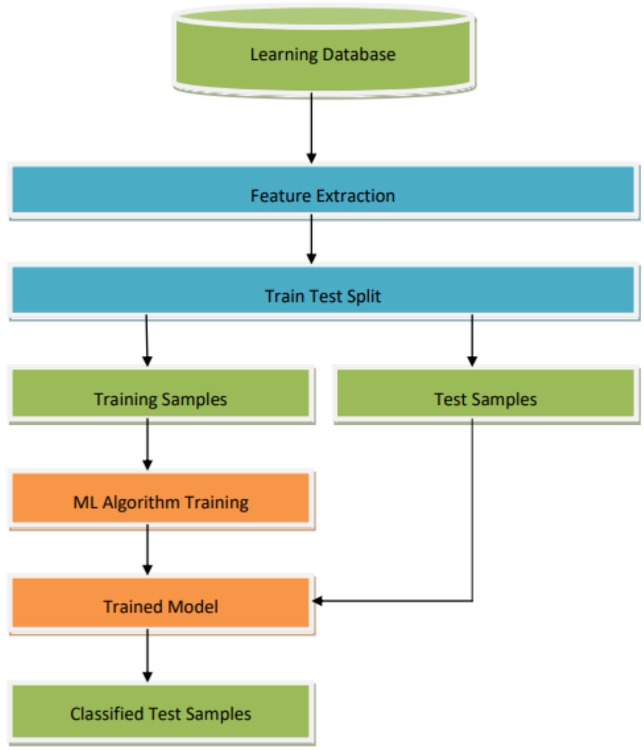


Figure 1 Plant disease detection model using plant leaf images

The model consists of the different components of the system. The model is designed to classify the crop leaf images according to the crop names, healthy, diseased and disease name. Therefore, we need to work with the multi-class classifier. The model accepts the input as images and produces the required outcomes.

Learning Database: that is an essential component of the proposed model. The learning data base includes the data and labels which work as the learning examples for machine. In this wok the plant village dataset has considered for experimentation. The dataset has a significant amount of data and it consist of 87K RGB color images of healthy and diseased crop leaves, which is distributed into 38 classes. The table 1 demonstrates the detail of available classes in dataset. Among the entire dataset we have used the three different crops namely corn, grapes and apple plant leafs. All these species has four sub classes thus we have 12 classes to be recognize. For constructing the final dataset we extract 40 images from each class and stored in subdirectories. Therefore the entire dataset contains a total of 480 sample images and 12 classes. All these images are taken in RGB format for utilizing the different image feature descriptor.

Feature selection: in this experiment, we consider three different kinds of image features, i.e. color, shape, and texture. The implemented methods for obtaining these features are discussed as:

1. Shape feature

In order to compute the shape feature of the images we have utilized two different techniques namely k-means and FCM for segmentation of the images:

The K-Means is a clustering algorithm which can be used for image segmentation. In this algorithm we select k random pixels of image as initial centroid, then measure the distance between each pixel and centroid. Based on the distance among centroid pixel and current pixel, the algorithm assigns a label to the pixel as nearest centroid pixel. In further steps the algorithm update the centroids by using average of all cluster pixels, and repeat this process until the stopping criteria [11]. Normally the following equation has used for identifying the stopping criteria:

$$\mathrm{E} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_{ij} - m_i \right\|^2$$

Where x_{ij} is pixel j of i-centroid, m_i is the centroid of i-class, n_i i is the number of pixels in i-class.

The K-means clustering follows the following steps:

Let the image I has the pixels $(x_j, x_z \ \ldots \ x_n)$ to create k number of clusters. And after processing the image pixels the algorithm returns k clusters and sum of error;

In first step we select k pixels of image as initial centroids say $(m_1, m_2, \ldots, m_k)$. Now we need to calculate the distance between each pixel x_i and centroid m_j, and then assign a label to each pixel based on distance:

$$\mathrm{d}(\mathrm{x_i}, \mathrm{m_j}) = \sqrt{\sum_{\mathrm{j}=1}^{\mathrm{d}} (\mathrm{x_i} - \mathrm{m_{j1}})^2}, \mathrm{i} = 1 \ldots \mathrm{N}, \mathrm{j} = 1 \ldots \mathrm{k}$$

Where, $d(x_i, m_i)$is the distance between pixel i and cluster pixel j.

Further we calculate the mean of pixel in each cluster to update the centroid pixel, using the blow equation:

Table 1 Description of Plant village dataset

S. No.	Plant	Disease	Images	Total image samples
1	Apple	Apple scab	2016	7771
		Black rot	1987	
		Cedar apple rust	1760	
		healthy	2008	
2	Blueberry	healthy	1816	1816
3	Cherry (including sour)	Healthy	1826	3509
		Powdery mildew	1683	
4	Corn (maize)	Cercospora leaf spot Gray leaf spot	1642	7316
		Common rust	1907	
		Healthy	1859	
		Northern Leaf Blight	1908	
5	Grape	Black rot	1888	7222
		Esca (Black Measles)	1920	
		healthy	1692	
		Leaf blight (Isariopsis Leaf Spot)	1722	
6	Orange	Haunglongbing (Citrus greening)	2010	2010
7	Peach	Bacterial spot	1838	3566
		Healthy	1728	
8	Pepper	Bacterial spot	1913	3901
		healthy	1988	
9	Potato	Early blight	1939	5702
		healthy	1824	
		Late blight	1939	
10	Raspberry	healthy	1781	1781
11	Soybean	healthy	2022	2022
12	Squash	Powdery mildew	1736	1736
13	Strawberry	healthy	1824	3598
		Leaf scorch	1774	
14	Tomato	Bacterial spot	1702	18345
		Early blight	1920	
		healthy	1926	
		Late blight	1851	
		Leaf Mold	1882	
		Septoria leaf spot	1745	
		Spider mites Two-spotted spider mite	1741	
		Target Spot	1827	
		Tomato mosaic virus	1790	
		Tomato Yellow Leaf Curl Virus	1961	

$$m_i = \frac{1}{N}\sum_{j-1}^{n_i} x_{ij} \;, i = 1,2,\ldots,K$$

Where N_i is the number of pixel in current cluster i; we repeat these steps until the objective function E results the desired values then algorithm produces final centroid pixels $(m_1, m_2, \ldots, m_k)$.

The second model which is being used here for image segmentation is FCM. The FCM (Fuzzy C-Means) is used to segment the image pixels into k groups. In FCM a membership function is used for creating the segments of images. The membership function is defined as:

$$M_{i,j} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_i\|}{\|x_i - c_k\|}\right)^{\frac{2}{n-1}}}$$

The membership function provides the bounding among image pixel and the centroid pixel. The membership values are used to decide a label to the pixel. Finally in this algorithm the stopping criteria are given using:

$$O_n = \sum_{i=1}^{N}\sum_{j=1}^{k} M_{i,j}^{j}\left\|x_i - c_j\right\|^2$$

Where, O_n is objective function, $M_{i,j}^{j}$ is membership function, $\left\|x_i - c_j\right\|^2$ is a distance function.

2. Color feature

The color feature shows better stability and insensitive to the rotation and zoom of image. Color information is used as a powerful tool in content-based image retrieval (CBIR). The goal is to retrieve all the images whose colors are similar to the query image. The feature we are going to use is called "Color Grid Moment". [13]

In order to compute the color features from image we convert the RGB format to the HSV. Then divide the image into 3x3 equal blocks. Using this we have nine blocks, then for all the nine blocks we compute its mean color using:

$$x' = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Where N is the number of pixels in each block, x_i is the pixel intensity of color channels. Then we compute variance for colors

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - x')^2$$

And skew-ness using:

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{N}(x_i - x')^3}{\left(\frac{1}{n}\sum_{i=1}^{N}(x_i - x')^2\right)^{3/2}}$$

In this way each block have 9 features, and entire image will have 81 features.

3. Texture feature

In order to measure the texture we used Local Binary Pattern (LBP). For a pixel in the image LBP is computed by comparing it with its neighbors:

$$LBP_{P,R} = \sum_{P=0}^{P=1} s\left(g_p - g_e\right)2^p$$

$$s(x) = \begin{cases} 0 & x \geq 0 \\ 1 & x < 0 \end{cases}$$

Where, g_e is the gray pixel value of the central pixel, g_p is the value of neighbors, P is the total number of neighbors, and R is the radius. In an image coordinate (0, 0), the g_p can be given as:

$$\left(Rcos\left(\frac{2\pi p}{P}\right), Psin\left(\frac{2\pi P}{P}\right)\right)$$

The neighbors that are not in the image is estimated using interpolation. Suppose the image is of size I*J after LBP each pixel is identified, a histogram to represent the texture:

$$H(k) = \sum_{i=1}^{I}\sum_{J=1}^{j} f\left(LBP_{p,r}(i,j), k\right), k \in [0,k]$$

$$f(x,y) = \begin{cases} 1 & x = y \\ 0 & otherwise \end{cases}$$

Where, K is the LBP pattern, U is value of LBP defined as the number of spatial transitions (bitwise 0/1 changes) in pattern

$$U(LBP_{P,R}) = \left|s(g_{p-1} - g_e) - s(g_0 - g_e)\right| + \sum_{p=1}^{p-1}\left|s(g_p - g_e) - s(g_{p-1} - g_e)\right|$$

The uniform LBP refer to the patterns which have limited transition ($U \leq 2$) in the circular binary presentation. The mapping from $LBP_{P,R}$ to $LBP_{P,R}^{u2}$ provides $P * (P-1) + 3$ distinct outputs, with a lookup table of 2^p elements, to achieve rotation invariance, which is defined as:

$$LBP_{P,r}^{riu2} = \begin{cases} \sum_{p=0}^{p-1} s(g_p - g_e)\ if U(LBP_{P,R}) \leq 2 \\ P+1\ otherwise \end{cases}$$

In this section we have described two shape features, one color and one texture feature measuring algorithms. thus we have four types of features to be used with the classification techniques. In addition, we prepared combination of these features in the following manner.

1. K-Means and LBP
2. K-means, LBP and CGM
3. FCM and LBP
4. FCM, LBP and CGM

Train and Test splitting: the image features and their combinations provide eight total different types of features for providing training to the ML algorithms. in order to create training and testing data we split the entire feature vectors into two parts training and testing. Therefore the 70% of features are randomly selected for the training of the ML model.

ML Model Training: in this phase the ML model has been employed to learn the feature patterns. Therefore we have implemented two classifiers namely SVM (support vector machine) and ANN (artificial neural network).

The SVM is considered as one of the most robust and accurate method. It requires small amount of examples for training, and is insensitive to the number of dimensions [8]. The SVM is trying to locate the best arrangement to distinguish between two classes. The idea of the "best" characterization has been acknowledged geometrically. For classifying a dataset, a simple technique is to find an isolating hyperplane $f(x)$ that goes between the two classes. When two examples are resolved, then new example x_n can be arranged by the function $f(x_n)$; x_n has a place with the positive class if $f(x_n) > 0$. However, there are a number of hyperplanes are possible but the SVM ensures the most effective margin between two samples by expanding the margin between the two classes. Geometrically, the hyper plane demonstrates a brief separation between two samples [9].

The second classifier is also a popular ML technique which is known as Artificial neural network (ANN). The ANN is a prominent technique for complex non-linear and multivariable data analysis. ANN is a data handling disciplinary subject, its foundations in the working of the human neural system. It is like our nervous system, which gets the data, translates and gives the output. ANN is made with a substantial number of small components called neurons, which are interconnected with each other. The ANN incorporates its capacity to decipher exact information and perceive the examples. Additionally, the speed and precision are higher. The element of ANN is no past information behind the procedure and thus that are recognized as discovery framework. Also, ANN can go up against every semantic variable that can't be estimated by ordinary techniques [6].

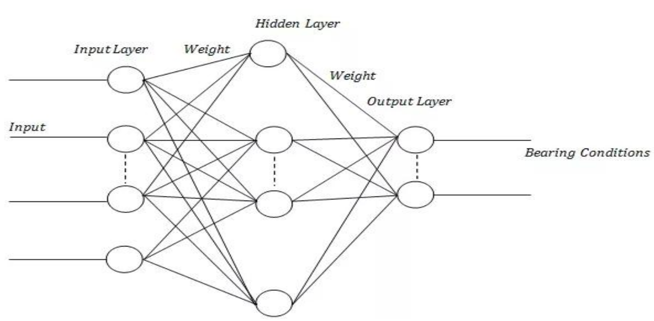


Figure 2 shows the basic Neural Network model

The introductory model of the ANN is demonstrated in figure 2. The main advantages of ANN include their high tolerance to noisy, and ability to classify patterns on which it is not trained. The concern is in the training phase which is focused on the weights of the ANN, which is used according to the transactions used. For each training input, the ANN receives, input and the expected output [7].

However there are two classifiers have implemented the SVM and ANN. The SVM is a binary classifier therefore; we need to extend the classifier for the multiclass classification. In this context, the One-Vs-All (OVA) strategy is used. In addition to that to implement the ANN technique the BPN (backpropagation neural network) is implemented.

Trained model: the ML algorithms are consuming the training sample provided as input and develop the data models which will be able to recognize the similar objects and patterns.

Test Samples: the validation or testing of the proposed plant disease detection model has carried out with the randomly selecting 30% of samples from the entire extracted features. The model evaluates each sample image in the test set and produces the class labels for each of the sample.

Classified samples: the test datasets are classified and their class names are predicted. Based on prediction the performance of the system is measured and reported in the next section.

III. RESULTS ANALYSIS

The aim of this study is to conduct a comparative performance study among the different image feature selection techniques and two well-known classifiers. The classifiers are SVM (support vector machine) and ANN (artificial neural network). Additionally, we have prepared a set of different feature's combinations for extraction of features from image which is also involved with these classifiers.

a. Performance Parameters

In order to conduct the performance analysis we consider the following parameters:

A. **Accuracy:** The accuracy of the ML technique describes the correctness of the learning model. The accuracy is a ratio of correctly classified and total samples. The following formula can be used for measuring the accuracy.

$$Accuracy = \frac{Total\ Correctly\ classified}{Total\ Samples} X100$$

B. **Error Rate:** The error rate of a ML model demonstrates the misclassified samples. It is the measurement of error of the ML model. That can be calculated using:

$$Error\ Rate = 100 - Accuarcy$$

Or

$$Error\ Rate = \frac{Total\ incorrect\ classified}{Total\ Samples} X100$$

C. **Time Consumptions:** The time consumption is also known as the time complexity. The amount of time utilized during the calculation of features and the learning of ML algorithm is measured hear as time consumption. The following equation can be used:

$$Time\ consumpion = End\ time - Start\ Time$$

D. **Memory Usages:** The memory usage is also known as the space complexity. That can be measured using the factors as:

$$memory\ usage = total\ space - free\ space$$

This section describes the used parameters, the next section describes the experimental analysis of the conducted experiments.

Table 2 Performance of the model

S. No.	Features	Accuracy		Error Rate		Training Time		Memory Used	
		SVM	ANN	SVM	ANN	SVM	ANN	SVM	ANN
1	K-means	87.7	89.2	12.3	10.8	276	288	13393	14663
2	FCM	92.5	92.8	7.5	7.2	224	239	13454	14544
3	LBP	85.3	87.7	14.7	12.3	123	143	13145	14338
4	CGM	83.8	84.1	16.2	15.9	124	149	13433	14176
5	K-Means + LBP	90.8	94.3	9.2	5.7	435	498	13584	14566
6	K-Means + LBP + CGM	93.2	98.5	6.8	1.5	553	587	13873	14354
7	FCM + LBP	92.3	95.8	7.7	4.2	367	408	13343	14221
8	FCM + LBP + CGM	96.1	99.1	3.9	0.9	487	510	13672	14426

b. Experimental Results

There are four performance indicators has reported in this work namely accuracy, error rate, time and memory usage. The accuracy of the implemented test scenarios using two ML algorithms are discussed first. In this experiment we use the 8 combinations of feature selection techniques from images, and used with the two classifiers SVM and ANN. Based on the conducted experiments the classification accuracy has reported in figure 2(A) and table 2. The X-axis of the diagram shows the different combinations of image features. Additionally, Y-axis contains the accuracy of the ML algorithm in terms of percentage (%). The results demonstrate the increasing combinations of features have improving the classification accuracy of both the ML algorithms. According to the findings, the performance of FCM and ANN is much effective as compared to the K-means and SVM based approach. Similarly the error rate of the classifiers with the combinations of different features is reported in figure 3(B) and table 2. The Y-axis shows the error rate of ML algorithms in percentage.

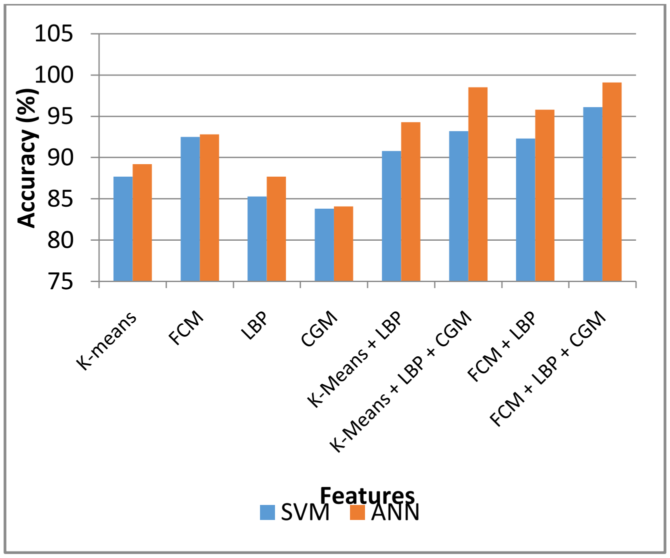
According to the observations, the FCM and ANN-based combinations provide less inaccurate results as compared to the SVM and k-means based algorithm. The next parameter is time consumption, which is demonstrated in figure 3(C) and table 2. In the bar graph Y-axis shows the consumed training time. According to the findings, the ANN consumes less amount of time for feature extraction and learning as compared to the SVM algorithm. In addition, we also find that K-means based segmentation approach consumes a higher amount of time as compare to FCM based approach. Finally the memory usage of the implemented techniques is reported in figure 3(D) and table 2. The mean values of different experiments are used as results. The figure shows the different utilized features in X-axis and the Y-axis shows the memory used. The memory usage is measured in terms of KB (kilobytes). According to the results the ANN consumes higher memory as compared to SVM in all the scenarios.

IV. CONCLUSION

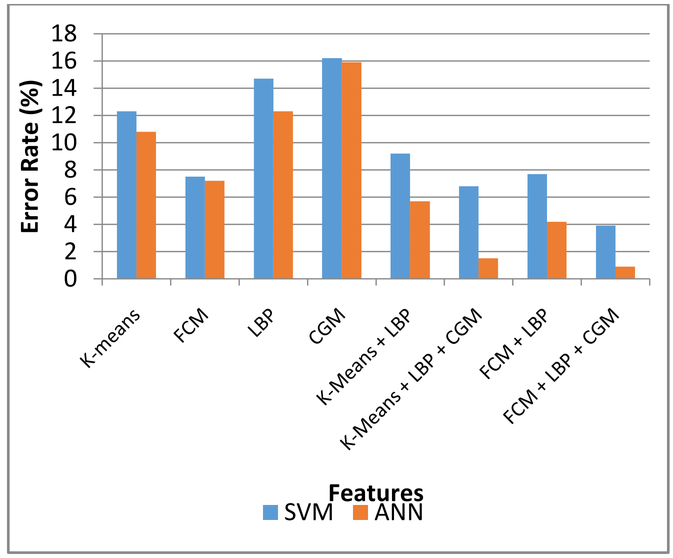
The aim of this paper is to design a machine learning model for farmer’s welfare. This model will be used in a smart farming application as a component. This component contributes as a tool to identify the disease in various different kinds of plants. The proposed technique utilizes the feature selection, training and testing modules. The images are processed through three types of features i.e. shape, color, and texture. In order to calculate shape features from the images, the K-means and FCM are used, for color features CGM is used, and for texture analysis, the LBP is used.

Additionally, the different combinations of features have also prepared, for more effective feature representation. Further to classify the features and their combinations we have implemented two classifiers SVM and BPN. By using a piece of a large open source plant village dataset we demonstrate the effectiveness of the proposed model. The given prototype has successfully works on these images to recognize the disease in three different plants namely Corn, Apple and grapes. According to experimental results we conclude the following facts which will help to design more effective and accurate disease detection modeling.

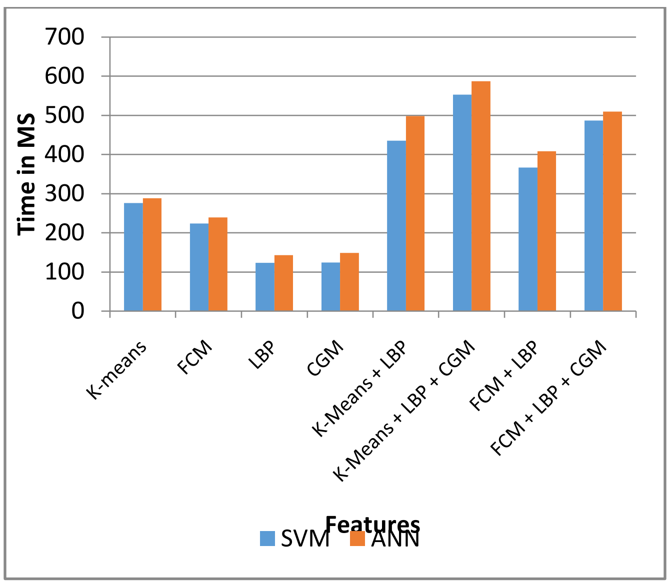
1. The combination of shape, color and texture feature is more effective than the individual feature-based classification
2. The combination FCM, LBP and CGM is more effective then K-means, LBP and CGM
3. The classification accuracy of the proposed model based on BPN is higher than the SVM.



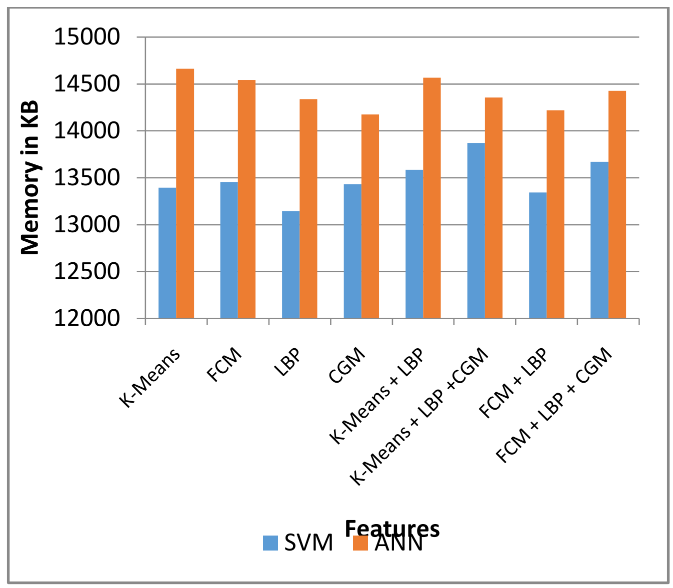
(A)



(B)



(C)



(D)

Figure 3 shows the performance of the implemented disease detection model for demonstrating the comparison among different image feature descriptor combinations and the classifiers

However the proposed model has working fine based on a very small amount of samples and with the limited number of crops. But the actual problem has a number of variations therefore the proposed work is motivated us to improve the currently contributed model based on the following directions:

1. The size of data in plant village data is very huge which can not be handled using the simple methods therefore we are proposing to work for the large size of dataset using the cloud such as google colab as well as for ease in processing the large data we are proposing to utilize the deep learning models
2. The current model provide the classification for only 12 classes of images we have tried to extend the evaluation for all the 38 classes of diseases and healthy plant leafs.

REFERENCES

[1] *"Agricultural Education"*, http://www.hillagric.ac.in/aboutus /vc/vc_addresses/pdf/2017/08.12.2017-Agri.Edu.Day-03.12. 2017.pdf

[2] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, D. Bochtis, *"Machine Learning in Agriculture: A Review"*, Sensors 2018, 18, 2674; doi:10.3390/s18082674
[3] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, W. Ouyang, *"T-CNN: Tubelets with Convolutional Neural Networks for Object Detection from Videos"*, Copyright 2017 IEEE
[4] M. Wang, Y. Wan, Z. Ye, X. Lai, *"Remote sensing image classification based on the optimal support vector machine and modified binary coded ant colony optimization algorithm"*, Information Sciences, 402, 2017, 50-68
[5] G. Wang, W. Li , M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, T. Vercauteren, *"Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning"*, IEEE Transactions on Medical Imaging, Vol. 37, No. 7, July 2018
[6] S. Yi, A. Andrzejak, D. Kondo, *"Monetary cost-aware checkpointing and migration on amazon cloud spot instances"*, IEEE Transactions on Services Computing 5, no. 4 (2012): 512-524.
[7] S. O. Danso, *"An Exploration of Classification Prediction Techniques in Data Mining: The insurance domain"*, Master's Degree in Advanced Software Engineering, School of Design, Engineering, and Computing. Bournemouth University, September, 2006.
[8] V. Kumar, J. Ghosh, D. J. Hand, *"Top 10 algorithms in data mining"*, Knowledge and Information System, PP. 1–37, (2008).
[9] V Vapnik, *"the nature of statistical learning theory"*, Springer, (1995) New York.
[10] H. Yang, L. Chan, I. King, *"Support Vector Machine Regression for Volatile Stock Market Prediction",* Intelligent Data Engineering and Automated Learning IDEAL, PP. 391- 396, Springer-Verlag Berlin Heidelberg, 2002
[11] N. Chaturvedi, A. Rajavat. *"Novel Approach of K-Mean Algorithm"*, The International Journal Of Engineering And Science (IJES) Volume 2 Issue 10, Pages 20-23, 2013
[12] Y. K. Dubey, M. M. Mushrif, *"FCM Clustering Algorithms for Segmentation of Brain MR Images"*, Hindawi Publishing Corporation Advances in Fuzzy Systems Volume 2016, Article ID 3406406, 14 pages
[13] Y. G. Jiang, J. Yang, C. W. Ngo, A. G. Hauptmann, *"Representations of Key point-Based Semantic Concept Detection: A Comprehensive Study"*, IEEE, 2008
[14] Z. Guo, L. Zhang, D. Zhang, *"A Completed Modeling of Local Binary Pattern Operator for Texture Classification",* IEEE transaction on image processing, 2010