International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

AI-BASED CROP IDENTIFICATION, DISEASE RECOGNITION AND PESTICIDE SUGGESTION

Prasad Salwe[1], Avanti Sakharkar[2], Rudrani Shete[3], Manisha Kowdiki[4]

[1]*B.Tech student, Electronics and Communication, MIT World Peace University, Maharashtra, India*
[2]*B.Tech student, Electronics and Communication, MIT World Peace University, Maharashtra, India*
[3]*B.Tech student, Electronics and Communication, MIT World Peace University, Maharashtra, India*
[4]*Professor(Pursuing Ph.D.), Electronics and Communication, MIT World Peace University, Maharashtra, India*

---***---

Abstract- Identification of the crop is an important aspect for preventing the losses of the agricultural product. The study of crop diseases means studying the visually observable patterns observed on the crop. Disease detection on crops is very unfavourable for sustainable agriculture. It is very difficult to monitor the crop diseases manually. It requires an enormous amount of work, professionals in the crop diseases, and requires extreme processing time. Hence, image processing is used for the detection of crop diseases. The proposed system involves five important steps: a) image capturing, b) image pre-processing, c) image segmentation, d) feature extraction and e) classification. Finally, features are trained with the machine learning algorithm and tested on unknown images. Results are being calculated based on qualitative and quantitative analysis. In qualitative analysis, pre-processing and segmentation methods are operated on the whole data set. In quantitative analysis, the accuracy of all machine learning algorithms is computed. This paper discussed the methods used for the detection of crop diseases using their leaf image. This paper also discussed some segmentation and feature extraction algorithms used in crop disease detection. The classification is done using Gradient Boosting Classifiers with the highest accuracy.

Key words: **Artificial Intelligence, Machine learning, Gradient boosting, Gray level co-occurrence matrix, true positive, true negative, false positive, false negative.**

NOMENCLATURE

FAO	Food and Agricultural Organization
GLCM	Gray-Level Co- Occurrence Matrix
SVM	Support Vector Ma chine
MDC	Minimum Distance Classifier
GBDT	Gradient Boosting Decision tree
KNN	K-Nearest Neighbor
CNN	convolutional neural network
SMOTE	Synthetic Minority Over-Sampling Technique
RGB	Red, Green , Blue
HSV	Hue, Saturation,Value
OPP	Opponent Color Space
RBF	Radial Basis Function
SGD	Stochastic Gradient Descent
GB	Gradient Boosting
ML	Machine Learning
ANN	Artificial Neural Network
XOR	Exclusive OR
MLP	Multilevel Perceptron
TRP	True positive
TRN	True Negative
FAP	False Positive
FAN	False Negative

1. INTRODUCTION

India is a cultivated country and about 70% of the population is involved in agriculture. Farmers have a large range of diversity for selecting various suitable crops and finding suitable pesticides for the plant. The timely diagnosis of plant diseases is important as it is challenging. In the early days, the monitoring and analysis of plant diseases were done manually by an expert in that field. This requires a tremendous amount of work and requires excessive processing time. To overcome this, Digital Processing Techniques can be used to detect the diseases of the plants. It is estimated that 30 to 40% of harvests are lost each year throughout the production chain due to diseased crops, representing a threat. Disease development in plants continues to have a great impact on society. At present, around 10-16% of global crop production Scheinis lost to pests—including insects, fungi, and bacteria. Epidemics may lead to disease injuries, which may lead to crop loss (damage) which, in turn, may lead to economic loss Zadoks and 1979; Zadoks 1985)[11]. As per the

2014 FAO world agriculture statistics, India is the world's largest producer of many fresh fruits, vegetables, major spices, fibrous crops, and staples. At a global scale, pathogens and pests are causing wheat losses of 10 to 28%, rice losses of 25 to 41%, maize losses of 20 to 40%, potato losses of 8 to 21%, and soybean losses of 11 to 32%, according to the study, published in the journal *Nature, Ecology & Evolution*[13].

Automatic detection of plant diseases is an essential research topic as it may prove benefits in monitoring large fields of crops and thus automatically detect the symptoms of diseases as soon as they appear on plant leaves. Therefore, looking for fast, automatic, less expensive, and accurate methods to detect disease by calculating leaf area through pixel number statistics. Most of the image processing algorithms are based on intensities. In an outdoor environment, the intensities of the images change. The machine learning approaches can be used to classify the plants in the healthy and diseased leaf. Disease detection in every aspect is important in today's world. This system helps the farmer to detect the disease on plants and helps to find the right solution for the disease. This motivates the development of the proposed system.

2. LITERATURE REVIEW

2.1 Related works

Different techniques had been conferred by various researchers in detecting and recognizing plant diseases. Shrutika Ingale et al. in their proposed system "Plant Leaf Disease Detection Recognition using Machine Learning'' [1] implemented image processing for automatic disease detection. They carried out a segmentation process using boundary and spot detection, masking of green pixels, and thresholding. Feature extraction was performed using Grey level co-occurrence matrix (GLCM) for texture features, shape features are calculated using aspect ratio, and also used venation for getting vein patterns. For classification, Support Vector Machine-SVM, Otsu threshold algorithm, and Back Propagation methods were used. SVM gave a high prediction accuracy. Back Propagation was easy to program and could adopt complex functions. K Means Clustering algorithm was able to give quick results.

R.Meena Prakash et al. in "Detection of Leaf Diseases and Classification using Digital Image Processing"[2] used k-means clustering for segmentation of the leaf and used statistical Grey level co-occurrence matrix(GLCM) for texture feature extraction for computing the spatial relationship between the pixels in the image which formed a matrix and gave the statistical measures. Support Vector Machine was used as a classification tool, by which the training vectors gained the classification accuracy of 0.9 to 1.0.

Anand. H. Kulkarni et al. proposed "Applying image processing technique to detect plant diseases"[3] wherein Gabor filter was used for the feature extraction. Image classification was done using an ANN classifier which used the union of color and texture features. It gained a recognition rate of up to 91%. Their analysis showed that the neural network's performance also depends on the quality of the sample image along with the number of features and hidden neurons in different input conditions so that the samples will be classified correctly corresponding to their classes.

S. Arivazhagan, et.al in "Detection of the unhealthy region of plant leaves and classification of plant leaf diseases using texture features"[4] used two methods for classification- Minimum Distance Classifier and SVM. In MDC, classification gain is calculated using the no. of images correctly classified and the total number of images that belong to a particular texture group. In SVM, a list of p numbers was viewed and a (p-1) hyperplane was considered, called a linear classifier. One hyperplane was selected, representing the largest distance between two classes so that the margin from it to the nearest data point on each side could be maximized. The overall accuracy gained by MDC was 86.77 and that of SVM was 94.74. Thus the accuracy results were improved by the SVM classifier.

K.R.N.V.V.D.Aravind et.al in the system "Classification of Healthy and Rot Leaves of Apple Using Gradient Boosting and Support Vector Classifier"[5] Gray Level Co-occurrence Matrix (GLCM) was adopted for texture feature extraction as its recognition rate was very fast. Global features color and shape are taken for classification, which was done using Gradient Boosting Algorithm and SVM. Accuracy scores for both the algorithms were calculated and compared and found out the confusion matrix. Gradient Boosting algorithm imparted less accuracy of 87% as compared to SVM with an accuracy score of 91%.

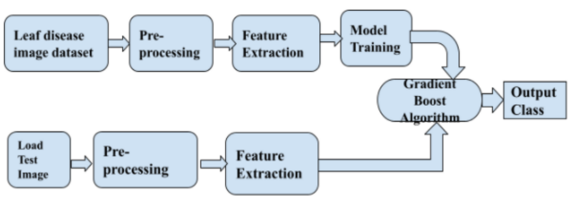
A paper "Classification of Plant leaf diseases using Machine Learning and image preprocessing techniques" by Pushkar Sharma et. al[6] gave the results for plant disease by testing four different algorithms. They tested algorithms namely logistic regression, KNN, SVM, and CNN. After analyzing the logistic regression model, an accuracy of 66.4% is obtained. For the KNN classifier, the accuracy found was 54.5%. The SVM classifier gave an accuracy of 53.5% and the deep learning model CNN was able to give a maximum accuracy of 98%. The results proved that CNN outperformed all other classifiers.

A research paper named "Corn Disease Identification Based on Improved GBDT Method'' by Tong Xiao et.al [7] combined the GBDT(Gradient Boosting Decision tree) algorithm and data balance algorithm to achieve the identification of corn disease. They compared the results with traditional classification models such as logistic regression, linear/RBF SVM, decision trees, random forest, BP propagation, and Naive Bayes. The GBDT algorithm gained a recognition rate of 92%. Thus their result analysis proved that the GBDT algorithm has the highest precision in the recognition of crop diseases.

2.2 Review

Since there are a lot of ways of crop disease recognition methodologies, more of the learning approach is needed to achieve the utmost precision in disease recognition. K-means clustering[2] is a slow learner. It does a good job only when the cluster has a kind of spherical shape and the numbers of clusters are to be predefined. Otsu threshold Algorithm[1] works well regardless of uniformity and shape measures but takes more time for processing. The GLCM method[5] can be extremely sensitive to changes such as scale, rotation of any image. The recognition rate of GLCM is very fast compared to its computation time. As the size of the dataset was increased, the performance of gradient boosting also rose. In ANN[3], the network can be reduced to a certain value of the error on the sample which indicates that the training has been finished and this value will not give any prime results. The efficiency of the network increases if there is the least number of hidden neurons. Logistic Regression[6] is the simplest classification algorithm function. CNN(Convolutional Neural network) is a very complex model and it requires better computational power. So far it was able to provide the highest accuracy. Synthetic Minority Over-Sampling Technique(SMOTE)[7] solves the problem which occurs due to data imbalance while undergoing image pre-processing. Bayesian classification algorithm[7] has a disadvantage in that sample attributes need to be balanced.SVM[4] works well with semistructured data such as images.If an appropriate kernel function is selected then it can even solve the complex problem. The training time required is huge for the larger dataset. It will not perform well if the number of features corresponding to data points surpasses training data sample numbers.

3. PROPOSED METHODOLOGY



3.1 Block Diagram

Fig-1: Workflow of the proposed system

Fig.1 represents the workflow of the entire system. It shows the leaf disease image dataset of 12862 images, pre-processing, feature extraction, model training, gradient boosting classifier, and final output result of classification. The Block Diagram is divided into two parts, the training-testing phase for generating machine learning models and the actual system for classifying images.

3.2 Dataset

Fig-2: Apple healthy

Fig-3: Apple Blackrot

3.2.1dataset example:

Fig-2. & Fig-3. shows the example of healthy and diseased leaves of our dataset.

3.2.2 Splitting dataset

The whole dataset consists of 12862 images of leaves of 4 different plant species. Our dataset is segregated into 12 different categories of which some are diseased images, and some are healthy images. We have taken apple, corn, grape, and tomato in our dataset to train the machine learning model. The dataset was then split into two parts: the training set, which comprises 75% data, and the test set, which comprises the remaining 25%. Table-1. shows the segregation of the dataset in 12 different classes.

Table-1: Segregation of the dataset in 12 different classes.

Database	Total images	Training	Testing
Apple Blackrot	621	497	124
AppleCedar applerust	275	220	55
Apple Healthy	1628	1299	329
Corn(maize) Gray leafspot	513	411	102
Corn(maize) Common rust	1192	954	238
Corn(maize)healthy	1157	925	232
Grape Blackrot	1180	944	236
Grape Esca (Black Measles)	1383	1107	276
Grape Healthy	423	339	84
Tomato Earlyblight	1000	800	200
Tomato Healthy	1588	1270	318
Tomato lateblight	1902	1521	381

3.3 METHODOLOGY

3.3.1IMAGE PROCESSING

Image pre-processing is defined as methods performed on raw images to remove unwanted distortions and to enhance image features important for further processing. In the developed model, image pre-processing and analysis. In our research, resizing images and converting from RGB to other formats is done. In this phase, RGB to gray conversion is carried out to obtain a 2D image. Pre-processed images are reduced image size and image crop to a given input. It processes and enhances the image to its needed color scale. The study uses gray-scale and HSV and resized images to 256x256 resolution for processing.

Step 1. Resize the 2D image to 256 × 256 sized images.

Step 2. Convert input image to 2D image by converting RGB image to Gray-scale image.

Step 3. Convert input image to 2D image by converting RGB image to HSV image.

3.3.2RGB TO HSV CONVERSION:

The R, G, B values are divided by 255 to change the range from 0-255 to 0-1:'

$$R' = R/255$$

$$G' = G/255$$

$$B = B/255$$

$$Cmax = \max(R', G', B')$$

$$Cmin = \min(R', G', B')$$

$$\Delta = Cmax - Cmin$$

Hue Calculation

$$H = \begin{Bmatrix} 0° & \Delta = 0 \\ 60° \times \left(\frac{G' - B'}{\Delta} mod6\right) & C_{max} = R' \\ 60° \times \left(\frac{B' - R'}{\Delta} + 2\right) & C_{max} = G' \\ 60° \times \left(\frac{R' - G'}{\Delta} + 4\right) & C_{max} = B' \end{Bmatrix}$$

Saturation Calculation:

$$S = \begin{Bmatrix} 0 & C_{max} = 0 \\ \frac{\Delta}{C_{max}} & C_{max} \neq 0 \end{Bmatrix}$$

Value calculation:

$$V = Cmax$$

3.3.3 IMAGE SEGMENTATION

Image segmentation refers to the conversion of an image into a collection of regions of pixels that are represented by a mask or a labeled image. By dividing an image into segments, we can process the important segments of the image instead of processing the entire image. Image segmentation techniques involve thresholding, clustering, graph-based segmentation, region growing, etc. It is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics. In leaf image segmentation, we have researched and used thresholding-based segmentation and masking technique in the model.

a)Masking: Masking of green pixels is carried out in this operation. Masking is the process that is underneath many types of image processing, including edge detection, motion detection, and noise reduction. Masking is an image processing method in which we set a particular value of the pixel of the green portion and if the pixel value exceeds the particular set value, it is discarded. The remaining portion is identified as a diseased region of the image. In this way, we will get the diseased part of the leaf as an extracted portion.

b) Threshold based Segmentation: The process of thresholding involves, comparing each pixel value of the image (pixel intensity) to a specified threshold. This divides all the pixels of the input image into 2 groups:

1. Pixels having intensity values lower than the threshold.
2. Pixels having an intensity value greater than the threshold.

The group having members with pixel intensity, greater than the set threshold, are assigned "Max Value", or in the case of a grayscale, a value of 255 (white). The members of the remaining group have their pixel intensities set to 0 (black). Binary image contains only values of ones and zeros. Then this binary image is multiplied with the original RGB color image. In this way, the infected portion is extracted.

$$dst(x,y) = \begin{matrix} maxVal & if\ src(x,y) > src \\ 0 & otherwise \end{matrix}$$

3.3.4 MORPHOLOGICAL OPERATIONS

After thresholding, morphological operations are performed on the leaf image. Morphological techniques probe an image with a small shape or template called a structuring element. The structuring element is positioned at all possible locations in the image and it is compared with the corresponding neighbourhood of pixels. Dilation and Erosion operations are used in our model to get the expected image of the diseased leaf. The erosion of a binary image *f* by a structuring element *s* (denoted $f \ominus s$) produces a new binary image $g = f \ominus s$ with ones in all locations (x,y) of a structuring element's origin at which that structuring element *s* fits the input image *f*, i.e. $g(x,y) = 1$ is *s* fits *f* and 0 otherwise, repeating for all pixel coordinates (x,y). The dilation of an image *f* by structuring element *s* (denoted $f \oplus s$) produces a new binary image $g = f \oplus s$ with ones in all locations (x, y) of a structuring element's origin at which that structuring element *s* hits the input image *f*, i.e. $g(x, y) = 1$ if *s* hits *f* and 0 otherwise, repeating for all pixel coordinates (x, y). Dilation has the opposite effect to erosion -- it adds a layer of pixels to both the inner and outer boundaries of regions.

- Flat dilation operator: local maximum over window W
  g⌈x, y⌊ ⌉⌋ = max W f⌈x, y⌊ ⌉{ { ⌋}} := dilate(f ,W)

- Flat erosion operator: local minimum over window W
  g⌈x, y⌊ ⌉⌋ = min W f⌈x, y⌊ ⌉{ { ⌋}} := erode(f ,W)

After thresholding, the leaf image is dilated and eroded three times with iterations of 3, 5, and 6 in the respective order. The output result will be the image with a diseased portion in a clear format concerning operations done before dilation and erosion. We must keep repeating the morphological operations of the image till we get the desired result of the image.

3.3.5 FEATURE EXTRACTION

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is many variables that require a lot of computing resources to process. The process of feature extraction is useful when we need to reduce the number of resources needed for processing without losing important or relevant information. Feature extraction can also reduce the amount of redundant data for a given analysis. Also, the reduction of the data and the machine's efforts in building variable combinations (features) facilitate the speed of learning and generalization steps in the machine learning process.). These new reduced sets of features should then be able to summarize most of the information contained in the original set of features. Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features.

In our research and project model, we have researched various features which are categorized into two types that are low-level features and high-level features. Low-level features are extracted directly from the original images, whereas high-level feature extraction is based on low-level features. The texture is a surface property. The image texture depends on the scale or resolution at which it is displayed. A texture with specific characteristics on a sufficiently small scale could become a uniform texture if it is displayed at a larger scale. So we have used texture also in the features used in our system.

Following are the features used in our model:

a) Aspect Ratio (shape feature); The length of the leaf is found by taking the Euclidean distance between the two points on either side of the long axis whereas breadth corresponds to the length of the minor axis.

$$Aspect\ ratio = \frac{Length\ of\ the\ leaf}{breadth\ of\ the\ leaf}$$

By using length and breadth of leaf aspect ratio is found.

b) Area (shape feature): The area is calculated by initially finding the area of one pixel.

$$Area = area\ of\ one\ pixel * total\ no.of\ pixels$$

c) Perimeter (shape feature): The perimeter of the leaf is given by the count of pixels having the leaf margin.

d) Rectangularity (shape feature):Rectangularity shows the similarity between a leaf and a rectangle.

$$Rectangularity = \frac{L * W}{A}$$

e)Color feature: Color space gives the color in the form of intensity value. We specify, visualize and create the color by using the color space method. There are different color feature extraction methods. We obtain the color feature by using the histogram color method. The color histogram represents the image from a different perspective. The image in which color bins of the frequency distribution are represented by color histogram and it counts the pixels which are similar and store it. Color histogram analyzes every statistical color frequency in an image. If the particular region is affected the skin lesions region changes the color effectively. Relative color histograms in different color spaces are constructed to identify the melanoma. The 3-D histogram is constructed for the color spaces such as RGB, LAB, HSV, HUE, and OPP [12]. RGB color space represents a mixture of Red, Green, and Blue. The color component is represented by the mixture coefficients of these three colors. The drawbacks of the color spaces are not perceptually uniform, and it depends on the acquisition setup. So, three colors are used which are RED, BLUE, and GREEN. The histogram is operated with all pixel values of these three colors of an image. At first, RGB is set to its values of the red-green and blue channels. Then the mean of all three-pixel values of three channels is calculated. Then standard deviations of all three color channels are calculated. In this way, color features are calculated.

f) Texture feature: The texture is a surface property. It is characterized by the spatial distribution of gray levels in a neighborhood. Since texture shows its characteristics both by pixel coordinates and pixel values, there are many approaches used for texture classification. The image texture depends on the scale or resolution at which it is displayed. A texture with specific characteristics on a sufficiently small scale could become a uniform texture if it is displayed at a larger scale. In statistical texture analysis, from the distribution of intensities, the texture features are obtained at a specified position relative to one another in an image. The statistics of texture are classified into first order, second order, and higher-order statistics.

- TWO TEXTURED FEATURES ARE USED:

A)GLCM FEATURE [1] - GLCM (also called Gray tone spatial dependency matrix) is a tabulation of the frequencies or how often a combination of pixel brightness values in an image occurs. The figure below represents the formation of the GLCM of the grey-level (4 levels) image at the distance d= 1 and the direction of 0°. Features are the statistical data of the Image. GLCM can also be formed for the direction of 45°, 90°, and 135°as shown in Fig-4 Second order below. From the centre (0,0) to the pixel 1 representing direction = 0 with distance d =1, to the pixel 2 direction = 45°with distance d = 1, to the pixel 3 direction = 90° with distance d = 1, and to the pixel 4 direction = 135°with distance d = 1. GLCM is the method that is used to extract different features from Gray and binary images. In the proposed approach following GLCM features are extracted:

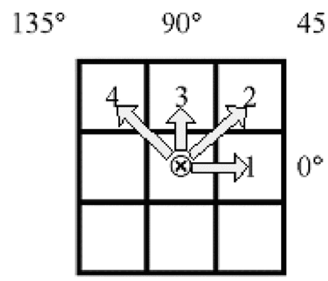


Fig-4: The direction of GLCM generation

GLCM is the method that is used to extract different features from Gray and binary images. In the proposed approach following GLCM features are extracted:

Table-2:-Features

Sr. No	Features	Description
1	Contrast = $\sum_{i,j} \lvert i-j \rvert^2 p(i,j)$	Measures the local variations in the gray-level co-occurrence matrix
2	Homogeneity = $\sum_{i,j} \frac{1}{1+(i,j)^2} p(i,j)$	The homogeneity measures of the closeness of the element distribution in GLCM-to-GLCM diagonals
3	Energy = $\sum_{i,j} p(i,j)^2$	The energy is the measure of uniformity between the pixels.
4	Entropy = $-\sum_{i,j} p(i,j)\log(p(i,j))$	The entropy of an image is a statistical measurement of the randomness of the pixel element.

5	Dissimilarity = $\sum_{i,j} \lvert i-j \rvert p(i,j)$	Dissimilarity is a measure that defines the variation of gray level pairs in an image

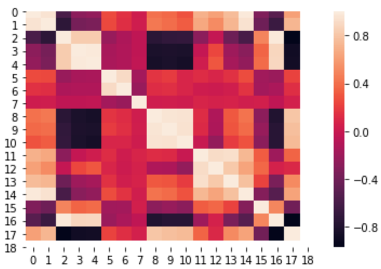
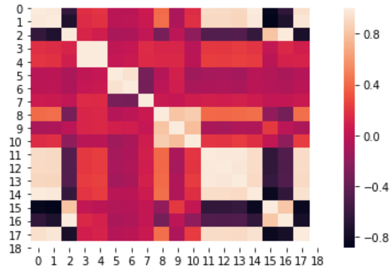
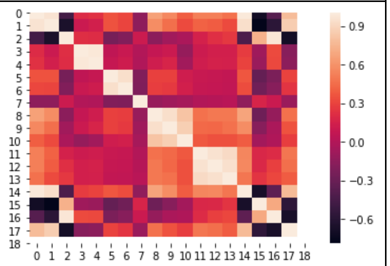
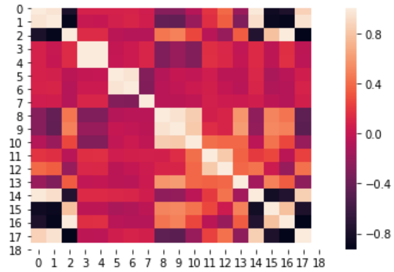
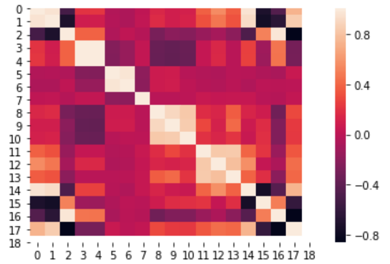
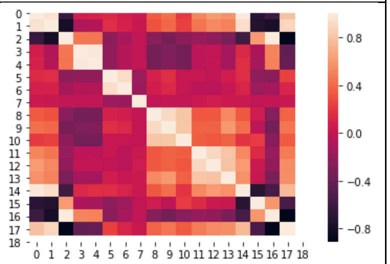
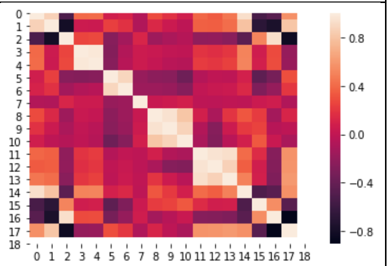
B) HARLICK FEATURE - GLCM is calculated for every image of ROI with orientations of different angles and considering eight and sixteen quantization levels of gray values and distance between two neighboring pixel one. Then from these GLCM's, Harlick features are calculated for every image. After averaging these Harlick features over all four orientation angles we get an averaged value of all five Harlick features for every image. Then by combining these values Harlick features a single feature vector that is formed for every print. Then we normalize this feature vector. In our model, we have used harlick features to extract values from color features and split the harlick feature into 4 features which are ht_mean_[1], ht_mean_[2], ht_mean_[4], ht_mean_[8].

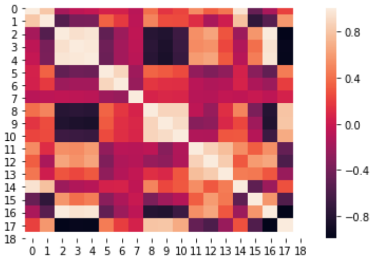
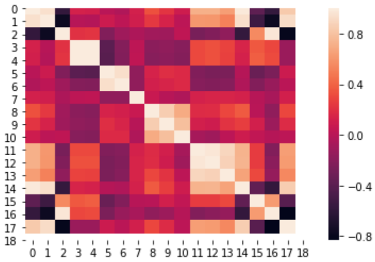
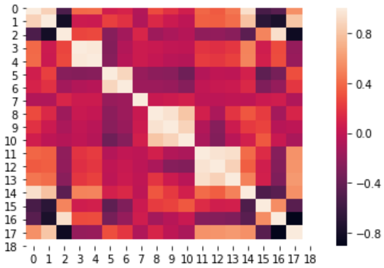
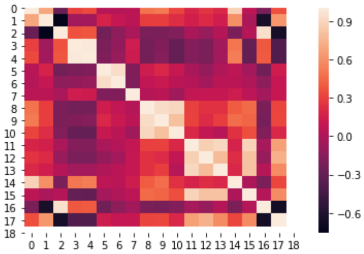
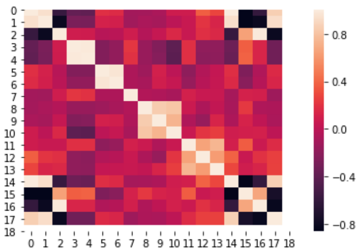
Concluding the feature extraction process, a total of 18 features are used in the model for the classification of diseased images.

3.3.6 FEATURE EXTRACTION

The selection of features is the method of selecting accurate features of an image to get the highest efficiency. In our research, we have experimented with 18 features in total which were found to be useful. So all 18 features are used for the model. Below **Table.3** is Heat-Map of all 12 classes of plant species and its 18 features and last column as a label of an image used.

Table-3: Heat-Map analysis of all the 18 features extracted from training set leaf images.

Sr. No	Classes	Heat map
1	Apple-Black rot	
2	Apple-Cedar appleru st	
3	Apple healthy	
4	Corn(m aize) Cercosp ora leafspot Graylea fspot	
5	Corn (maize) Commo nrust	
6	Corn (maize) healthy	
7	Grape Black rot	

8	Grape Esca (Black Measles)	
9	Grape-healthy	
10	Tomato Early blight	
11	Tomato healthy	
12	Tomato Late blight	

Following are the labels for 18 features:

1.Contrast
2. Dissimilarity
3. Homogeneity
4. Entropy
5. Energy
6. Area
7. Perimeter
8. Aspect Ratio
9. Red Color Intensity
10. Blue Color Intensity
11. Green Color Intensity
12. Red Color Standard Deviation
13. Green Color Standard Deviation
14. Blue Color Standard Deviation
15. Harlick Texture mean_ [1]
16. Harlick Texture mean_ [2]
17. Harlick Texture mean_ [4]
18. Harlick Texture mean_ [8]
19. Labels of images (not specific feature)

3.3.7 CLASSIFICATION

Classification refers to the process of recognizing, understanding, and grouping into required categories or "sub-categories". Using pre-categorized training datasets, machine learning programs are used to classify datasets. These classifiers are machine learning algorithms. Supervised machine learning algorithms have experimented within the model. We have researched 10 machine learning algorithms and experimented based on different parameters of result analysis.

We have experimented with 10 machine learning algorithms to identify the correct algorithm for the system which are as follow:

1) Support Vector Machine (SVM):
Support vector ma- chines (SVMs) are a set of supervised learning methods used for classification, regression, and outlier detection. SVM creates a hyper-plane that creates a boundary between the types of data. In 2-dimensional space, this hyper-plane is nothing but a line. In SVM, we plot each data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data. Next, find the optimal hyperplane to separate the data. The training algorithm maximizes the margin between training data and class boundary. The result depends on training data called support vectors. It is effective in a high- dimensional space where several dimensions are greater than the number of training data. We have used 3 types of SVM in the model for experimentation.

- Polynomial Kernel

  F(x, xj) = (x.xj+1) ^d
  Here '.' shows the dot product of both the values experimenting classification model and d denotes the degree.
  F(x, xj) representing the decision boundary to separate the given classes.

- Gaussian Radial Basis Function (RBF)
  F(x, xj) = exp (-gamma * ||x - xj||^2)

  The value of gamma varies from 0 to 1. You must manually provide the value of gamma in the code. The most preferred value for gamma is 0.1

- Sigmoid Kernel

F(x, xj) = tanh (αxay + c)

2) Gaussian Naïve Bayes:

Gaussian-NB implements the Gaussian Naive Bayes algorithm for classification. Gaussian Naive Bayes supports continuous-valued features and models each as conforming to a Gaussian (normal) distribution. The likelihood of the features is assumed to be Gaussian. When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be-

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(\frac{(x_i - \mu^y}{2\sigma_y^2}\right)$$

3) Decision Tree:

A decision tree algorithm is a super- vised learning algorithm. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. Based on the comparison we follow the branch corresponding to that value and jump to the next node. Entropy controls how a Decision Tree decides to split the data. It affects how a Decision Tree draws its boundaries. Entropy values range from"0 to 1", less the value of entropy more it is trusting.

$$\boldsymbol{Entropy = -p_i * log_2 p_i \dots .* log_2 p_n}$$

Here, n is the number of classes. Entropy tends to be maximum in the middle with a value up to 1 and minimum at the ends with a value up to 0.

4) KNN(K-nearest neighbour):

K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most like the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well-suited category by using K- NN algorithm. The K- NN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems. KNN used Euclidian distance to measure the nearest distance.

Euclidean Distance between A1 and B2 =

$$B2 = \sqrt{(X2 - X1)^2 + (Y2 - Y1)^2}$$

5) Gradient Boosting:

GB is an ensemble-based technique, and it is based on learning multiple models. The knowledge is to combine more than a few weak models into a single strong model. Gradient boosting is a mixture of the Adaboost method and gradient descent method. It enhances the differential loss function and forms the model in a forward fashion. This algorithm is highly customized for a specific application. The advantage of the Adaboost is that it boosts the outliers near classification boundaries. GB is the ML technique used for both regression and classification and helps increase the classifier's accuracy. The predictive model can be improved in two basic ways embracing either by feature engineering by applying the algorithms (boosting) straight away. There are many boosting algorithms such as Ada-boost, XG-Boost, GB, Gentle Boost, etc. Each BA has its basic math. The steepest descent direction is given by the loss negative gradient function:

$$-g_m(y) = -\left[\frac{\partial L(x, f(y))}{\partial f(y)}\right] f(y) = f^{(m-1)(y)}$$

6) Multilayer Perceptron:

The perceptron is very useful for classifying linearly separable data sets. They encounter serious limitations with data sets that do not conform to this pattern as discovered with the XOR problem. The XOR problem shows that for any classification of four points that there exists a set that is not linearly separable. The Multilayer Perceptron (MLPs) breaks this restriction and classifies datasets that are not linearly separable. They do this by using a more robust and complex architecture to learn regression and classification models for difficult datasets.

7) Logistic regression:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

$$\frac{1}{(1 + e^{-value})}$$

Where e is the base of the natural logarithms and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

$$y = e^{(b_0 + b_1 * x)/(1 + e^{b_0 + b_1 * x})}$$

8) K-means Clustering:

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), where each observation is a d-dimensional real vector, k- means clustering aims to partition the n observations into k (n) sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ to experimenting classification model minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

Euclidean Distance between A1 and B2=

$$\sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2}$$

Where μ_i is the mean of points in S_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster.

9) Adaboost Classifier:
AdaBoost is short for Adaptive Boosting and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”. The AdaBoost technique follows a decision tree model with a depth equal to one. AdaBoost is nothing but the forest of stumps rather than trees. AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well. The AdaBoost algorithm is developed to solve both classification and regression problems.

10) Stochastic gradient Descent(SGD):
Stochastic gradient is an iterative method for optimizing an objective function with suitable smoothness properties. It can be regarded as a
stochastic approximation of gradient descent optimization, since it replaces the actual gradient especially in high-dimensional optimization problems this reduces the computational burden,
achieving faster iterations in trade for a lower convergence rate. Gradient descent is an iterative algorithm that starts from a random point on a function and travels down its slope in steps until it reaches the lowest point of that function.

- Classification algorithm used in all classifiers:
  1. Fit a simple linear regression or decision tree on data (I have chosen decision tree in my code) [call x as input and y as output.
  2. Calculate error residuals. Actual target value, minus predicted target value [e]= y-y_predicted1]
  3. Fit a new model on error residuals as target variable with
     same input variables [call it e1_predicted]
  4. Add the predicted residuals to the previous predictions[y_predicted2=y_predicted1+e1_predictd]
  5. Fit another model on residuals that is still left. i.e. [e2= y-y_predicted2] and repeat steps 2 and 5 until it starts overfitting, or the sum of residuals becomes constant. Overfitting can be controlled by consistently checking accuracy on validation data.

4. RESULTS AND DISCUSSION

The proposed system dataset was implemented in Python, and the result analysis was carried out. The parameters have been tuned by the trial and error method, in machine learning algorithms. It is characterized by repeated, varied attempts, which are continued until success. Here, the plant village dataset and some own manually made datasets together combined with the dataset for the model were considered, which consists of both diseased and healthy images. For performing the experiment, the dataset was split into training and testing images which were 75% and 25% respectively. Further, the dataset is divided into 12 classes of plant species having both diseased and healthy images. The classification finally was operated by a gradient boost algorithm having max accuracy. The feature vectors made are then processed for the classification phase. A Support Vector Machine is used and 12 other algorithms have experimented with the classification model. The data is analyzed on SVM’s 3 Kernels-Lineal, RBF, and Polynomial Kernel. The performance of the proposed model was compared with different result parameters and measures. The measures such as accuracy, precision, F1 score, recall, and confusion matrix were analyzed and considered for the performance.

4.1 Pre-processing and segmentation results:

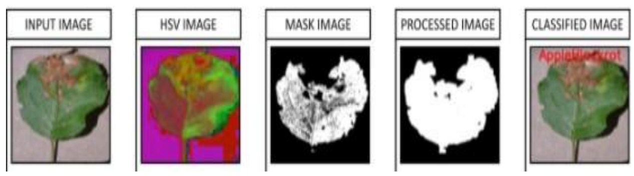


Fig-5: Qualitative analysis on Apple Blackrot disease image

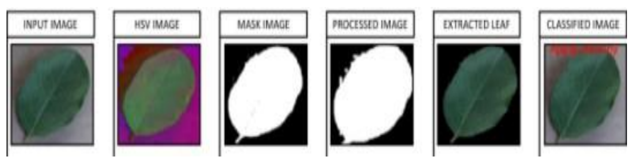


Fig-6: Qualitative analysis on Apple Healthy image

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

4.2 Performance measures:

For performance measures, true positive (TRP), true negative (TRN), false positive (FAP), and false negative (FAN) parameters are considered which are as follows:

a) Precision: It is "the ratio of positive observations that are predicted exactly to the total number of observations that are positively predicted".

$$Pre = \frac{TRP}{TRP + FAP}$$

b) F1 score: It is defined as the "harmonic mean between precision and recall. It is used as a statistical measure to rate performance".

$$F1 = \frac{Sen * Pre}{Pre + Sen}$$

c) Sensitivity: It measures "the number of true positives, which are recognized exactly".

$$Sen = \frac{TRP}{TRP + FAN}$$

d) Accuracy: The term accuracy is the "ratio of the observation of exactly predicted to the whole observations".

$$Acc = \frac{TRP + TRN}{TRP + TRN + FAP + FAN}$$

Table-4: Overall performance analysis and quantitative analysis of all classifiers:

Classifiers	F1 Score	Precision	Recall
SVM_RBF	0.039	0.276	0.143
SVM_Poly	0.035	0.02	0.141
SVM_Sig	0.035	0.02	0.141
GNB	0.418	0.484	0.448
DT	0.705	0.703	0.706
KNN 3	0.386	0.404	0.389
KNN 5	0.412	0.415	0.418
KNN 7	0.421	0.421	0.428
SGD	0.11	0.1	0.176
GB	0.801	0.803	0.803
MLP	0.161	0.521	0.239
LR	0.391	0.368	0.455
k-mean	0.04	0.086	0.026
Adaboost	0.295	0.424	0.327

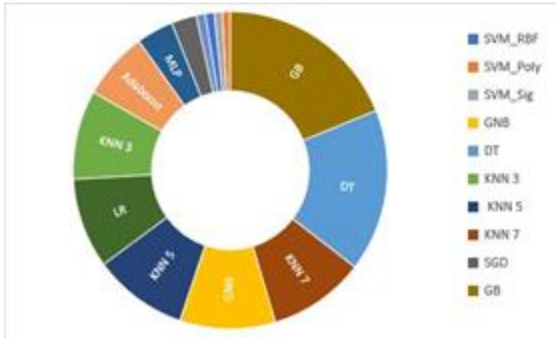


Chart-1: Sunburst chart performance analysis of above Table 4

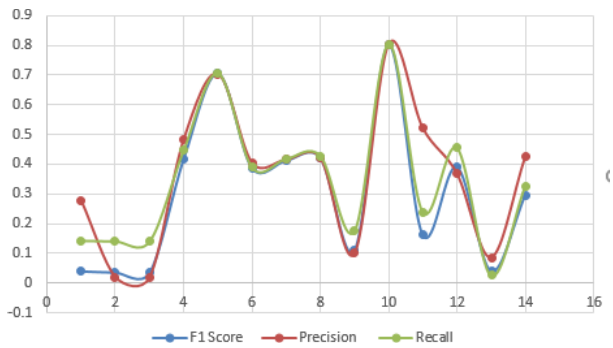


Fig-7: Sunburst chart performance analysis in the form of graph

4.3 Accuracy Comparison

Table-1: Accuracy Measures

Classifiers	Apple	Corn	Grapes	Tomato	Combined
SVM_RBF	63.06	41.79	44.74	41.9	14.96
SVM_Poly	63.06	39.55	14.7	41.73	14.31
SVM_Sig	63.06	41.79	44.74	41.73	14.31
GNB	76.78	83.99	68.31	53.79	44.71
DT	87.53	95.76	74.9	82.88	69.89
KNN 3	74.02	82.8	63.87	63.32	38.58
KNN 5	73.57	83.86	66.53	65.43	41.44
KNN 7	75.22	83.73	66.66	65.85	41.97
SGD	63.06	64.55	58.68	58.09	21.64
GB	94.59	98.54	85.17	88.02	80.02
MLP	39.33	74.07	56.78	52.36	25.3
LR	70.57	75.79	66.66	68.71	43.71
K-means	20.27	18.78	13.3	38.36	1.91
Adaboost	73.72	73.28	71.3	73.03	31.63

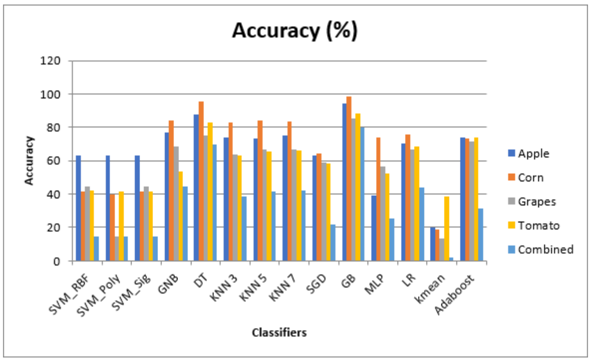


Fig-8: Accuracy Comparison Chart

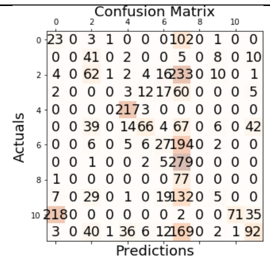
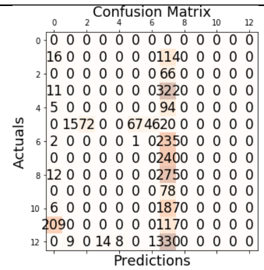
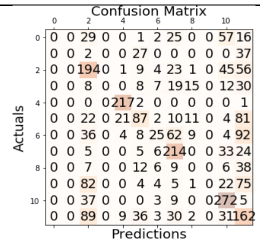
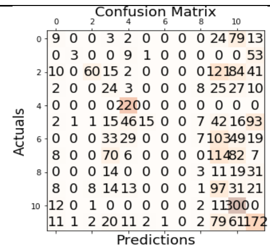
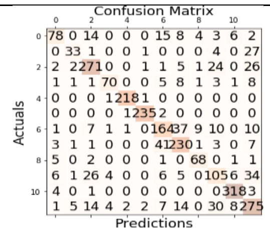
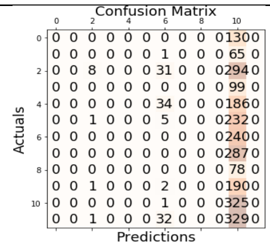
4.4 Confusion Matrix

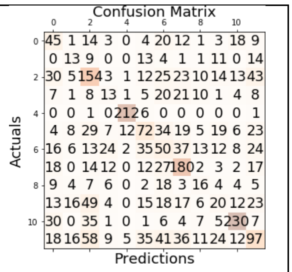
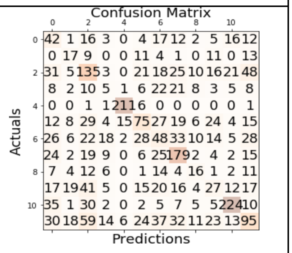
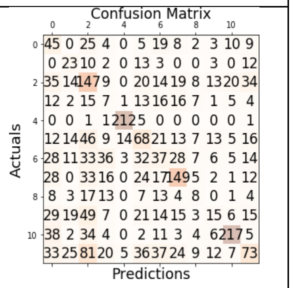
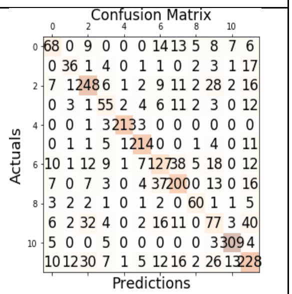
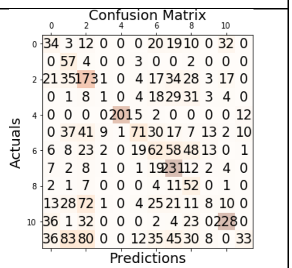
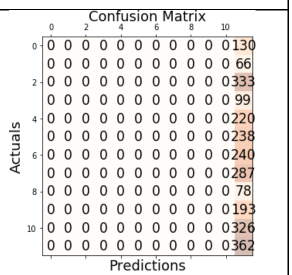
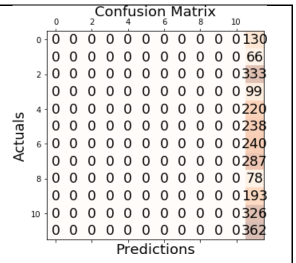
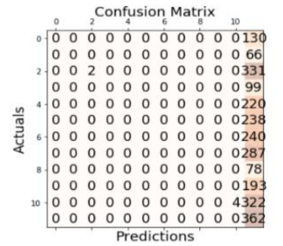
Following are the confusion matrix of all classifiers experimented with:

Results are being calculated based on qualitative and quantitative analysis. Different result analyses are being carried out to test the performance of the system. In qualitative analysis, pre-processing and segmentation methods are operated on the whole dataset of 12 plant species images for experimentation. The above **fig.5. & fig.6.** Shows the pre-processing and segmentation of apple black rot and apple healthy images. Thus the qualitative analysis is executed.

In quantitative analysis, the accuracy of all machine learning algorithms is computed. The comparison between the accuracy of the 14 classification algorithms are represented in **Fig-8 & Table-5**, confusion matrix of all algorithms are evaluated to compare among the 14 algorithms for features extraction selection and it reveals in that in each case, gradient boost algorithm outperforms all other algorithms in detecting and classifying the diseases. For performance analysis, measures like F1 score, precision, recall, and confusion matrix are calculated. On further comparison of all the measures, the f1 score of SVM sigmoid is lowest at 0.034 highest of the gradient boosting algorithm at the value of 0.801. The gradient boosting algorithm is selected for the model as it gives the highest accuracy 80.02% as the lowest was 14% by the SVM polynomial algorithm.

Table-6: Confusion matrix

Sr.No	Algorithms	Confusion Matrix
1	Adaboost Classifier	
2	K-Mean clustering	
3	Logistic Regression	
4	Multi-Layer Perceptron	
5	Gradient Boosting	
6	stochastic gradient descent	

7	K nearest Neighbours (7)	
8	K nearest Neighbours (5)	
9	K nearest Neighbours (3)	
10	Decision Tree	
11	Gaussian Naive Bayes	
12	SVM (SIGMOID)	
13	SVM (POLYNOMIAL)	
14	SVM (RBF)	

Suggestion of remedies are done on a research basis based on factual information. There are 12 classes of plant species where the model is trained. According to the classes, sugges- tions of remedies are given which are as follows:

1.Apple balckrot class = Captan , Lime sulphur, Copper Fungicide, Oil Fungicides
2.AppleCedarapplerust class = Chlorothalonil, Copper, mancozeb,Myclobutanil, propiconazole, Sulphur
3.Apple Healthy class = No remedies Required
4.Corn(maize) Cercosporaleafspot Grayleafspot = Foliar spray, tabuconazole, Methomyl, Sporrin
5.Corn(maize) Common Rust = Oxadixyl,Urea fungicide, Vinclozolin, Nystatil
6.Corn(Healthy) = No remedies Required
7.GrapeBlackrot =Luna fungicide,sercadis plus, acrobat-complete, signum
8.GrapeEsca(BlackMeasles) =Acrisio, cabrio, merivol, fluridone
9.Grapeleaf(Healthy) = No remedies Required
10.TomatoEarlyblight = Carbendazim, hexaconazole, met- alaxyl, Benomly
11.Tomato(Healthy) = No remedies Required
12.TomatoLateblight = Mancozep, maneb, kitazil, tride-morph

5. CONCLUSION AND FUTURE SCOPE

In this paper, different crop disease classification techniques are surveyed. The main motive was to develop an AI-enabled system for the detection of crop disease as an application of smart agriculture to monitor the disease, suggest correct remedies, and improve a wide range of automated agriculture. Thus artificial Intelligence is adopted which makes the use of machine learning for the classification of crop leaf diseases. This proposed system successfully recognized the diseases of the crop and implied the recommendation of pesticides accordingly. Among all the classifiers experimented with, the Gradient boosting classifier model gives 80.02% accuracy for a database used with the GLCM approach. The healthy and diseased crops can be distinguished with the help of this model. Apple, Corn, Grapes, and Tomato species are operated for testing the model. Hence, relevant diseases for these crops are taken for identification. In the future, this experimenting system can be improved by reducing the number of features used for the classification process keeping the classification accuracy rate higher, also to acquire less computational time and high precision, Deep Learning can be used. This work can be widened by integrating a huge number of diseases of various species of crops.

6. REFERENCES

[1] S. Ingale, V. B. P. B. and, "Plant Leaf Disease Detection Recognition using Machine Learning", Interna- tional Journal of Engineering Research & Technology (IJERT) 8 (6) (2019) 1179–1182.

[2] R. Prakash, G. P. Saraswathy, G. Ramalakshmi, K. H. Mangleshwari, T. Kaviya, "Detection of leaf Diseases and Classification using Digital Image Pro- cessing", 2017International Conference on Innova- tions in Information, Embedded and Communication Systems (ICIIECS) (2017).

[3] Anand.H.Kulkarni, A. P. R. K, "Applying image processing technique to detect plant diseases", Interna- tional Journal of Modern Engineering 2 (5) (2012) 3661–3664.

[4] S. Arivazhagan, R. N. Shebiah, S. Ananthi, S. V. Varthini, "Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features", Agric Eng Int: CIGR Journal15 15 (1) (2013) 211–217.

[5] K. R. N. V. V. D. Aravind, S. P. Shyry, Y. Felix, "Classification of Healthy and Rot Leaves of Apple Using Gradient Boosting and Support Vector Classi- fier", International Journal of Innovative Technology and Exploring Engineering 8 (12) (2019) 2868–2872.

[6] P. Sharma, P. Hans, S. C. Gupta, "Classification Of Plant Leaf Diseases Using Machine Learning And Image Preprocessing Techniques", 10th International Conference on Cloud Computing (2020) 480–484.

[7] Xiao, Tong, Liu, et al., "Corn Disease Identification Based on improved GBDT Method", 6th Interna- tional Conference on Information Science and Control Engineering (ICISCE) (2019) 215–219.

[8] W. Huang, Q. Guan, J. Luo, J. Zhang, J. Zhao, D. Liang, L. Huang, L. Huang, D. Zhang, "New Optimized Spectral Indices for identifying and Monitoring Winter Wheat Diseases" (2014).

[9] S. B. Patil, et al., "LEAF DISEASE SEVERITY MEASUREMENT USING IMAGE PROCESSING", International Journal of Engineering and Technology 3 (5) (2011) 297–301.

[10] S. N. Ghaiwat, P. Arora, "Detection and Classification of Plant Leaf Diseases Using Image processing Techniques: A Review", International Journal of Recent Advances in Engineering & Technology 2 (3) (2014) 15–19.

[11] S. Savary, A. Ficke, , et al., Crop losses due to diseases and their implications for global food pro- duction losses and food security, in: Food Security, Springer Science+Business Media B.V. & Interna- tional Society for Plant Pathology, 2012, pp. 4–4.

[12] P. S. B. Dhaygude, M. N. P. Kumbhar, "Agricultural plant Leaf Disease Detection Using Image Processing", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 2 (1) (2013) 599–602.

[13] P. Kan-Rice, et al., "Pests and pathogens place global burden on major food crops" (2019).

[14] R. M. Badnakhe, P. R. Deshmukh, " An Applica- tion of K-Means Clustering and Artificial Intelligence in Pattern Recognition for Crop Diseases", International Conference on Advancements in Information Technology20 20 (2011) 57–65.

[15] S. Bashir, N. Sharma, "Remote Area Plant Disease Detection Using Image Processing", IOSR Journal of Electronics and Communication Engineering 2 (6) (2012) 31–34. doi:10.9790/2834-0263134.

[16] Piyush Chaudhary et al. "Color Transform Based Approach for Disease Spot Detection on Plant Leaf", International Journal of Computer Science and Telecommunications, 3(6)(2012)65-70.

[17] Smita Naikwadi, Niket Amoda" ADVANCES IN IMAGE PROCESSING FOR DETECTION OF PLANT DISEASES", International Journal of Application or Innovation in Engineering & Management, 2(11)(2013)168-175.

[18] Arti N. Rathod, Bhavesh Tanawal, Vatsal Shah" Image Processing Techniques for Detection of Leaf Disease", International Journal of Advanced Research in Computer Science and Software Engineering, 3(11)(2013)397-399.

[19] Sanjay B. Patil et al. "LEAF DISEASE SEVERITY MEASUREMENT USING IMAGE PROCESSING", International Journal of Engineering and Technology, 3(5)(2011) 297-301.

AUTHOR BIOGRAPHY


PRASAD SALWE


AVANTI SAKHARKAR


RUDRANI SHETE


MANISHA KOWDIKI