

Analysis of Crop Disease Detection with SVM, KNN and Random Forest Classification

Rupali Kale and Sanjay Shitole

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

December 20, 2022

Analysis of Crop disease detection with SVM,KNN and Random forest classification

MRS. RUPALI KALE

Department of Computer Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, India

MR. SANJAY SHITOLE

Department of Computer Science and Technology, Usha Mittal Institute of technology, SNDT University, India

Due to an uneven climatic condition crops are being affected which leads to decrease in agriculture yield. It greatly affects global agricultural economy. However, the condition becomes more worse when diseases are identified in crops. Agriculture plays a vital role in every country's economy. Thus, there is a need to identify the crop disease before it is visible on a crop so that disease can be avoided by using appropriate measures. The traditional way of identifying a crop disease is through observation by naked eyes. But as it requires large number of experts and continuous monitoring of crop it will be costly for large fields. Hence, an automatic system is required which can not only examine the crops to detect disease but also can classify the type of disease on crops. The proposed system determines disease from an input image. The input image has to go through following stages: Image Acquisition, Image pre-processing, Image segmentation, Feature Extraction, and Classification in order to determine diseased crop and accordingly provides remedy for that disease. Infected crop image is taken as input in Image Acquisition stage. In Image pre-processing stage noise is removed from the input image by applying gaussian blur filter and non-local means denoising algorithm. Also, the background of image is eliminated using Thresholding algorithm. To extract Region of Interest (ROI) i.e. infected region from input image K-means Clustering algorithm is used. Then color, texture and shape features are extracted from ROI and features are further send to the classification stage. Three different classification algorithms namely Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Random Forest are implemented for classification out of which Support Vector Machine Algorithm is found to be best in terms of accuracy. Hence, classification is carried out by using Multivariate Support Vector Machine algorithm which detect disease present in crop accurately. In this way, the proposed system detects a disease from the given input image.

Keywords: Image Acquisition, Image pre-processing, Image Segmentation, Feature Extraction, K-means Clustering Algorithm, Support Vector Machine, K-Nearest Neighbors, Random Forest

1. INTRODUCTION

In a developing world 80% of the agriculture is done by small holder farmers and nearly 50% of crops are destroyed yearly by diseases. It may damage crop, lower the quality of fruit and sometime even wipe out the entire harvest. Thus, detection of plant disease at an early stage plays a vital role. The traditional method of crop disease detection is through naked eye observation. But it will be very costly to monitor large farms as it requires continuous monitoring of crops and large number of experts. Thus, farmer must be able to identify the crop disease just before it is visible on a crop. This can be done by automatic detection of crop disease just by seeing symptoms on a crop. It will be easier as well as cheaper for large farms.

Image processing is a technique which is used to perform certain operations on an image, in order to enhance the image or extract useful information from it. Advancement of new technology like Image processing has a vast application in biological field. Most of the farmers are small and don't have much resources. Due to lack of resources they are not able to improve agricultural productivity. Also, most of them may don't know the amount of fertilizer to be used which will lead to an unbalanced use of fertilizer and it may affect the crop.

A user-friendly web application may help farmer to identify crop disease. The farmer has to click a picture of crop and upload it on a system. The image is first pre-processed in order to remove the noise and enhance an image. Then the pre-processed image is segmented in order to get region of interest. After segmentation the features are extracted from the region of interest. These features are then fed to classification algorithm so as to classify the disease. Thus, the farmer will be able to see the disease present in a crop and the amount of pesticides/insecticides required to control the disease. This method will be cheaper as well as beneficial for large farms.

2. LITERATURE REVIEW

Manisha Bhange, H.A. Hingoliwalab at Reference [1] proposed a web-based tool that helps farmers for identifying fruit disease by uploading fruit image to the system. Segmentation is done by using K-means algorithm and features are extracted on parameters such as color, morphology, and CCV. Next, SVM is used for classification to classify the image as infected or non-infected. Out of three features extracted best results are obtained using morphology.

Shashank Chaudhary, Upendra Kumar, Abhishek Pandey at Reference [2] focused mainly on the leaf diseases. The paper presents a comprehensive view on the various researches done in contemporary domain of Crop diseases. Here, the analysis was done by segmentation techniques such as Otsu's and K-means clustering.

Ch Ramesh Babu, Dammavalam Srinivasa Rao, V. Sravan Kiran, N. Rajasekhar at Reference [3] presents the k-mean clustering procedure for the segmentation of input images. The GLCM (gray-level co-occurrence matrices) procedure is utilized which extracts textural features from the input image. Also implemented KNN (k-nearest neighbors) algorithm for image classification and produced classification accuracy from 70 to 75% for different inputs.

Mrunmayee Dhakate, Ingole A. B. at Reference [4] proposed methodology uses Artificial Neural Network (ANN). Feature extraction is done by using Grey level co-occurrence matrix (GLCM) approach and trained the artificial neural network using Back propagation Algorithm. In the beginning, several images have been used for training of neural network and remaining images are utilized as test images to test accuracy from the outcomes. The mean accuracy of proposed methodology is to be about 90%.

Savita N. Ghaiwat at Reference [6] presents survey on different classification techniques that can be used for plant leaf disease classification. For given test example, k-nearest-neighbor method is seeming to be suitable as well as simplest of all algorithms for class prediction. SVM was found competitive with the best available machine learning algorithms in classifying high-dimensional data sets.

Pooja Kantale, Shubhada Thakare at Reference [7] presents comparative study on different disease detection techniques for pomegranate. Segmentation is done by k-means algorithm and features are extracted based on the parameters such as color, morphology, texture, etc.

Johnny L. Miranda, Bobby Dioquino Gerardo, Bartolome T. Tanguilig Iii at Reference [8] presents an automatic detection and extraction system, different image processing techniques were used to detect and extract the pests in the captured image. The authors used background modeling to detect the presence of insect pests in the captured image, and a median filter was used to remove the noise produced by different lighting conditions.

Rupesh G. Mundada, Dr. V. V. Gohokar at Reference [10] proposes a software prototype system for early pest detection on the infected crops in greenhouse. Images of the infected leaf are captured by a camera with pan tilt and zoom and processed using image processing techniques to detect presence of pests. This paper contains all the information about features like entropy, mean, standard deviation, contrast, energy, Correlation, eccentricity and other features used for feature extraction. SVM classifier is used for classification.

Saradhambal G., Dhivya.R, Latha.S, R. Rajesh at Reference [13] implements an innovative idea to identify the affected crops and provide remedy measures to the agricultural industry. Disease detection involves steps like image acquisition, image preprocessing, image segmentation, feature extraction and classification. Proposed system is used to detect the plant diseases and provide solutions to recover. It also has voice navigation system, so a person with lesser expertise in software should also be able to use it easily. Sharath D M, Akhilesh, S Arun Kumar, Rohan M G and Pratap C. at Reference [14] proposed a system for computerized leaf or fruit identification and disease classifier. There are two stages, first is Training stage in which image acquisition, pre-processing, feature extraction, segmentation, classification and calculation of percentage infection is performed. After segmentation using grab cut segmentation (GCS), canny edge detection (CAD) is used for identification of the disease stiff region in the fruit. Author has created his own dataset by downloading images from google images(2598) and done processing to detect the diseased leaf, which increased the accuracy up to 31%. They evaluated the performance with ResnetV2 and MobileNet models.[17] David Ireri et al [1] developed a computer vision system for defect discrimination and grading in tomatoes using machine learning and image processing. Defected regions were detected by an RBF-SVM classifier using the LAB color-space pixel values. A combination of color and texture features achieved the highest accuracy in all the grading categories in image features evaluation. However, the grading accuracy decreased as the number of grading categories increased.[18]

Piedad, *et al.* [19] applied three machine learning methods such as Artificial Neural Network (ANN), Support Vector Machine (SVM) and Random Forest are used for the tier based classification of banana. The four features including color and shape features of banana are used for the classification of banana. The three machine learning methods provides the considerable performance in the classification. The Random Forest classifier has the higher accuracy in the classification of banana compared to other two models. The three models have lower performance in classifying the reject class and have considerable performance in other classes.

Arun Kumar R *et al* [20] developed a Multi Class Grading and Quality Assessment of Pomegranate Fruits Based on Physical and Visual Parameters. The research work developed a unique Effective Quality Assessment (EQA) algorithm comprising of a holistic approach towards the grading and quality assessment of pomegranate fruits. The research work also proposes a unique Effective Quality Assessment (EQA) algorithm comprising of a holistic approach towards the grading and quality assessment of pomegranate fruits. However, the developed model was not applicable for real time applications.

3. PROPOSED METHODOLOGY

The step by step description of proposed methodology is as follows:

3.1 Algorithm/Framework/Technology:

Table I. Algorithm/Framework/Technology				
Task	Algorithm/Framework/Technology			
Remove Noise from Image	Non-Local Means Denoising, Gaussian Blur Algorithm			
Remove Background	Image Thresholding Algorithm			
Extracting Region of Interest (ROI)	K-Means Clustering Algorithm			

Extracting Textural Features	Gray Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP)		
Extracting Color Features	Applied Built-In Open-CV functions (Mean, Standard Deviation) on RGB color channels		
Extracting Shape Features	Open-CV Shape Descriptor (Hu Moments)		
Feature Scaling	Standard Scalar of Scikit Learn Library		
Classification	Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest		

3.2 System Block Diagram:



Fig. 1. System Block Diagram

Block diagram shows implemented modules to accurately detect disease present in given input image of leaf or fruit:

3.2.1 Image Acquisition:

In Image Acquisition Process, Images of the infected areas of crops are captured by user using camera. This image is in RGB (Red, Green and Blue) form.

Here Apple crop is chosen for demonstration. Following first 3 images shows the diseases present in Apple crop leaf and last one image is of healthy leaf:



Leaf Black Rot Leaf Cedar Rust

Leaf Scab Healthy Leaf Fig. 2. Apple Leaf Diseases

3.2.2 Image Pre-processing:

In image pre-processing step, removal of noise and distortion from image takes place. Non-local means denoising for preserving textures algorithm of scikit-image library and Gaussian Blur algorithm is used to remove noise from image. Also, Image thresholding Algorithm is used to remove unnecessary background from input image.

Here Apple Leaf with Cedar Rust as Disease is chosen. Pre-processing of that input leaf

image is as follows:



3.2.3 Image Segmentation:

Segmentation is a process of partitioning the image into different parts where each part has its own common similarities. Here, the segmentation is carried out by using k- Means Clustering algorithm with value of k as 4. From research it is found that k-Means Clustering Algorithm gives accurate clusters for value of k=3 or k=4.

Before Segmentation, input image is converted into $L^*a^*b^*$ color space. L* corresponds to lightness of the image, a* corresponds to Red / Green value of image and b* corresponds to Blue / Yellow value of the image. We take a* component of image and perform segmentation on that component as this component gives more accurate clusters in case of leaf images as this component corresponds to Red / Green value of the image. Segmentation of pre-processed image from earlier stage is as follows:



Fig. 4. Image Segmentation

3.2.4 Separate Clusters from Segmented Image:

To extract Region of Interest (ROI) that is infected part of the leaf image, first separation of 4 clusters from segmented image takes place which is as follows:



Fig. 5. Saperation of clusters from segmented image

Here, Cluster 0 contains majority of diseased part; hence it is the Region of Interest (ROI).

3.2.5 Feature Extraction:

Feature Extraction is a process of extracting numeric features from input image based on its texture, shape and color.

a) Textural Features:

It includes Energy, Entropy, Contrast, Dissimilarity, Correlation, Homogeneity, Angular Second Moment, Inverse Difference Moment features. Here GLCM (Gray Level Cooccurrence Matrix) algorithm and LBP (Local Binary Patterns) is used to extract textural features.

b) Color Features:

It includes mean and standard deviation features of channels of color space (RGB). Used open-cv built in algorithms like 'mean' and 'StdDev' to extract color features.

c) Shape Features:

Used OpenCV shape descriptor (Hu Moments) to extract 7 different types of morphological moments from image.

Here, Features are extracted from ROI to increase accuracy in classification. Further these features undergo scaling using standard scalar of scikit-learn library.

3.2.6 Classification:

Classification is a process of classifying input image into respective class based on its features. Here, input image is accurately classified into one of the four below mentioned classes. In case of Apple crop there are 4 different classes based on diseases.

Class 0: Leaf Scab

Class 1: Leaf Black Rot

Class 2: Leaf Cedar Rust

Class 3: Healthy Leaf

Three different classification algorithms namely Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Random Forest are implemented for multivariate classification out of which SVM is found to be best in terms of accuracy. Hence, Multivariate Support Vector Machine (SVM) algorithm is used for classification in presented disease detection system.

a) Support Vector Machine Classifier:

Support Vector Machine Algorithm (SVM) is supervised machine learning algorithm which is generally used for classification task. SVM can be used for both binary classification and multivariate classification. In this paper Multivariate SVM is chosen as there are more than 2 classes that has to be predicted. SVM has an ability to handle multiple categorical and continuous data points. The objective of SVM algorithm in classification is to find a maximum marginal hyperplane in an N-dimensional space such that it distinctly classifies data points. Then Support Vectors are drawn in iterative manner on endpoints of similar featured data points so that error can be minimized. Distance between two parallel support vectors is called as marginal distance. Support vectors are placed such that marginal distance is maximum and vectors are parallel to each other. Then Hyperplane is drawn between support vectors which classifies data points into different classes. SVM is highly preferred classification algorithm as it produces significant accuracy with less computation power.

b) K-Nearest Neighbor Classifier:

K-Nearest Neighbor Algorithm (KNN) is simple supervised machine learning algorithm. KNN can be used for both regression and classification. In this paper, KNN is implemented for multivariate classification. KNN assumes the similarity between new incoming data points and available datapoints and then puts the new data points in respective class according to similarity of already available data points. The most important step in KNN is to choose value of K. Small value of K can result into overfitting of data points where as larger value of K can be computationally expensive. In this paper, value of K is chosen according to general rule of thumb which says that value of k should be odd and nearer to sqrt(n) where n is number of samples in training dataset. **c) Random Forest Classifier:**

Random Forest algorithm is popular supervised machine learning algorithm which uses decision trees for its implementation. Random Forest Algorithm can be used for both

regression and classification. In this paper, Random Forest algorithm is implemented for multivariate classification. This algorithm is based on concept of ensemble learning. Ensemble learning means combining multiple classifiers output in order to improve performance of model. Accuracy of Random Forest classifier is directly proportional to the number of decision trees that it uses in its implementation.

The advantage of Random Forest algorithm is that it increases accuracy of the model and eliminates overfitting of the dataset.

GUI Demonstration of presented system of disease detection is as follows:



Fig. 6. GUI Demonstration of presented Disease Detection System

In this GUI, from the 4 clusters of input image cluster 1 is ROI whose number is manually entered by user resulting into prediction of respective disease.

4. RESULTS AND DISCUSSIONS

Results and their analysis are important part of any research work. In this presented work plant leaf disease dataset is gathered from "Kaggle – Plant Village Leaf Dataset". But in this paper demonstration is shown for only Apple crop leaf diseases. First dataset is partitioned into two sets: training dataset and testing dataset based on 80:20 rule. Then as discussed above, this dataset goes through different stages like image pre-processing, image-segmentation, feature extraction and classification. In Image Segmentation stage, K-means Clustering algorithm is used with value of k=4 because more accurate clusters are formed at this value. And hence Region of Interest is successfully extracted from infected crop image. In Classification stage, Model is trained by using 3 different classification algorithms namely Support Vector Machine Classifier, K-Nearest Neighbor Classifier and Random Forest Classifier out of which Support Vector Machine Classifier is found to be best in terms of performance. The comparative performance analysis of these 3 classification algorithms is as follows:

Performance Matrices:

To analyze the performance of classification algorithms various methods can be used. Here Confusion Matrix is used to analyze performance of the model.

Confusion Matrix:

Confusion Matrix is a N x N matrix where N is number of classes. It is a table with 4 different combinations of predicted and actual values. That combinations are based on nature of prediction that is positive or negative and output of prediction that is true or false. 4 different combinations are True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Confusion Matrix is used for performance

measurement of classification algorithms. Many performance parameters like Precision Score, Recall Score, F1 Score are dependent on Confusion Matrix. As discussed in classification section, here there are 4 different classes as Apple crop have 4 different leaf diseases. Hence confusion matrix will be of size 4 x 4.

Confusion matrix for Support Vector Machine Classifier is as follows:



Fig. 7. Confusion Matrix for SVM Classifier

Confusion matrix for K-Nearest Neighbor and Random Forest Classifier is as follows:



Fig. 8. Confusion Matrix for KNN Classifier



Here, columns are predicted values and rows are actual values for 4 different classes. From the above confusion matrices, performance parameters can be determined like Precision Score, Recall Score and F1 Score.

a) Precision Score:

Precision Score is a measure of actually positive classes out of total correctly predicted positive classes.

Precision Score = TP / (TP + FP)

b) Recall Score:

Recall Score is a measure of correctly predicted classes out of total positive classes. Recall Score = TP / (TP + FN)

c) F1 Score:

F1 Score is a measure of both Precision Score and Recall Score at a same time. It is a harmonic mean of Precision Score and Recall Score.

F1 Score = (2*Precision Score*Recall Score) / (Precision Score + Recall Score)

Comparative Performance analysis of Classifiers for each class in terms of performance parameters is as follows:

Class 0 (Leaf Scab): Table II. Values of Performance Parameters for Class 0

	SVM	KNN	Random Forest
Precision	0.95	0.71	0.77
Recall	0.56	0.56	0.47
F1 Score	0.70	0.63	0.59

Class 1 (Leaf Black Rot): Table III. Values of Performance Parameters for Class 1

	SVM	KNN	Random Forest
Precision	0.79	0.76	0.73
Recall	0.89	0.76	0.89
F1 Score	0.84	0.76	0.80

Class 2 (Leaf Cedar Rust):

Table IV. Values of Performance

Parameters for Class 2			
	SVM	KNN	Random
Precision	0.76	0.66	0.78
Recall	0.94	0.85	0.85
F1 Score	0.84	0.74	0.82

Class 3 (Healthy Leaf):

Table V. Values of Performance Parameters for Class 3

i uluiteters for cluss 5			
	SVM	KNN	Random
Precision	0.96	0.95	0.89
Recall	1.0	0.91	0.95
F1 Score	0.98	0.93	0.92

From the above comparisons, the average performance parameters score for each Classifier is as follows:

	SVM	KNN	Random
Precision	0.86	0.77	0.80
Recall	0.85	0.77	0.79
F1 Score	0.84	0.76	0.78

Table VI. Average Values of Performance Parameters

The below comparative performance analysis clearly shows that Support Vector Machine Classifier gives highest performance among three. Hence, in this paper Multivariate Support Vector Machine algorithm is used for classification.



Fig. 10. Comparative Performance Analysis of Classifiers

5. CONCLUSIONS AND FURTHER WORK

This paper is focused on detection of Crop Diseases. From dozens of crops only Apple Crop is demonstrated in this paper. First Region of Interest is successfully extracted from input infected crop image using k-means clustering algorithm with value of k as 4. Then Feature Extraction, Feature Scaling is performed on found Region of Interest. Then these features are used to train the classification model with Multivariate Support Vector Machine (SVM) Classifier which gives highest performance for testing dataset as compare to other classifiers. In this way, diseases are successfully predicted from input image of infected crop. This early stage detection of disease will result into increase in yield of crop.

Further work will be focused on implementing Convolutional Neural Network algorithm for classification and its comparative performance analysis with Multivariate Support Vector Machine algorithm. After getting perfect model, this disease detection system will be deployed using cross platform web application. Hence, user can access it from any platform at any time.

REFERENCES

[1] BHANGE MANISHA, HINGOLIWALAB H. A., Smart Farming: Pomegranate Disease Detection Using Image Processing / 2015 / ScienceDirect Publications.

[2] CHAUDHARY SHASHANK, KUMAR UPENDRA, PANDEY ABHISHEK, Crop Plant Disease Detection Using Image Processing, 2019, IJITEE.

[3] CH BABU RAMESH, RAO DAMMAVALAM SRINIVASA, KIRAN V. SRAVAN, RAJASEKHAR N., Assessment of Plant Disease Identification using GLCM and KNN Algorithms, 2020, IJRTE.

[4] DHAKATE M. AND INGOLE A. B., Diagnosis of pomegranate plant diseases using neural network, 2015, Fifth National Conference on Computer Vision.

[5] GARCIA JAYME, BARBEDO ARNAL, KOENIGKAN LUCIANO VIEIRA, Thiago Teixeira Santos (2016), Identifying multiple plant diseases using digital image processing,147pp104-116.

[6] GHAIWAT SAVITA, Detection and Classification of Plant Leaf Diseases Using Image Processing Techniques, 2014, IJRATE.

[7] KANTALE POOJA, THAKARE SHUBHADA, A review on Pomegranate Disease Classification Using Machine Learning and Image Segmentation Techniques, 2020.

[8] MIRANDA JOHNNY L., GERARDO BOBBY DIOQUINO, TANGUILIG BARTOLOME T., Pest Detection and Extraction Using Image Processing Techniques, 2014, International Journal of Computer and Communication Engineering.

[9] MORE S. AND NIGHOT M., An agro advisory for pomegranate field using wireless sensor network, 2016, International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune.

[10] MUNDADA RUPESH G., DR. GOHOKAR V. V., Detection and Classification of Pests in Greenhouse Using Image Processing, April 2013, IOSRJECE.

[11] PATIL S. S. AND THORAT S. A., Early detection of grapes diseases using machine learning and IoT, 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), Mysore

[12] ROCHA, ANDERSON, et al. Automatic fruit and vegetable classification from images, Computers and Electronics in Agriculture 70.1 (2010): 96-104.

[13] SARADHAMBAL G., DHIVYA R., LATHA S., R. RAJESH, Plant Disease Detection and Its Solution Using Image Classification, 2018, International Journal of Pure and Applied Mathematics.

[14] SHARATH D M, AKHILESH, KUMAR S ARUN, G. ROHAN M. AND C. PRATAP, Image based Plant Disease Detection in Pomegranate Plant for Bacterial Blight, 2019.

[15] SINGH V., MISRA A. K. (2017) Detection of plant leaf diseases using image segmentation and soft computing techniques, 4PP41-49, 2017.

[16] TRIPATHI M. K. AND MAKTEDAR D. D., Recent machine learning based approaches for disease detection and classification of agricultural products, 2016 International Conference on Computing Communication Control and automation (ICCUBEA), Pune

[17] DAVINDER SINGH*, NAMAN JAIN*, PRANJALI JAIN*, PRATIK KAYAL* SUDHAKAR KUMAWAT,, NIPUN BATRA, PlantDoc: A Dataset for Visual Plant

Disease Detection, IIT, Gujrat

[18]IRERI, D., BELAL, E., OKINDA, C., MAKANGE, N. AND JI, C., 2019. A computer vision system for defect discrimination and grading in tomatoes using machine learning and image processing. *Artificial Intelligence in Agriculture*, *2*, pp.28-37.

[19] Larada, J.I., Pojas, G.J. and Ferrer, L.V.V., 2018. Postharvest classification of banana (Musa acuminata) using tier-based machine learning. Postharvest Biology and Technology, 145, pp.93-100.

[20] KUMAR R, A., RAJPUROHIT, V.S. AND BIDARI, K.Y., 2019. Multi Class Grading and Quality Assessment of Pomegranate Fruits Based on Physical and Visual Parameters. *International Journal of Fruit Science*, *19*(4), pp.372-396.

Authors' Details:

Mrs. Rupali Kale (Assistant Professor),

Department of Computer Engineering, Shah & Anchor Kutchhi Engineering College, University of Mumbai, Maharashtra, India;

Mr. Sanjay Shitole

Department of computer science and technology, Usha Mittal Institute of technology, SNDT University, India