Dissertation on
PLANT LEAF DISEASE DETECTION AND IDENTIFICATION

Thesis submitted towards partial fulfilment
of the requirements for the degree of

Master of Technology inIT (Courseware Engineering)

Submitted by
SAIKAT GHOSH

EXAMINATION ROLL NO.: **M4CWE22019**
UNIVERSITY REGISTRATION NO.:**154501 of 2020-21**

Under the guidance of
Prof. Ranjan Parekh

School of Education Technology
Jadavpur University

Course affiliated to
Faculty of Engineering and Technology
Jadavpur University
Kolkata-700032
India

2022

M.Tech IT (Courseware Engineering)

DEDICATION

To my

Respected

Guideand

Beloved Parents

Course affiliated to
Faculty of Engineering and Technology
Jadavpur University
Kolkata, India

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled **"PLANT LEAF DISEASE DETECTION AND IDENTICICATION"** is a bonafide work carried out by **SAIKAT GHOSH** under our supervision and guidance for partial fulfillment of the requirements for the degree of Master of Technology in IT (Courseware Engineering) in School of Education Technology , during the academic session 2021-2022.

R. Pareth 24/6/2022

SUPERVISOR
School of Education Technology
Jadavpur University,
Kolkata-700 032

Professor
School of Education Technology
Jadavpur University
Kolkata - 700 032

R. Pareth 24/6/2022

DIRECTOR
School of Education Technology
Jadavpur University,
Kolkata-700 032

Director
School of Education Technology
Jadavpur University
Kolkata - 700 032

Subeny Chakraborty 24/06/2022

DEAN - FISLM
Jadavpur University,
Kolkata-700 032

Dean
Faculty of Interdisciplinary Studies, Law & Management
Jadavpur University, Kolkata-700032

M.Tech. IT (Courseware Engineering)
Course affiliated to
Faculty of Engineering and Technology
Jadavpur University
Kolkata, India

<u>CERTIFICATE OF APPROVAL</u> **

This foregoing thesis is hereby approved as a credible study of an engineering subject carried out and presented in a manner satisfactory to warranty its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not endorse or approve any statement made or opinion expressed or conclusion drawn therein but approve the thesis only for purpose for which it has been submitted.

Committee of final examination for evaluation of Thesis

** Only in case the thesis is approved.

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research work by the undersigned candidate, as part of his **Master of Technology in IT (Courseware Engineering)** studies.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by this rule and conduct, I have fully cited and referenced all materials and results that are not original to this work.

NAME: SAIKAT GHOSH

EXAMINATION ROLL NUMBER: M4CWE22019

THESIS TITLE: PLANT LEAF DISEASE DETECTION AND IDENTIFICATION

SIGNATURE: Saikat Ghosh DATE: 24/06/2022

Acknowledgement

I feel fortunate while presenting this dissertation at School of Education Technology, Jadavpur University, Kolkata, in the partial fulfilment of the requirement for the degree of Master of Technology in IT (Courseware Engineering).

I hereby take this opportunity to express warm thanks from the core of my heart to my guide and mentor Dr. Ranjan Parekh for being so kind and supportive throughout the course period. His advice and guidance have paved the path of success of this thesis. He has encouraged me and has guided me with all possible suggestions and constructive criticism. I will remain grateful to him, for my entire life.

I would like to express my warm thanks to Prof. RANJAN PAREKH, Director of School of Education Technology for her support and advices. I would also like to thank Dr. Matangini Chattopadhyay, Dr. Saswati Mukherjee and Mr. Joydeep Mukherjee. My thanks and appreciation goes to my classmates from M.Tech IT (Courseware Engineering) and MMD. I do wish to thank all the departmental support staffs and everyone else who has different contributions in this dissertation.

Finally, my special gratitude to my mother and Anu who stood for me at all my odds and have supported me well enough to achieve this height.

Date: 24/06/2022

Place: KOLKATA

Saikat Ghosh

SAIKAT GHOSH

M.Tech IT (Courseware Engineering)

School of Education Technology

Jadavpur University

Kolkata-700032

LIST OF FIGURES

LIST OF TABLES

\

TABLE OF CONTENTS

Executive Summary

Agriculture is the most common occupation in India and plays an important role. It provides sustenance for all humans, even if the population grows rapidly. Agriculture is the main source of income for rural people, but when plants are in the developing stage, they become infected with a variety of unusual diseases. These disorders are difficult to detect with the naked eye. With the advancement of technology, there is a need to develop procedures for quickly identifying diseases that are not visible to the naked eye.
Identifying plant diseases is essential for preventing yield and quantity losses of the rural product.

The idea behind the paper is to carry awareness among the farmers and to reduce sicknesses in plant leaves. It is very difficult to identify plant diseases manually as it requires an excessive processing time. As a result, image processing is used for the detection of plant
diseases. Disease detection involves the steps like image acquisition, pre-processing, segmentation, feature extraction and classification. This paper mentioned a few segmentation and feature extraction algorithms used inside the plant disease detection.

In acquisition of images, the samples of Plant leaves having problems are considered. With those samples of sick leaves, farmers will easily find the illnesses based on the early signs. First of all in pre-processing, the samples of leaves are resized and then Histogram Equalization is used to improve the quality of the samples and then image segmentation is accomplished. Then descriptors viz., Principal Component Analysis (PCA), Grey Level Co-occurrence Matrix(GLCM), Mean, Standard Deviation(SD), Entropy, RMS, Variance, Kurtosis and Skewness are used to extract informative features from leaf samples. Ultimately, the extracted features are categorized using machine learning approaches which include K-Nearest Neighbor (K-NN) and Quadratic Support Vector Machine(Q-SVM). To avoid the problems of having two or more one-of-a-kind diseases that could have identical chances of having similar

diseases, an extra inspection step is added to the algorithm at the final stage. Here, at the final stage the extra inspecting step checks for a newly added feature i.e; Speeded Up Robust Feature (SURF) which is calculated and Classified Perfectly.

With an average accuracy of 98.91 percent, the proposed system could correctly detect and categories the examined disease. MATLAB software is used to implement the proposed system.

1.INTRODUCTION

1.1 OVERVIEW

Agriculture productivity is a key factor in this financial prosperity of India. The productivity of the agricultural sector accounts for between 70 and 80 percent of the Indian financial system. Due to the significant supplies of food, agriculture is given a unique priority by all governments, whether in developing or advanced nations.

There are several environmental elements that have an affect on agriculture, including microorganisms, viruses, fungi, and non-biotics that are mostly based on water, air, temperature, and other environmental factors. Therefore, damage to the environment's vegetation might result in a significant loss in output, which would ultimately affect the economy.

For portraying early indications and symptoms, leaves might be the most important component of many flowers. Numerous illnesses affect flowers and result in significant manufacturing losses, posing a threat to people's ability to go about their daily lives normally. To find the extremely common plant illnesses that have a serious impact on performance deterioration, we are utilizing a human visual examination with our unaided eyes. This is the issue that is now most prevalent and common.

This conventional method leaves a lot of room for errors, with farmers trying to find the leaf-based disease through visual inspection as a high risk of errors in some cases turning to experts. Many laboratory-based techniques, like polymerase chain reaction, reduction in food production, pest management, and hyper spectral techniques, are known for detecting illnesses but they are very time-consuming and expensive.

Another issue is that the majority of the produce is wasted. Fields is situated in a rural location where farmers must go great distances to locate expertise. Picture processing delivers accuracy, high speed, doesn't cost a lot of money to perform, and takes more time than what is provided by pros. From the very beginning of our daily life cycle to the specific moment they are scheduled to be harvested, the vegetation has to be more closely monitored for illnesses.

The first step in the process of identifying plant diseases is to utilise the naked eye, which is a laborious and error-prone procedure. A requirement for the automated detection approach to find the plant diseases is there, which speeds up the process. Several ways have been included as a result of technological improvement to overcome the guide detection of plant illnesses.

The ability to watch a wide variety of crops and automatically identify diseases from symptoms that appear on plant leaves makes automated detection of plant illnesses a crucial research topic.

As a result, robot steering for disease control and automated diagnosis of plant disease using picture processing approach offer improved accuracy. Visible identification, in contrast, takes substantially less time and is far less accurate. This automated technology is intended to eliminate the drawbacks of manual methods. A high-resolution mobile smartphone camera or a common virtual digital camera might be used to take the photo.

This image is sent to the machine as input so it can determine the capabilities of the leaf. The process for the device includes several different processes, including segmentation, function extraction, identification, and classification.

1.2 IMAGE PROCESSING

Image processing is the process of converting a physical image to a digital representation and then conducting operations on it to extract relevant information.

When implementing specific specified signal processing algorithms, the image processing system normally treats all images as 2D signals.

Image processing can be divided into five categories:

1. Visualization - Look for objects in the image that aren't visible.
2. Object Recognition - Identify or detect objects in an image.
3. Enhancement and restoration - From the original image, create an upgraded image.
4. Pattern recognition - Examine the numerous patterns that surround the image's objects.
5. Image Retrieval - Search and browse a big collection of digital images that are comparable to the original image.

1.3 FEATURE EXTRACTION

The feature extraction phase is interceded by a sub module called feature detection in which, the best feature amongst all the features relevant and experimented in required cases of recognition is chosen to design the final version of the system.

One of the most essential characteristics that may be utilized to classify and recognize things is texture.

GLCM stands for Grey Level Co-occurrence Matrices and is a statistical approach. For texture classification, it is an ancient and widely used feature extraction approach. It is an essential texture classification feature extraction approach that computes the connection between pixel pairs in the picture. The Gray-Level Co-occurrence Matrix (GLCM) is a spatial dependency matrix with many levels. Co-occurrence is a term that is regularly used in the literature without a hyphen.

Co-occurring Gray Levels The GLCM =Gray matrix is created using Matrix (I). The grey scale intensity value I appears horizontally adjacent to a pixel with the value j in this example. The number of times a pixel with value I occurred horizontally adjacent to a pixel with value j is specified by each element (I, J) in GLCMS.

1.3.1 <u>Contrast</u>: Over the whole image, returns a measure of the intensity contrast between a pixel and its neighbor. [0 (size (GLCM, 1)-2] Range = [0 (size (GLCM, 1)-2]. For a constant picture, contrast is zero.

$$\boldsymbol{Formula} = \sum_{\boldsymbol{i,j}} |\boldsymbol{i}-\boldsymbol{j}|^{\mathbf{2}}\boldsymbol{p}(\boldsymbol{i},\boldsymbol{j}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots..\mathbf{1}$$

1.3.2 <u>Correlation</u>: Returns a metric for how closely a pixel is connected to its neighbors throughout the entire picture.

[- 1 1] is the range. A fully positively or negatively correlated picture has a correlation of one or -one. For a constant picture, correlation is null.

$$\boldsymbol{Formula} = \sum_{\boldsymbol{i,j}} (\boldsymbol{i}-\boldsymbol{\mu i})(\boldsymbol{j}-\boldsymbol{\mu j})\boldsymbol{p}(\boldsymbol{i},\boldsymbol{j})/\boldsymbol{\sigma i \sigma j} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots..\mathbf{2}$$

1.3.3 <u>Energy</u>: The sum of squared items in the GLCM Range = [0 1] is returned. For a constant picture, energy equals one. The location of the property Uniformity, uniformity of energy, and angular second moment are all terms used to describe energy.

$$\boldsymbol{Formula} = \sum_{\boldsymbol{i,j}} \boldsymbol{p}(\boldsymbol{i},\boldsymbol{j})^{\mathbf{2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..3$$

1.3.4 <u>Homogeneity</u>: Returns a value that indicates how near the GLCM's element distribution is to the GLCM diagonal. [0 1] is the range. For a diagonal GLCM, homogeneity is 1.

$$\boldsymbol{Formula} = \sum_{\boldsymbol{i,j}} \boldsymbol{p}(\boldsymbol{i},\boldsymbol{j})/\mathbf{1}+|\boldsymbol{i}-\boldsymbol{j}| \quad \ldots\ldots\ldots\ldots\ldots\ldots\mathbf{4}$$

Textural properties such as contrast, correlation, energy, entropy, and homogeneity may be computed from the produced GLCMs.

Calculate the Color feature values of Skewness, Kurtosis, Entropy, Standard Deviation, Mean, Variance, and RMS to extract the illness symptoms.

1.3.5 <u>Mean</u>: M = mean (A) gives the average of the items of A along the first array dimension with a size greater than one.

$$\boldsymbol{Formula: X = \sum_{i=1}^{n} (X_i)/n \; \ldots\ldots\ldots\ldots\ldots\ldots\ldots 5}$$

1.3.6 <u>Standard Deviation</u>: A is made up of N scalar observations for a random variable vector. The square root of the variance is the standard deviation.

Some standard deviation definitions utilize N instead of N-1 as a normalization factor, which you may indicate by setting w to 1.

$$\boldsymbol{Formula: \sigma = \sum (X_i)/n \; \ldots\ldots\ldots\ldots\ldots\ldots\ldots 6}$$

1.3.7 <u>Skewness</u>: The asymmetry of the data around the sample mean is measured by skewness.

When skewness is negative, data spreads out more to the left than to the right of the mean. The data spreads out farther to the right if the skewness is positive.

$$\boldsymbol{Formula: s = E(x-\mu)^3/\sigma^3 \; \ldots\ldots\ldots\ldots\ldots\ldots\ldots 7}$$

Where, the mean of x is μ, the Standard Deviation of x *E*(*t*) represents the expected value of the quantity *t is* σ .

1.3.8 <u>Kurtosis</u>: Kurtosis is a measure of how prone a distribution is to outliers. The normal distribution has a kurtosis of 3.

Kurtosis larger than 3 indicates that the distribution is more prone to outliers than the normal distribution; kurtosis less than 3 indicates that the distribution is less prone to outliers.

Some definitions of kurtosis deduct 3 from the calculated value, resulting in kurtosis of 0 for the normal distribution.

$$\boldsymbol{Formula}\text{: } \mathbf{k} = \boldsymbol{E}(\boldsymbol{x} - \boldsymbol{\mu})^{4} / \boldsymbol{\sigma}^{4} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..\mathbf{8}$$

Where, μ is the mean of x, σ is the Standard Deviation of x $E(t)$ represents the expected value of the quantity t.

1.3.9 <u>Entropy</u>: Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image.

$$\boldsymbol{Formula}\text{: } \mathbf{E} = \boldsymbol{sum}(\boldsymbol{p}.* \boldsymbol{log2}(\boldsymbol{p})) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \mathbf{9}$$

Where, p contains the normalized histogram counts returned from histogram equalization.

1.3.10 <u>Variance</u>: Variance is a measurement of the spread between numbers in a data set. The term variance refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean and thus from every other number in the set.

$$\boldsymbol{Formula}\text{: } \boldsymbol{V} = 1/(N-1) \sum_{i=1}^{N} |\boldsymbol{A}_{\boldsymbol{i}} - \boldsymbol{\mu}|^{2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..\mathbf{10}$$

Where, $\boldsymbol{\mu}$ is the mean of A.

$$\boldsymbol{\mu} = \mathbf{1}/\boldsymbol{N} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} (\boldsymbol{A}_{\boldsymbol{i}}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\mathbf{11}$$

1.3.11 RMS: Root Mean Square (RMS) is the square root of the mean square, which is the arithmetic mean of the squares of a group of values. RMS is also called a quadratic mean and is a special case of the generalized mean.

$$\boldsymbol{Formula}: \boldsymbol{RMS} = \sqrt{}(\mathbf{1}/\boldsymbol{N}) \sum_{n=1}^{N} |\boldsymbol{x_n}|^2 \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \mathbf{12}$$

1.4 SVM & KNN CLASSIFIER

The speed and accuracy of plant disease detection machine-learning systems are two crucial qualities that must be addressed.

KNN is one of the simplest forms of machine learning algorithms mostly used for classification. It classifies the data point on how its neighbor is classified. KNN classifies the new data points based on the similarity measure of the earlier stored data points.

The K Nearest Neighbor (KNN) classifier is a slow learner, meaning it can train and test at the same time. The KNN classifier is an instance-based classifier that classifies unknown instances by linking unknown to known examples using distance or similarity functions. It takes the K closest points and assigns the unknown instance to the majority class.

Support Vector Machine (SVM) is machine learning technique which is basically used for classification. It is a kernel-based classifier; it was developed for linear separation which was able to classify data into two classes only.

SVM has been used for different realistic problems such as face, gesture recognition, cancer diagnosis voice identification and glaucoma diagnosis.

The Support Vector Machine, or SVM, is a popular Supervised Learning technique that may be used to solve both classification and regression issues. However, it is mostly utilized in Machine Learning for Classification difficulties.

1.4.1 K-nearest Neighbors

The K-nearest neighbours (k-NN) technique is a prominent non-parametric approach for classification and regression issues.

In this case, k-NN is employed to distinguish infected leaves from healthy ones. The input to k-NN is two vectors, one from test set 1 and one from train set 1. A k-NN classifier produces a class of all Diseased leaves.

If K = 1, the item is allocated to the class of the object's single nearest neighbour. In this case, one member of test set1 is compared to all members of train set1.

The class for that member of test set1 is given to the class in train set1 with which it has the greatest number of nearest neighbours. The training dataset is used to categories a testing dataset in k-nearest neighbour classification.

The following is a description of the algorithm:

a) Determine the number of neighbours.

b) Using Euclidian distance or any other acceptable distance (in this instance Euclidian), K closest neighbours of the test data member to be categorized are chosen (say a new data point).

c) The number of data points in each category is determined among these K neighbours.

d) The new data point is assigned to the category with the most neighbours. The distance utilized here is the Euclidian distance.

Euclidean distance is defined as follows for two n-dimensional vectors P = p1, p2,..., pn and Q = q1, q2,..., qn:

$$\boldsymbol{d}(\boldsymbol{P},\boldsymbol{Q}) = \sqrt{\sum_{i=1}^{n}} (\boldsymbol{P_i} - \boldsymbol{Q_i})^2 \ldots\ldots\ldots\ldots\ldots\ldots.\mathbf{13}$$

All Diseased photos are saved as Testing Dataset 2 after K-Nearest Neighbors.

1.4.2 <u>Classifier Selection of Knn for Detection Part:</u>

a) The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions.

b) The quality of the predictions depends on the distance measure. Therefore, the KNN algorithm is suitable for applications for which sufficient domain knowledge is available. This knowledge supports the selection of an appropriate measure.

c) KNN is the better choice for applications where predictions are not requested frequently but where accuracy is important.

1.4.3 Support Vector Machine (SVM)

SVM is a supervised machine learning technique that may be used to solve classification and regression issues.

However, it is most commonly employed in classification jobs. Each data item is represented as a point in an n-dimensional space (where n is the number of features), with the value of each feature being the value of a specific coordinate.

The SVM algorithm's goal is to identify a hyperplane in an N-dimensional space that clearly classifies the input points.

The categorization is then accomplished by locating the hyper-plane with the greatest margin of difference between the two classes.

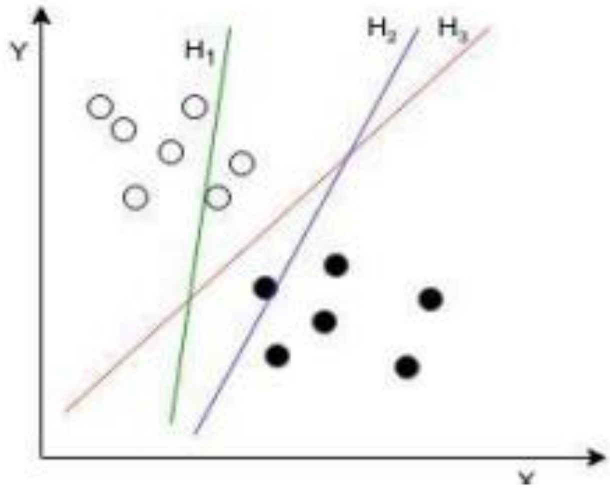


Figure1: Of the three hyperplanes (H1, H2, and H3), H3 has the greatest margin of separation from the training data.

To choose that hyperplane as the ideal hyperplane, the sum of distances between two border points of two classes must be the greatest.

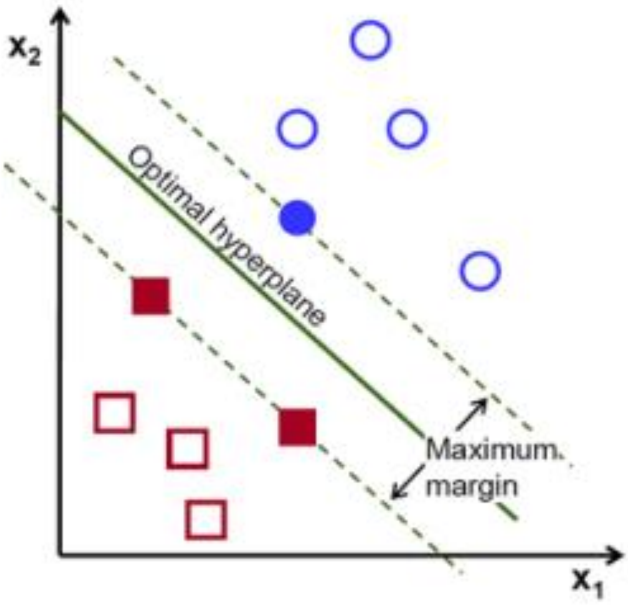


Figure2: Choice of Optimal Boundary

SVM has the benefit of dealing with points that are most unlike to those in its group. It is a risk taker, which makes it an extremely efficient algorithm. We used linear SVM in this investigation and got decent results.

1.4.4 Selection of Quadratic SVM for Classification Part:

a) To separate the data non-linearly, a dual optimization form that is a quadratic decision function is applied.

b) It is demonstrated that Q-SVM may be used in a quadratic optimization context. This configuration does not necessitate the usage of a dual form or the Kernel trick.

c) QSVM improves the performance of conventional SVM algorithms by leveraging the capabilities of quantum technology and quantum software.

The Q-SVM classification is performed between the training and testing datasets. The training feature set is used to train the Q-SVM model, while the testing feature set is used to validate the trained Q-SVM model's accuracy.

The output of the Q-SVM model thus only shows the classified diseased images.

1.4.5 Speeded Up Robust Features (SURF):

The SURF (Speeded Up Robust Features) technique is a quick and robust approach for local, similarity invariant picture representation and comparison.

SURF employs an integer approximation of the determinant of the Hessian blob detector to detect interest spots, which may be calculated with three integer operations using a precomputed integral picture. The total of the Haar wavelet response around the place of interest serves as its feature descriptor.

The SURF approach's key attraction is its quick computing of operators using box filters, which enables real-time applications such as tracking and object identification.

The SURF algorithm follows the same concepts and phases as SIFT, but the specifics in each step change. The algorithm is divided into three parts: identification of interest points, description of the nearby neighborhood, and matching.

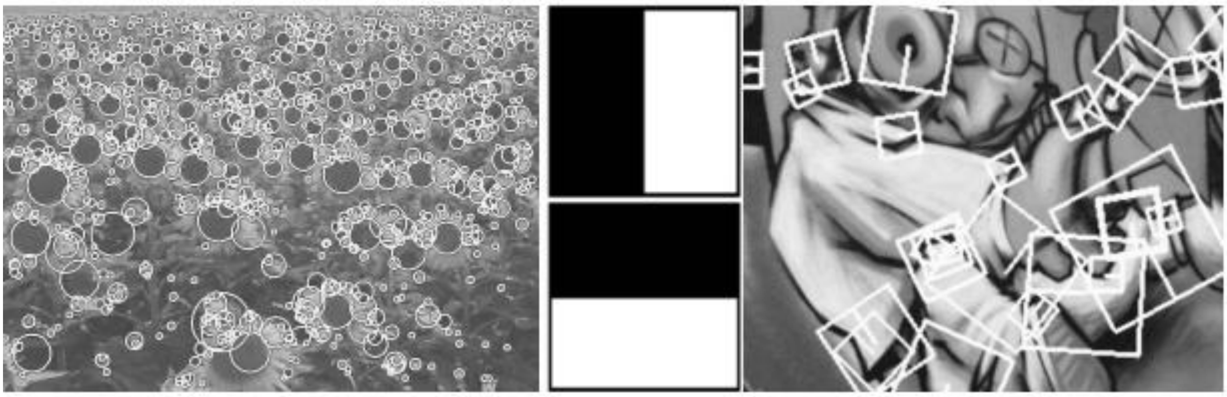


Figure 3: Detected interest points in a sunflower field on the left. This type of image vividly demonstrates the nature of the characteristics from Hessian-based detectors. SURF uses Haar wavelet types in the middle. Right: A detail of the Graffiti scene demonstrating the size of the description window at various sizes.

The first step in extracting the description is to create a square region centered on the interest point and oriented along the orientation chosen in the previous section. This change is not required for the upright variant. This window is 20s in size. Figure 3 shows examples of such square zones.

The territory is divided into smaller 4 4 square sub-regions on a regular basis. This keeps critical spatial information within. We calculate a few basic characteristics at 55 regularly spaced sample points for each sub-region. For convenience, we refer to dx as the horizontal Haar wavelet response and dy as the vertical Haar wavelet response (filter size 2s). The terms "horizontal" and "vertical" are defined in this context in respect to the orientation of the selected interest point.

To improve resistance to geometric deformations and localization errors, the responses dx and dy are first weighted with a Gaussian (= 3.3s) centred at the interest point. The wavelet responses dx and dy are then averaged over each subregion to generate the initial set of entries in the feature vector. We additionally extract the total of the absolute values of the responses, |dx| and |dy|, to bring in information about the polarity of the intensity changes.

As a result, given its underlying intensity structure, each sub-region possesses a four-dimensional descriptor vector v = (dx, dy, |dx|, |dy|).

This yields a descriptor vector of length 64 for each of the 44 sub-regions. The wavelet responses are unaffected by light bias (offset). In order to achieve contrast invariance (a scale factor), the descriptor is converted into a unit vector.

Figure 4 depicts the descriptor qualities for three unique picture intensity patterns inside a subregion.

Combinations of such local intensity patterns might result in a separate description. We experimented with fewer and more wavelet features, such as d2 x and d2 y, higher-order wavelets, PCA, median values, average values, and so on, to arrive at these SURF descriptors. The recommended sets were chosen after careful consideration.

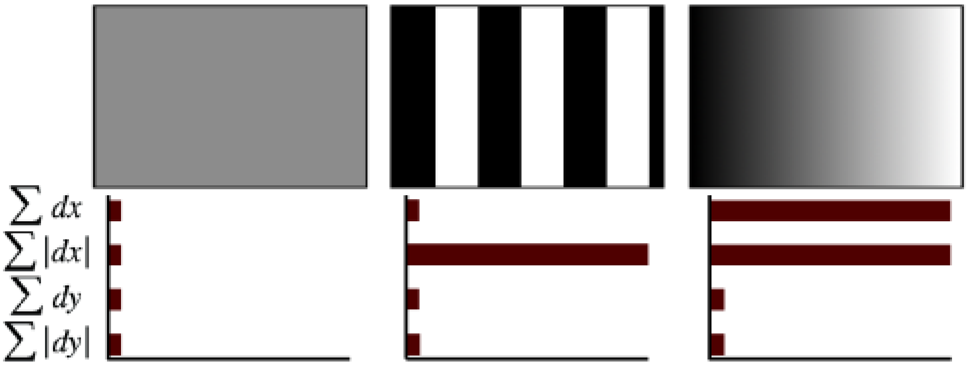


Figure 4: The nature of the underlying intensity pattern is represented by the descriptor elements of a sub-region. Left: In the event of a homogenous zone, all values are low. Middle: When frequencies in the x direction are present, the value of |dx| is high, but all others stay low. If the intensity steadily increases in the x direction, both dx and |dx| are large.

1.4.6 <u>Feature Selection of SURF</u>:

a) The SURF technique (Speeded Up Robust Features) is a quick and robust approach for local, similarity invariant picture representation and comparison.

b) In many other local descriptor-based techniques, interest points are specified as conspicuous features from a scale-invariant representation of a given picture.

c) The SURF approach's principal attraction is its quick computing of operators using box filters, which enables real-time applications such as tracking and object identification.

1.5 CHALLENGES

The following challenges were chosen as the most significant:

- The background frequently comprises components that make appropriately segmenting the region of interest where the symptoms show extremely challenging.

- Capture conditions are difficult to manage, which may result in photos with difficult-to-predict properties, making disease identification more difficult.

- Most symptoms don't have well defined boundaries, instead fading into normal tissue over time, making it difficult to distinguish between healthy and diseased areas.

- A disease's characteristics vary greatly depending on its stage of development and, in some cases, where it is found on the plant.

- Symptoms caused by many diseases may appear at the same time, either physically separated or mixed into a "hybrid" symptom that is difficult to distinguish.

- Different diseases might cause very identical symptoms, forcing approaches to rely on extremely minor distinctions to distinguish between them.

The first two challenges are external to the problem, whereas the other four are internal to it. Although the focus is on plant disease identification, the issues highlighted here are also relevant for disease severity measurement, and some references on the subject are included. In fact, the only significant difference is that while precisely describing symptoms is not always necessary for disease identification, it is essential for determining severity.

All of the other difficulties mentioned above are nearly equal in relevance for both issues, especially since illness identification may be an essential intermediate step for determining severity, especially when numerous diseases are predicted to coexist.

1.6 APPLICATIONS

Automatic detection of plant diseases is critical because it can aid in the monitoring of huge fields of crops and, as a result, automatically detect disease symptoms as soon as they emerge on plant leaves.

As a result, finding a quick, automatic, less expensive, and accurate technique of detecting plant disease cases is critical. Plant disease detection and recognition based on machine learning can provide a wealth of information for identifying and treating illnesses in their early stages.

Visual or naked eye identification of plant diseases, on the other hand, is highly costly, inefficient, imprecise, and challenging. It also necessitates the assistance of a qualified botanist.

An android application will prove to be a cost-effective solution for plant detection needs, assisting farmers in keeping plants healthy and safe while also protecting the health of both plants and people.

The creation of an Android application that will assist farmers and the general public in detecting plant diseases. The user or farmer must use his or her Android smartphone to take a picture of the sick leaf. The application's diseased leaf image will be forwarded to a database server for additional processing. The relevant features from the diseased leaf will be extracted using image feature extraction, and the results will be presented on the android device.

1.7 OBJECTIVE OF THE WORK

We can reduce insect infestations by utilizing the right pesticides and cures. By using correct size reduction techniques, we may minimize the size of the photographs while maintaining a high level of quality. We can expand the efforts of the aforementioned authors such that the system also displays the disease's cure.

The major goal is to use image processing to identify plant diseases. It also suggests the name of the insecticide to apply after the disease has been identified. It also pinpoints the insects and pests that are to blame for the outbreak. Aside from these dual goals, this drone saves a lot of time. The model's budget is quite costly for small-scale farming, but it will be cost effective in large-scale farming.

It accomplishes each of the outputs by successively completing each of the processes. As a result, the key goals are:

1) To create a system that can accurately detect crop disease and pest.
2) Create a pesticide database for each pest and illness.
3) To give treatment for the sickness that has been identified.

1.8 ORGANIZATION OF THE THESIS

Section 1 presents an introduction and overview of this thesis. Previous works are mentioned in Section 2. Section 3 presents the proposed approach. Section 4 and 5 presents experimental results and analysis & comparative analysis respectively. Section 6 is the conclusion & future scope.

2. LITERATURE REVIEWS

To improve the accuracy of the results, researchers focused on new methodologies in Machine Learning (ML) algorithms for detecting leaf illnesses. Every method is important and focuses on the path of machine learning applications, as well as difficulties faced by farmers.

[1] In the year 2019, N. Kanaka Durga and G. Anuradha presented a work titled "Plant Disease Identification Using SVM and ANN Algorithms." They employed the Histogram of Oriented Gradient (HOG) operation to forecast characteristics and feed those into the classification model in this research. Then they screen the leaves for disease and send the results to the farmer through text messaging. Take the leaves of tomato and maize plants and use SVM and ANN algorithms to identify the disease, resulting in a more efficient and accurate result. SVM and ANN classifiers are used to examine the tomato and corn crops. SVM accuracy for tomato crop is 60-70 percent, while ANN accuracy is 80-85 percent. When it comes to maize, SVM delivers an accuracy of 70-75 percent, whereas ANN gives an accuracy of 55-65 percent.

[2] In the year 2018, Nay Chi Htun's Yin Min Oo delivered a paper titled "Plant Leaf Disease Detection and Classification Using Image Processing." Using digital image processing techniques, this research offered an approach for analyzing and diagnosing plant leaf diseases. The suggested approach successfully detected and classified four important plant leaf diseases, including Bacterial Blight and Cercospora Leaf Spot, Powdery Mildew, and Rust, according to the experimental results. This system's major goal was to increase the accuracy of automatic plant disease identification. The proposed approach successfully detected and classified the plant illness with an accuracy of 88.2 percent, according to experimental results.

[3] In the year 2019, Md. Rasel Mia, Sujit Roy, Subrata Kumar Das, and Md. Atikur Rahman gave a talk titled "Mango Leaf Diseases Recognition Using Neural Network and Support Vector Machine." A Neural Network Ensemble (NNE) for Mango Leaf Disease Recognition is presented in this paper (MLDR). Mango trees are susceptible to a variety of diseases, and identifying them has been difficult in the past due to the fact that disease detection has been done manually. This research aims to use machine learning to detect the symptoms of plant illnesses more quickly than a manual monitoring system. Trained data are created using a classification technique that collects photos of leaves that have been afflicted by various diseases. A machine learning system has been developed to automatically upload and match fresh photos of afflicted leaves with learned data in order to determine the symptom of mango leaf diseases. With an average accuracy of 80%, the suggested system was able to detect and categories the illness under investigation.

[4] T. R. Ganesh Babu, S. Priya, J. Gopi Chandru, M. Balamurugan, J. Gopika, and R. Praveena delivered a paper titled "Prediction and Analysis of Plant-Leaf Disease in Agriculture Using Image Processing and Machine Learning Techniques." They used a machine learning strategy such as multi-layer feed-forward neural networks and image processing techniques in this study to forecast the causes of plant leaf disease in agriculture fields. A novel paradigm for identifying plant leaf diseases based on a deep Convolutional neural network is also presented in this research (strong CNN). The Deep CNN model is trained using a free data set of 39 different plant leaf groupings and background pictures. Picture flipping, gamma correction, injection noise, PCA colour enhancement, rotation, and scaling were among the six types of data increase procedures employed. They discovered that increasing the amount of data in the model can improve its performance.

[5] A paper on "Leaf Disease Detection and Fertilizer Suggestion" was delivered by Indumathi.R, Saagari.N, Thejuswini.V, and Swarnareka.R. In this work, the system identifies the afflicted leaf area as well as the illness that caused the leaf to be harmed. This is accomplished through the use of image processing; there are systems that can forecast leaf illnesses. To improve the accuracy of illness identification in the leaf, our

method employs K-Medoid clustering and the Random Forest algorithm. To detect the afflicted area of the leaf, the picture is first pre-processed, and then the clustering approach is used. The accuracy is measured and the sickness is found using 13 characteristics such as Entropy, RMS, Variance, Mean, SD, Smoothness, Kurtosis, Contrast, Correlation, Skewness, IDM, Energy, and Homogeneity. The system's accuracy is excellent, as is its capacity to identify illness. In this system, the time it takes to compute the illness in the diseased leaf is lowered, and the memory use is also reasonable.

[6] A paper titled "Recognition of Diseases in Paddy Leaves Using KNN Classifier" was delivered by Suresha M, Shreekanth K N, and Thirumalesh B V. In this research, geometrical parameters such as Region, Major axis, Minor axis, and Perimeter of the affected area of the leaves were used to offer a method for identifying Blast and Brown Spot illnesses. The data was classified using the global threshold approach and the KNN classifier. For the suggested technique, a score of 76.59 percent was attained.

[7] A paper titled "Plant Disease Detection Using Machine Learning" was delivered by Shima Ramesh, Niveditha M, Pooja R, Prasad Bhat N, Shashank N, Mr. Ramachandra Hebbar, and Mr. P V Vinod. The data sets developed in this article are used to identify healthy and sick leaves using Random Forest. The proposed study contains many implementation steps, including dataset construction, feature extraction, classifier training, and classification. To categories the infected and healthy photos, the produced datasets of diseased and healthy leaves are combined and trained under Random Forest. We utilize the Histogram of an Oriented Gradient to extract characteristics from a picture (HOG). Overall, we can identify illness in plants on a massive scale by utilizing machine learning to train big data sets that are publicly available. In terms of accuracy, the algorithm was put up against other machine learning models. The model was trained with 160 photos of papaya leaves using a Random forest classifier. With about 70% accuracy, the model could categories.

[8] A study titled "Classification of Chili Leaf Disease Using the Gray Level Co-occurrence Matrix (GLCM) and the Support Vector Machine (SVM) Methods" was delivered by Yuslena Sari, Andreyan Rizky Baskara, and Rika Wahyuni. The Gray

Level Co-occurrence Matrix (GLCM) feature extraction approach is used in this work to construct a method for identifying chili leaf disease into a classification system. The Support Vector Machine (SVM) approach was then used to classify the data. The total accuracy of the output categorization of illness diagnoses in chili was 88 percent. The results show that the Gray Level Co-occurrence Matrix (GLCM) and Support Vector Machine (SVM) methods for extracting features may be used to diagnose chili plant illness.

[9] A presentation titled "Prediction of Apple Leaf Diseases Using Multiclass Support Vector Machine" was delivered by Soarov Chakraborty, Shourav Paul, and Md. Rahat-uz-Zaman. This method effectively distinguishes between infected and non-infected apple leaves. Preprocessing the picture with image processing techniques such as the Otsu thresholding algorithm and histogram equalization is the first step in the identification process. The disease type from the original leaf picture is recognized with 96 percent accuracy using the image segmentation region of the diseased portion, and a Multiclass SVM detects the disease type from 500 photographs using the image segmentation region of the infected part. It also shows the proportion of that sick apple leaf image's entire affected area.

[10] A paper titled "Classification of Diseased Potato Leaves Using Machine Learning" was delivered by Sakshi Sharma, Vatsala Anand, and Swati Singh. Using potato leaf pictures, this research proposes a technique for illness identification that combines image processing and machine learning. Two potato leaf diseases, early blight and late blight, are categorized in this study using a variety of machine learning methods, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Nave Bayes, and Decision Tree. The sick leaf pictures are filtered with a Gaussian filter, and the appropriate Region of Interest (ROI) is found with the K means clustering technique. To categories the two potato leaf diseases, researchers utilized a variety of machine learning classifiers, the best of which is the support vector machine, which has a 92.9 percent accuracy.

[11] Jyoti et al looked studied a number of manual identification or infectious agent detection strategies for plant disease. The older approach discussed in the publications took more time and was more costly. The relevance of automated systems in agriculture, according to the author, is important for improving accuracy and boosting the yield of food items.

[12] In the year 2021, P. Chaitanya Reddy, M. Ayyappa Reddy, Rachakulla Mahesh Sarat Chandra, Mahesh T R, and Sindhu Madhuri G delivered a paper entitled "Detection of Plant Leaf-based Diseases Using Machine Learning Approach." The identification of leaf-based illnesses is investigated in this study utilizing Support Vector Machine (SVM) and Random Forest techniques. Root Mean Square Error (RMSE), Peak Signal Noise Ratio (PSNR), Disease Affected Area of the Leaf by Euclidian Distance technique, and Accuracy findings are compared to help farmers with less time, cheap cost, and increased agriculture output. This study used two plant leaf datasets, with dataset-1 achieving a 75 percent accuracy rate and dataset-2 achieving an 88 percent accuracy rate.

[13] Ayswarya et al suggested a technique for spraying herbicide on weeds to kill them by reducing the amount of herbicide sprayed on the plant. To reduce the grey level or conduct clustering-based thresholding on input binary pictures, the suggested system employs Otsu's approach. The suggested technology eliminates weeds more quickly than a human procedure while simultaneously increasing food output.

[14] Rima Herlina et al suggested a method that focuses on interpreting insect detection of a leaf picture in the early phases of crop damage prevention. Many approaches, such as DBSCAN and NN models, have been presented for early pest identification. The overall accuracy gained based on the pest dataset is roughly 96 percent, and it is used in research activities.

[15] Mohammed A. Hussein et al. suggested a method for detecting plant diseases that includes a knowledge base and categorization. In the proposed system, 799

example photos were utilized to develop knowledge, and an SVM classifier was employed to classify the plant disease with an accuracy of 88 percent.

[16] Vijai Singh et al suggested an automated detection and segmentation approach for leaf disease categorization. For the detection of leaf diseases, many categorization systems are examined. The suggested technique optimized the outcomes utilizing classification and identification of plant leaf diseases with very little computing effort.

[17] Prof. Sanjay B. Dhaygude et al. presented a texture-based classification strategy for detecting leaf illness using statistics. Colored co-occurrence matrix is used to segment images. To get satisfactory results, the texture-based parameters will be compared to normal leaf texture values.

[18] Kumar and Jayasankar et al. proposed a novel approach for handling the segmentation process of a system by using deformable models for detection of leaf diseases, but it uses image processing and feature-based extraction algorithms for efficiently extracting color information using the k-means algorithm.

[19] Harshadkumar B Prajapati and colleagues proposed a novel model for identifying and categorizing diseased rice in input photos. Author introduced centric feeding strategy to produce accurate results using k-means clustering as a segmentation technique. SVM is accomplished using a multiclass model and is based on categories such as shape, color, and texture.

[20] Suhaili Kutty et al. devised a method for classifying watermelon leaf diseases Anthracnose and Downey Mildew. Based on the RGB color component, this region of interest must be determined from an infected leaf sample. The median filter is then used to minimize noise and segment the data. The neural network pattern recognition toolkit is used for categorization. Based on its RGB mean color component, the proposed technique obtained 75.9% accuracy.

[21] Sanjeev Sannaki et al. want to use image processing and artificial intelligence techniques on photos of grape plant leaves to identify the condition. Downy mildew and powdery mildew of grape leaf are the principal illnesses they categories. To increase accuracy, masking is used to eliminate backdrop. Anisotropic Diffusion is utilized to preserve the information of the impacted area of the leaf. The k-means clustering algorithm is used to segment the data. The Gray Level Co-occurrence Matrix is used to extract features after segmentation. Finally, a Feed Forward Back Propagation Network classifier is used to classify the data. They just employed the Hue feature, which produces a more precise result.

[22] The support vector machine technique was utilized by Akhtar et al. for the classification and detection of rose leaf diseases such as black spot and anthracnose. The authors employed the threshold approach for segmentation, and the threshold values were determined using Ostu's algorithm. In this method, eleven haralick features based on DWT, DCT, and texture are extracted, which are then combined using an SVM technique to get a high accuracy value.

[23] Ms. Kiran R. Gavhale and colleagues demonstrated a variety of image processing algorithms for extracting damaged leaf parts. Pre-processing includes image enhancement in the DCT domain and color space conversion. After that, the k-means clustering algorithm is used to segment the data. The GLCM Matrix is used to extract features. SVM with radial basis kernel and polynomial kernel is used to classify citrus leaf canker and anthracnose disease.

[24] For the categorization of cotton leaf disease data, Bhog and Pawar used the neural network idea. K-means clustering was utilized for segmentation. Red spot, white spot, yellow spot, Alternaria, and Cercospora are some of the diseases that affect cotton leaves. The MATLAB toolbox was used to conduct the experiments. The identification accuracy of the K-Mean Clustering technique utilizing Euclidean distance is 89.56 percent, and the execution time is 436.95 seconds.

[25] The essential processes of image processing to identify and categories illness in plants were explained by Sachin Khirade and A. B. Patil. Image capture, picture preprocessing, image segmentation, feature extraction, and classification are all stages involved. Methods for segmentation include otsu's approach, transforming RGB images to HIS models, and k-means clustering. The k-means clustering approach is the most accurate of them all. Following that, features such as color, texture, morphology, and edges are extracted. Among these, the extraction of morphological features yields the best results. Following feature extraction, classification is carried out utilizing Artificial Neural Network and Back Propagation Neural Network techniques.

[26] Usama Mokhtar et al. devised a method for detecting illnesses on tomato leaves, and the diseases are Powdery mildew and Early blight. For picture improvement, several approaches such as smoothness, noise removal, image scaling, image isolation, and background removal were used. Gabor wavelet transformation is used in feature extraction and classification for feature vectors. In SVM, the Cauchy Kernel, Laplacian Kernel, and Invmult Kernel are used for output decision and illness detection training.

[27] According to Ranjan et al., diagnosis is primarily visual and involves exact judgement as well as scientific procedures. Unaffected and diseased leaves are photographed. Color HSV characteristics are retrieved as a consequence of segmentation, and an artificial neural network (ANN) is then trained to discriminate between healthy and sick samples. The accuracy of ANN categorization is 80 percent higher.

[28] Color transformation, thresholding of green pixels, segmentation, texture feature extraction using grey level co-occurrence matrix GLCM, and classification are all part of the methods presented by Kulkarni et al. in 2018.

[29] For plant leaf diseases, Kajale provides a method for detecting and computing textural information. The processing system consists of four basic steps: color picture conversion to HSI, green pixels masking and removal using a predefined threshold value, pre-processed image segmentation and extraction of important segments, and lastly texture information extraction. The texture information is used to analyses the diseases that are present on the plant leaf.

[30] According to Garg et al., support vector machine analysis is a very promising AI technology that may be used to handle a variety of classification problems. Support vector regression is a kind of SVM that is used to tackle regression issues (SVR). SVR is particularly popular among researchers since it allows the solution model to be generalized.

3.Proposed Approach

3.1 Motivation

A recognition problem deals with identity of the test database, when as compared against the educated statistics saved in the train database, with which the system has been modelled. The instinct in the back of plant leaf disease recognition is the equal. In this observation 3 special and most famous instances of plant ailment recognition which are -----------------------

1. In the related works, have not been able to identify Healthy leaves from Diseased leaves.

Figure5: Image of Healthy and Diseased leaf

2. Also, if the taken images are illuminated then the diseases were unable to identify from those images.

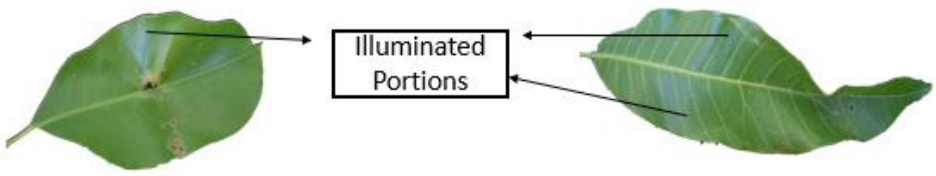


Figure6: Image of Illuminated portion

3. There are problems of having two or more one-of-a-kind diseases that could have identical chances of having similar diseases.

To lay out an effective robust system, these boundaries have to be minimized as a result the aim of this study is to cope with these challenges and discover a powerful technique to those troubles so that those technologies can be applied in real world.

3.2. System Components

In order to accomplish the task of plant leaf disease recognition, the system must have the following components.

1.Model Database
2.Feature Extractor Module
3.Classifier Module

1)Model Database

In this case, the model database is made of four sub databases. i)Train Database1. ii)Test Database1, iii) Train Database2. iv)Test Database2. In the Train set1 there are all healthy leaves samples from all classes, in the Test set1 there are all healthy & disease leaves samples from all classes, in the Train set2 there are all diseased leaves samples from all classes & In the test set2 there are all diseased leaves samples after classification between Train set1 & Test set1 from all classes. All features required have been extracted from these Train and Test leaves. so that the system learns about individual classes based on the feature extracted.
The aim of the system is to label each of these test data and classify each of them to the correct class they belong to. The plant leaves from test data also undergoes the process of feature extraction. Recognition system deals with classification, where test data is compared against train data and the required task is achieved.

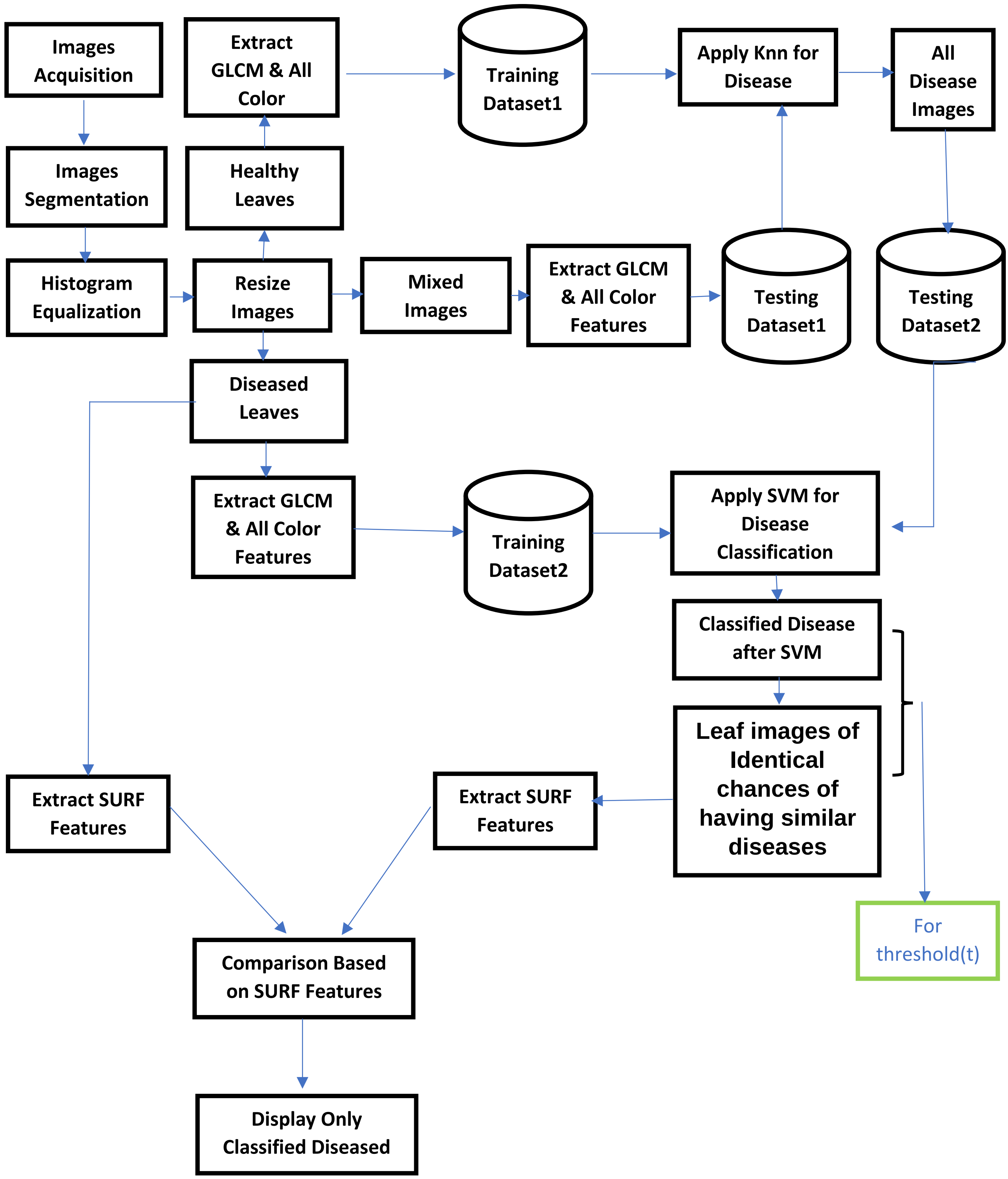
2)Feature Extractor Module.

In this phase, the raw leaf disease data collected from several sources are processed. Leaf disease can be represented with many features. But all the features are not useful for all tasks. So, for a given system in design, features must be chosen wisely. In this phase, the leaf samples from train set as well as test set undergoes the feature extraction techniques i.e, Grey Level Co-occurrence Matrix (GLCM), Mean, Standard Deviation (SD), Entropy, RMS, Variance, Kurtosis and Skewness. This gives the feature vector and with the help of these texture and color features, plant diseases are classified into different types. After 1st step of classification another feature is introduced for removing the problem of identical chances of having similar diseases that is Speeded Up Robust Feature (SURF).

3)Classifier Module

This module may be regarded as the ‘decision’ module of any recognition system. In this phase the representatives of the train set1 and test set1 are compared against each other. Each sample from the test set1 is compared against all the samples of the train set1 and with which all diseased leaves are sampled from all healthy leaves. Then from the test set2 is compared against all the samples of the train set2 and with which the maximum similarity is achieved the samples are labelled to be representative of that class. Due to the problems of having two or more one-of-a-kind diseases that could have identical chances of having similar diseases a threshold is employed to determine the samples which are misclassified and then another classification is done by the extra Robust features. Finally leaf samples are classified as their belonging classes.

3.3 Block Diagram



3.4 ALGORITHM Steps

Image acquisition is the initial stage in the suggested method. All Leaf Disease Affected and Healthy Photos are created in this method by taking images and saved in.png format.

The purpose of the preprocessing stage is to enhance picture data by segmenting background, noise, and unwanted distortions.

The **Histogram Equalization** is done for resolving the Illumination problem for the images.

The images stored in .png format are resized to standard dimensions.

After resizing images, they are separated into 3 types i.e,

a) Healthy images
b) Diseased Images
c) Mixed Images of Disease and Healthy

The **Feature Extraction** phase is the first crucial phase in the process of recognition of any object be it image, sound or video. In this phase, as explained earlier the processed raw data is transformed into some mathematical entities representing one or few features of each sample. In this case, each sample is represented by 11 features that is again represented by a vector that is feature vector.

The different color features are extracted from Mean, SD, Entropy, RMS, IDM, Smoothness, Skewness, Variance, Kurtosis. The Texture Features are extracted by using
contrast, correlation, Energy & Homogeneity.

After extracting all Color and Texture Features,

1.The Healthy Images are stored as Training Dataset1.

2.The Mixed images are stored as Testing Dataset 1.

3. All Diseased Images are stored as Training Dataset 2.

Therefore, the final feature vector is GLCM, RMS, Skewness, Mean, SD, Entropy, Variance, Kurtosis.

3.5 Classification:

In every recognition task, this is a critical step. In this stage, the features from the test and train datasets are compared to determine similarity in order to accomplish classification.

There are numerous classifiers available, but selecting the most effective one for a given problem that helps the system achieve its goal, has greater accuracy than others, and resists overfitting while providing repeatability is a hard effort involving many tests and observations.

In this work, two of the most common and widely used machine learning classifiers are employed: K-NN (number of neighbors = 1) for detection and SVM (Quadratic) for classification.

3.5.1 Selection of Parameter for K-NN Classifier:

Euclidian Distance is the most important parameter in the K-NN Classifier.

The basic goal of KNN is to locate the query point's closest neighbours. This algorithm argues that comparable items are close together; in other words, if X is positive in a collection of points, there is a good likelihood that the point closest to X is positive as well.

To calculate the distance between successive locations, we typically employ the Euclidian distance measuring approach. It's most commonly used to calculate the distance between two real-valued vectors.

When we need to determine the distance between real numbers such as integers, floats, and so on, we utilize Euclidean distance.

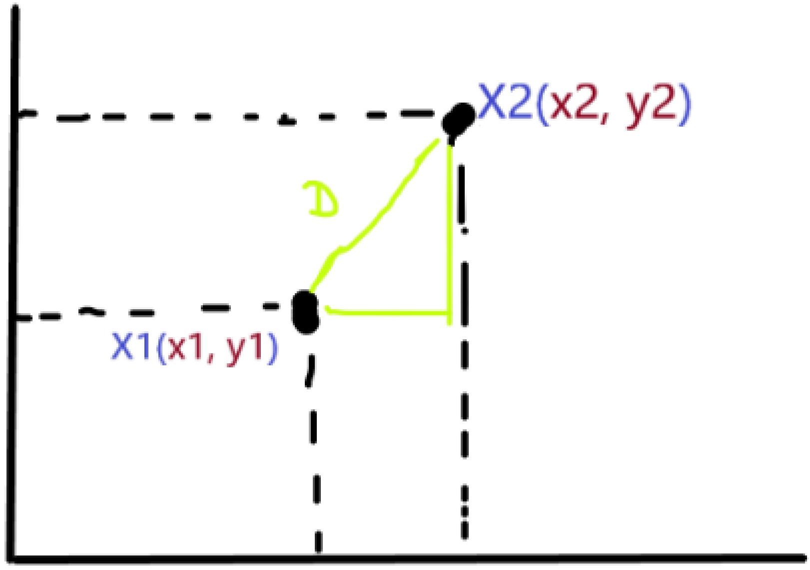


Figure7: Euclidian distance

You can see in the following graphic that there are 2-Dim data X1 and X2 stored at certain locations in 2 dimensions, assuming X1 is at (x1,y1) and X2 is at (x2,y2). We have two dimensions of data, therefore F1 and F2 are two features, and D is the shortest line between X1 and X2.

If we want to calculate the distance between these data points, we can use the Pythagoras theorem, which can be stated as: You already know that we're looking for the distance between any two points, so we can write it as: **d = || X1 – X2 ||**

$$d = \sqrt{\sum_{i=1}^{N}} |\mathbf{Xi} - \boldsymbol{Yi}|^2 \quad \dots\dots\dots\dots\dots\dots\dots\mathbf{14}$$

So, for the detection component, we utilized the KNN Euclidian distance parameter with a value of 1.

After K-NN Classification, all diseased and healthy leaves are appropriately categorized.

The Q-SVM classifier is then employed for better categorization of those diseased leaves. This Classifier will divide the leaves into five categories.

Parameters for SVM classification must now be selected.

3.5.2 Selection of Parameters for SVM Classifier:

One of the most well-known learning methods for classification and regression issues is the support vector machine (SVM). The complexity and performance of prediction models are heavily influenced by SVM parameters such as kernel parameters and Gama parameters.

As a result, the penalty parameter and kernel parameters are used in SVM model selection. However, without understanding the fundamental nuances, these settings are frequently chosen and exploited like a black box.

The following are the many reasons for employing the Kernel parameter in an SVM classifier for classification.

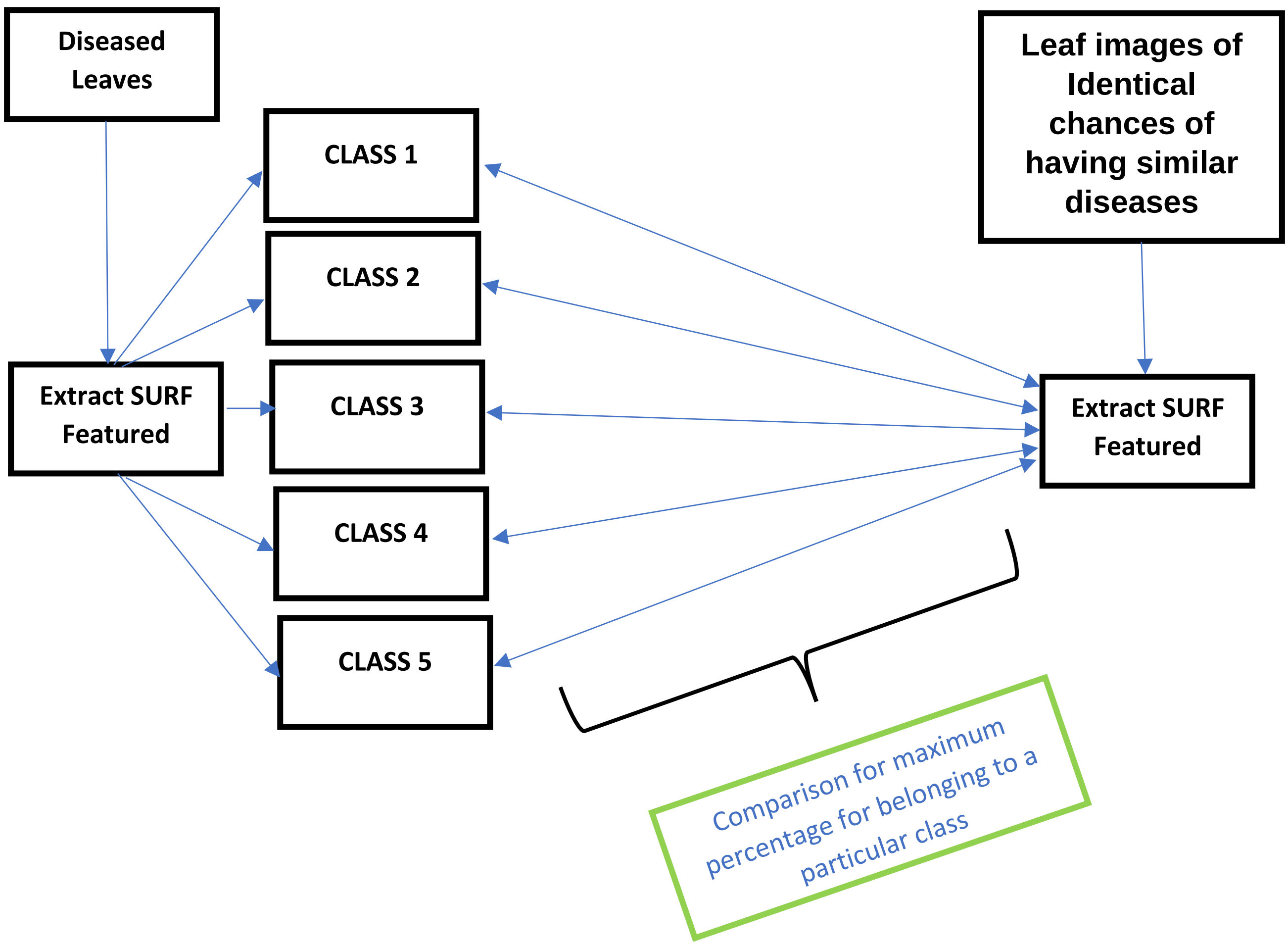
(1) Under the structural risk minimization concept, it has a high degree of generality.
(2) Kernel techniques can be used to solve non-linear issues.
(3) After inserting slack variables, it is resilient to noisy occurrences.
(4) It generates sparse solutions since the best hyper-plane is determined only by support vectors.
(5) Convergence is ensured.

After SVM classification there is a problem of two or more one-of-a-kind diseases that may have equivalent possibilities of having comparable diseases.

Due to the issue of two or more one-of-a-kind diseases that may have equivalent possibilities of having comparable diseases, another feature extraction method called Speeded Up Robust Features (SURF) of Diseased Images has been developed.

Also, after using the threshold (>=839), discover the SURF Features of the Leaf pictures of Identical odds of having identical illnesses.

3.5.3 Comparison Based on SURF Features:



Extracted SURF Features of Diseased Leaves are classified into five classes. Now, the retrieved SURF Features of the leaf pictures with equivalent probability of having similar diseases and those five Classes are compared.

In the comparison the extracted SURF features will give the maximum accuracy that is tens to 99.99% or above for belonging to a particular class.

Then the maximum accuracy containing Class will be Class the infected leaf.

The result of the Comparison of SURF characteristics then only shows the classified sick Images.

3.6 Chapter Summary:

In summary, the first stage in the suggested technique for Plant leaf Disease Identification is Image Acquisition. All Leaf Disease Affected & Healthy Photos are created in this method by taking images and saved in.png format.

The purpose of the preprocessing stage is to enhance picture data by segmenting background, noise, and reducing unwanted distortions.

Histogram Equalization is used to solve the image's Illumination issue.

Images saved in.png format are scaled to normal sizes. After resizing, photos are classified into three types: healthy images, diseased images, and mixed disease and healthy images.

The following methods are used to extract features from all Healthy pictures, Disease images, and Mixed images. Texture is one of the most essential characteristics that may be used to categories and identify items. Grey Level Co-occurrence Matrices (GLCM) are an ancient and widely used texture categorization feature extraction approach. Textural properties such as correlation, energy, entropy, and homogeneity may be computed from the produced GLCMs. Calculate the Color Feature values of Skewness, Standard Deviation, Homogeneity, Contrast, Correlation, Kurtosis, Energy, Entropy, Mean, Variance, and RMS to extract illness and non-disease symptoms.

The Healthy Images are saved as Training Dataset 1 for the detection phase, while the Mixed Images are saved as Testing Dataset 1. The Euclidean Distance between the features of each Training Dataset 1 and each Testing Dataset 1 is determined in the first classification. The smallest Euclidean Distance among all calculated Euclidean Distances is used to classify the testing picture. As a result, following the first classification, all diseased photos are saved as Testing Dataset 2.

All Diseased Images are saved as Training Dataset 2 for the Classification Part. The SVM is used for classification, regression, or other tasks in the second classification. The SVM classification is performed between the training and testing datasets. The training feature set is used to train the SVM model, while the testing feature set is used to validate the trained SVM model's accuracy. The output of the SVM model then only shows the classified sick images.

Another feature extraction, SURF features, is extracted for the Leaf pictures of identical risks of suffering comparable illnesses. Extracted SURF Features of Diseased Leaves are classified into five classes. Now, the retrieved SURF Features of the leaf pictures with equivalent probability of having similar illnesses and those five Classes are compared. The result of the Comparison of SURF characteristics then only shows the classified sick Images.

4.EXPERIMENTS AND OUTCOMES

This section is an important element of the study since it contains all of the findings and charts that lead us to our conclusion. This section aided us in selecting an acceptable feature and classifier for the system.

The majority of the experiments are carried out in the MATLAB environment. The versions used are MATLAB R2016a.

4.1. Experimentations of Detection part of diseased leaves:

At first the detection of Healthy leaves from the Mixed leaf Images are done then only the diseased images are Classified into Different classes.

4.1.1 Dataset for Detection part:

All Leaf Disease Affected & Healthy Image are generated by capturing images and stored in .jpg format. The most early and one of the vital steps of designing a system is the collection of relevant, solid, reliable dataset required for the process.

In this case, there are **five classes** that is used as dataset for Healthy leaves and Mixed leaves are ---------

a) **mango,**

b) **Arjun**

c) **Alstonia Scholaris**

d) **Guava**

e) **jamun**

The total No. of Dataset are **2348 of all leaves** that is Healthy, Diseased & Mixed leaves.

There are total of **1080 Healthy leaves** of such classes and **108 Mixed leaves** that is mixture of all Healthy and Disease leaves are present for the Detection part.

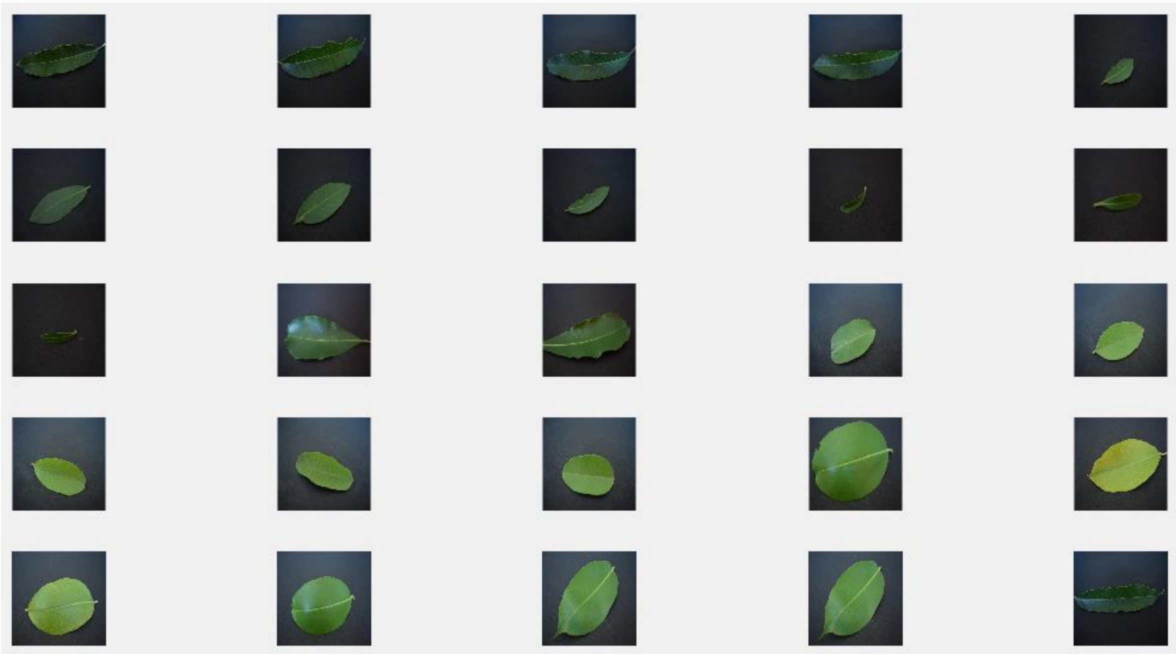


Figure8: Sample Images of Training Dataset1

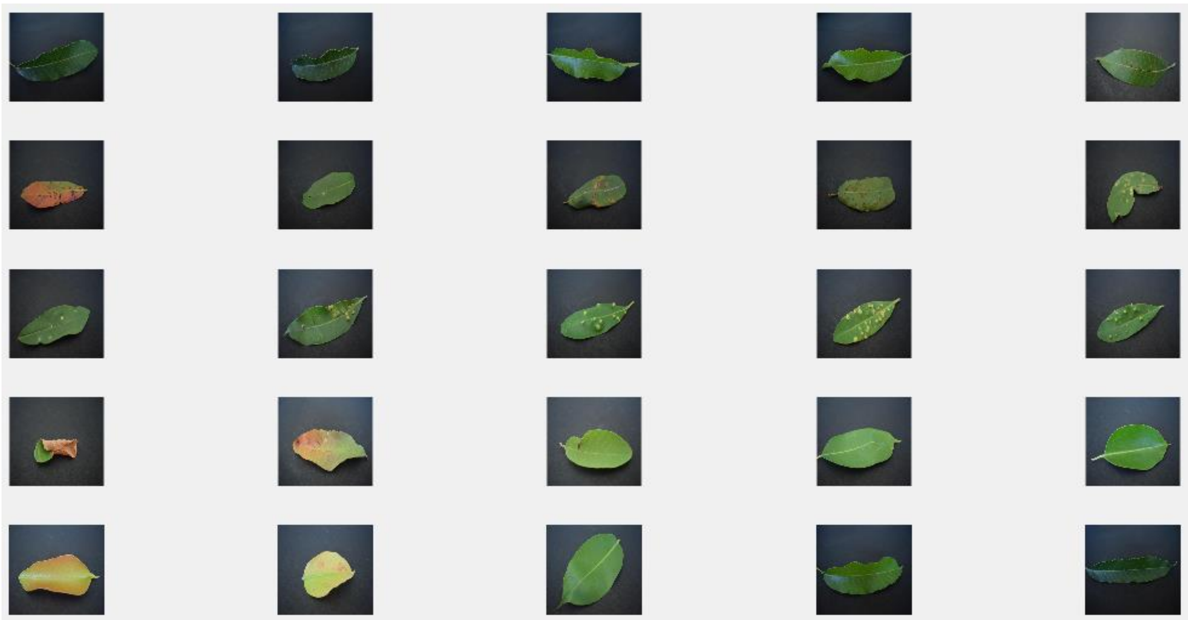


Figure9: Sample Images of Testing Dataset1

Here, are the images of 5 types of Diseased leaves and their names.

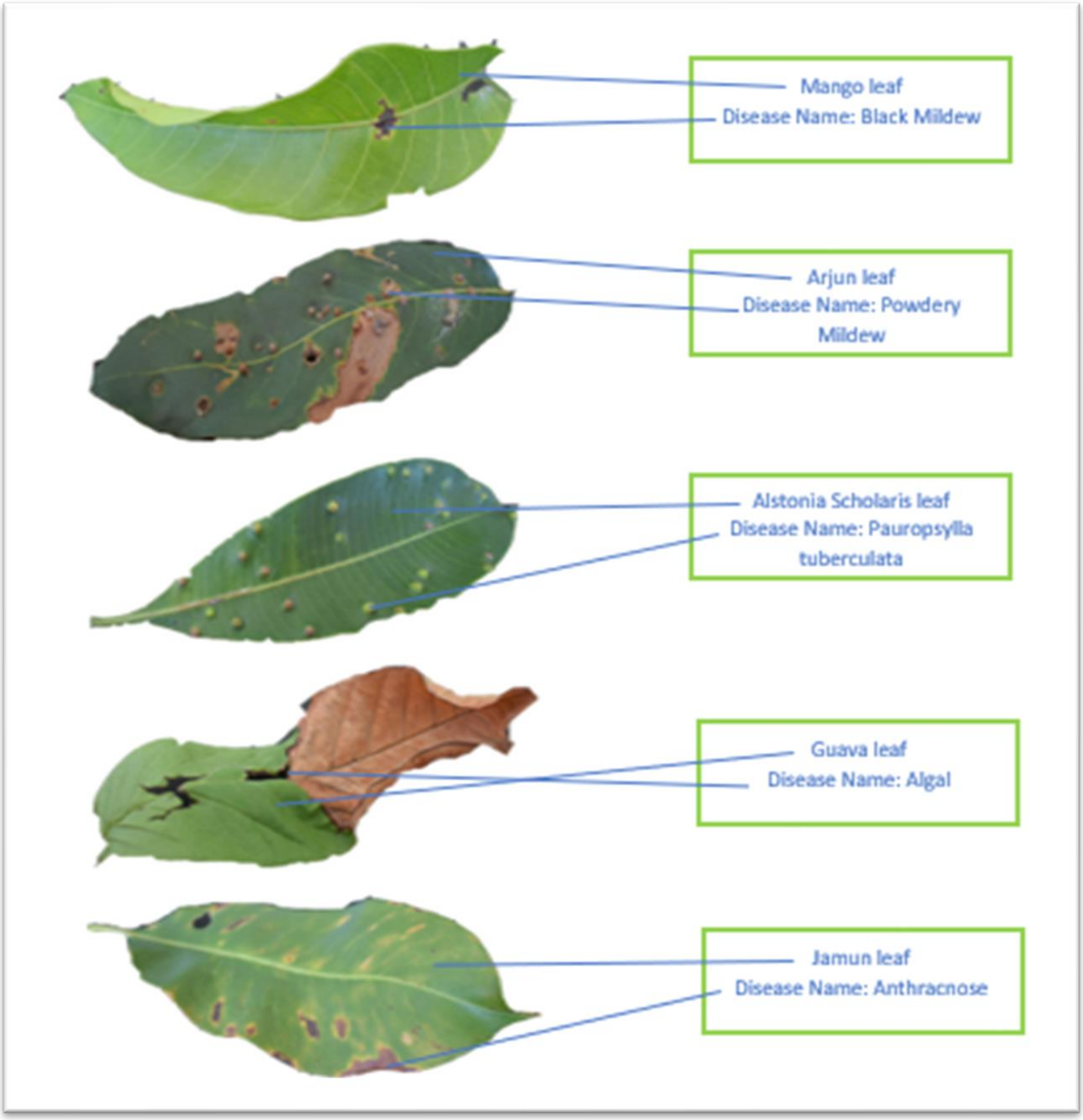


Figure10: Leaf Diseases and Their Names

For Detection part, Only Knn classifier is used to separate the Healthy Leaves from the mixed leaf images to the only Diseased leaves.

Then Diseased Leaves are used for another classification in the next part.

4.1.2 <u>Experimentations</u>

The experiment was carried out with the training dataset -1(1080) and the testing dataset-1(108) based on the characteristics, namely texture and all color features.

This section detects sick leaves among the mixed leaves and stores them in testing dataset-2.

Only the K-nearest neighbour Classifier is employed in this case to classify diseased leaves.

TABLE 1: Tabulation of Results of First Experiment

Total Dataset	Training Dataset-1	Testing Dataset-1	Feature Extraction	Classification	Accuracy
2348	1080	108	GLCM, Mean, SD, Entropy, RMS, IDM, Smoothness, Skewness, Variance, Kurtosis	KNN Classifier	100%

The table portrays how many test samples are classified correctly amongst the 108 test samples used in this experiment.

Here, training Dataset 1 contains 1080 Healthy leaves and Testing Dataset 1 contains 108 Mixed leaves.

After Classification based on the features the gained accuracy is 100%.

That means all the Diseased images from the mixed images are Classified Correctly.

Here, the Feature plot of Testing Dataset 1 are shown below.

As there is only Healthy Leaves in training dataset 1, so it is the feature plot of all healthy leaves.

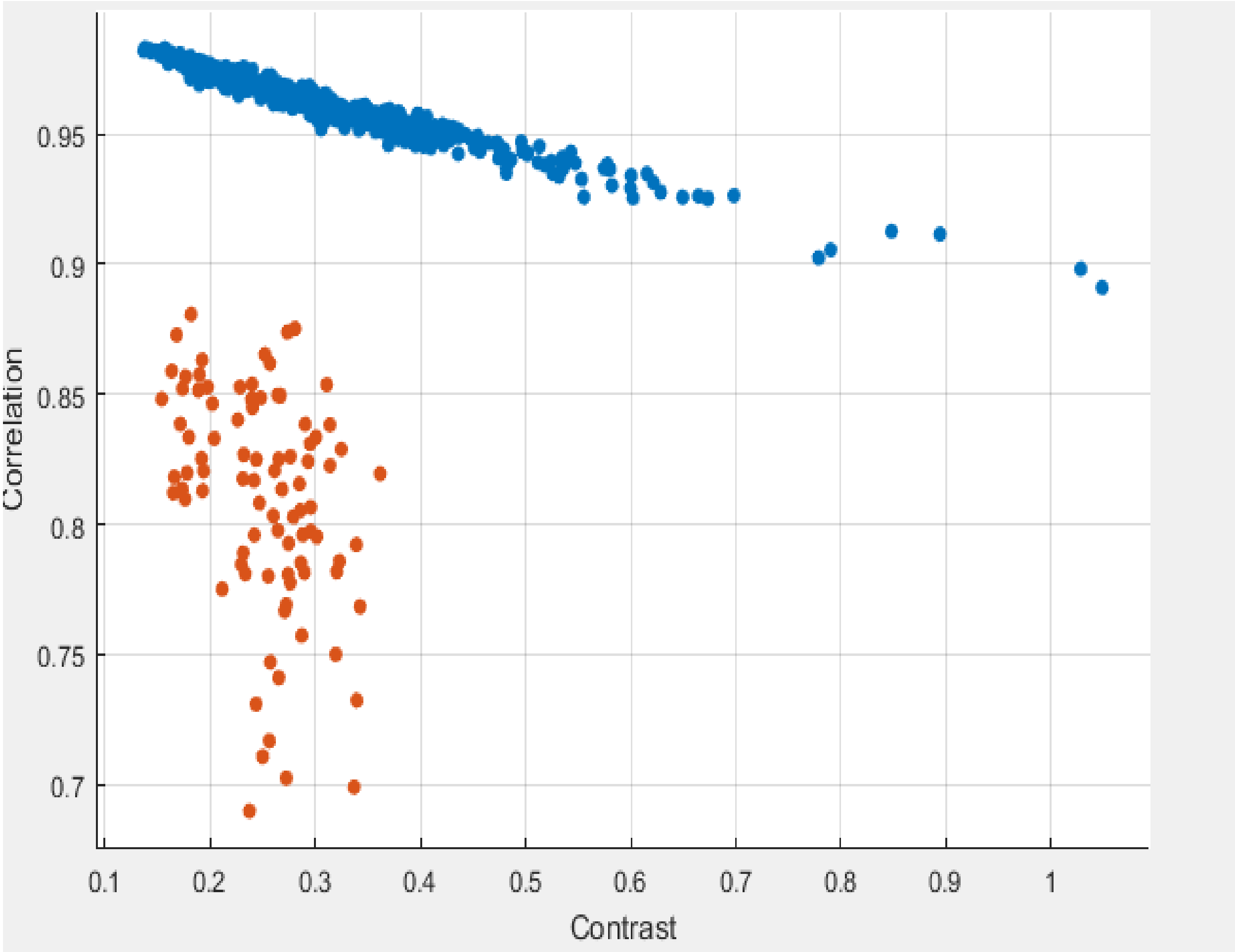


Figure11: feature plot of testing dataset 1

Here, it is the feature plot of testing dataset 1 i.e, Mixed images of healthy and disease. the Red ones are the Diseased leaves feature and the Blue ones are the Healthy leaves feature.

Now, after feature extraction of all color features that is Mean, SD, Entropy, RMS, IDM, Smoothness, Skewness, Variance, Kurtosis and all texture features of GLCM, the tasting dataset 1 are classified using the K-nearest neighbors Classifier.

Therefore, all diseased leaves from the mixed images are stored in Testing Dataset2.

Here is the confusion matrix for the Classification.

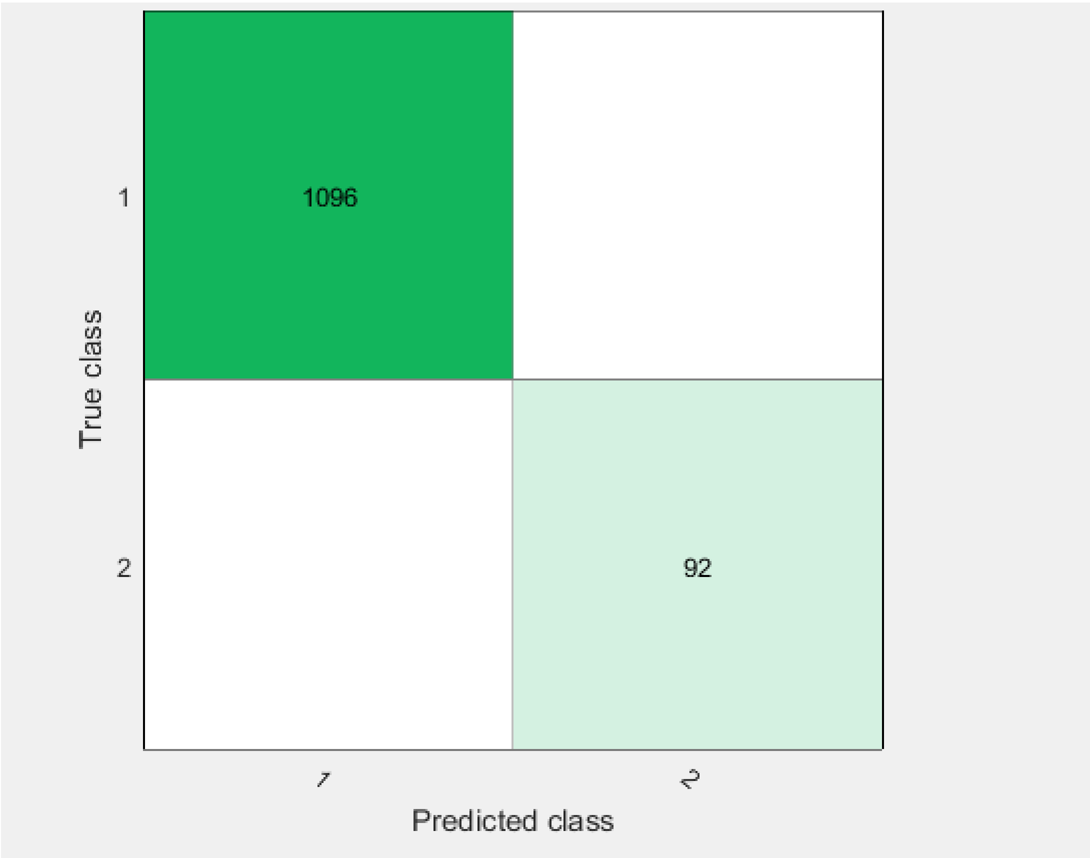


Figure12: Confusion matrix after classification

So, after classification it is clear that all diseased leaves are now in testing dataset 2. So the accuracy is 100%.

4.1.3 Dataset for Classification part:

All Leaf Disease Affected & Healthy Image are generated by capturing images and stored in .jpg format. The most early and one of the vital steps of designing a system is the collection of relevant, solid, reliable dataset required for the process.

In this case, there are **five classes** that is used as dataset for all Diseased leaves and Mixed leaves are ---------

a) **mango,**
b) **Arjun**
c) **Alstonia Scholaris**
d) **Guava**
e) **jamun**

The total No. of Dataset are **2348 of all leaves** that is Healthy, Diseased & Mixed leaves.

There are total of **1160 Diseased leaves** of such classes and other is detected by detection part that is **92 diseased leaves in Testing dataset 2** for the Classified part.

From 1160 Diseased Leaves 252 are Mango, 221 are Arjun, 222 are Alstonia Scholaris, 129 are Guava, 336 are jamun.

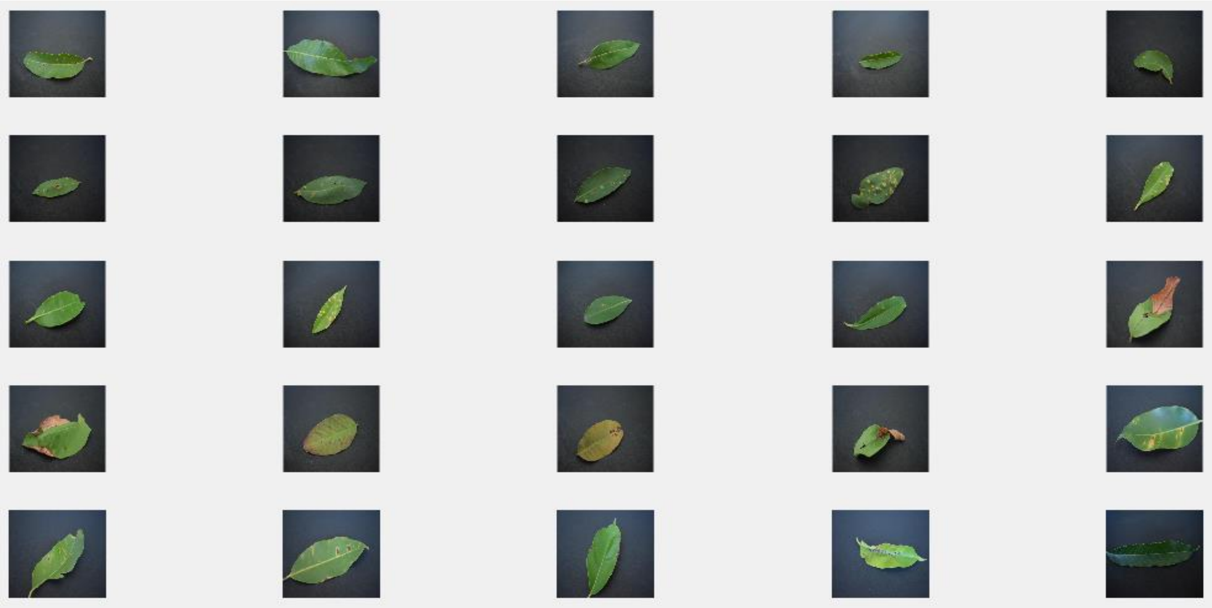


Figure13: Sample Images of Testing Dataset2

Figure14: Sample Images of Testing Dataset2

For Classification part, Only QSVM classifier is used to separate the classes of Testing dataset 2.

4.1.4 Experimentations

The experiment was conducted using the training dataset -2(1160) and the testing dataset-2(92) based on the characteristics, namely texture and all color features.

After QSVM, unhealthy leaves are discovered from the mixed leaves and saved in Classified Disease.

Here Only Quadratic Support Vector Machine Classifier is used for classification of five different classes that is Mango, Arjun, Alstonia Scholaris, Guava, Jamun.

Total Dataset	Training Dataset-2	Testing Dataset-2	Feature Extraction	Classification	Accuracy
2348	1160	Only Diseased Leaves	GLCM, Mean, SD, Entropy, RMS, IDM, Smoothness, Skewness, Variance, Kurtosis	SVM Classifier	89.60%

TABLE 2: Tabulation of Results of Second Experiment

The table portrays how many test samples are classified correctly amongst the testing dataset 2 samples used in this experiment.

Here, training Dataset 1 contains 1160 Diseased leaves and Testing Dataset 2 contains 92 Diseased leaves which is classified.

In this Classification the features are same as the 1st experiment using Knn.

The features are GLCM, Mean, SD, Entropy, RMS, IDM, Smoothness, Skewness, Variance, Kurtosis.

After Classification based on the features the gained accuracy is 89.60%.

That means all the Diseased images from the Training Dataset 2 images are Classified most accurately.

Now all feature plots as well as classification plots will show for better identification of classes.

Here, the Feature plots of Training Dataset 2 and Testing Dataset 2 are shown below.

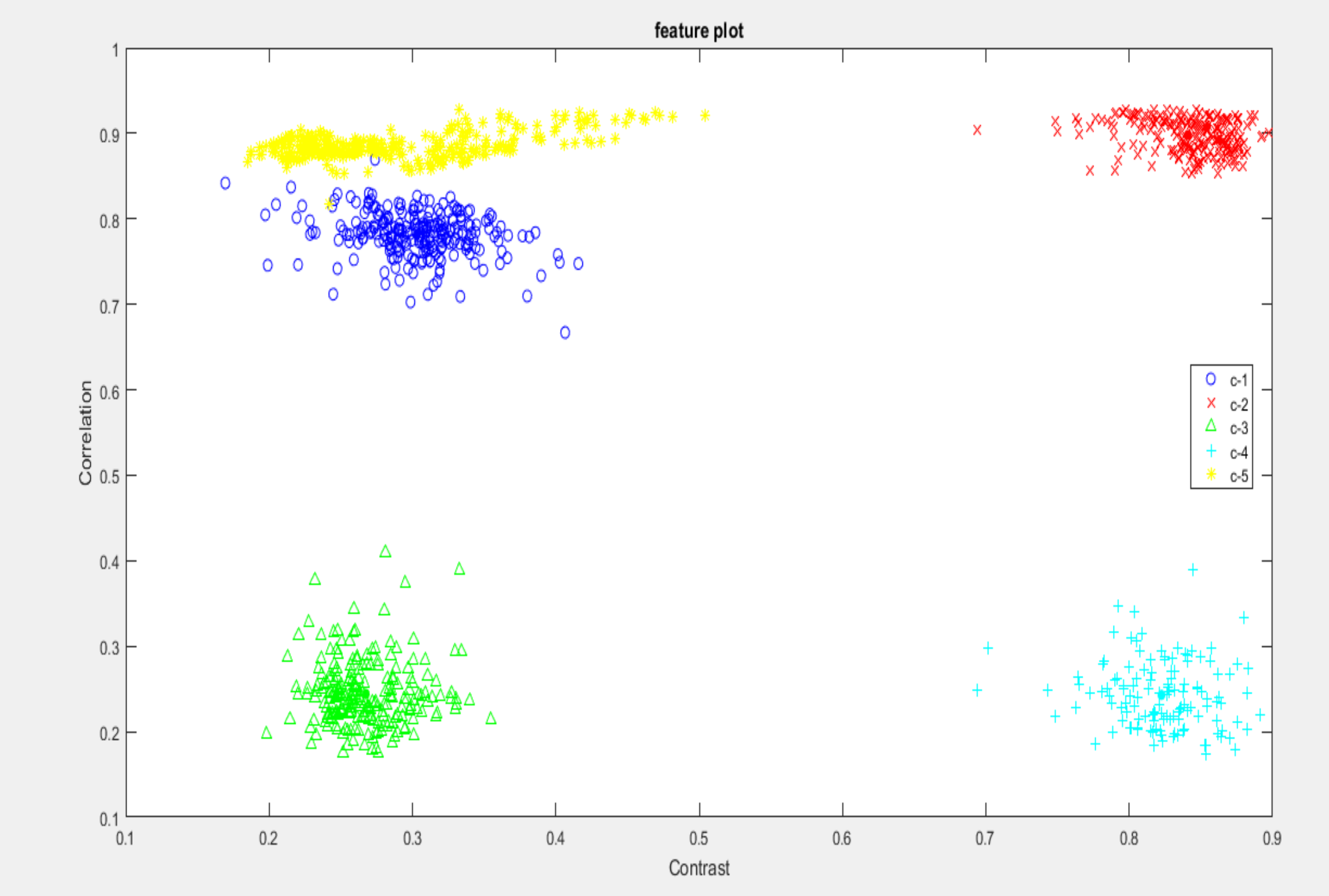


Figure15: feature plot of training dataset2 i.e, only Diseased leaves

As there is only Diseased Leaves in training dataset 2, so it is the feature plot of all Diseased leaves.

Here Blue represent the Class 1 that is Mango, Red represents the Class 2 that is Arjun, Green represents the Class 3 that is Alstonia Scholaris, Light Blue represents the Class 4 that is Guava & Yellow represents the Class 5 that is Jamun.

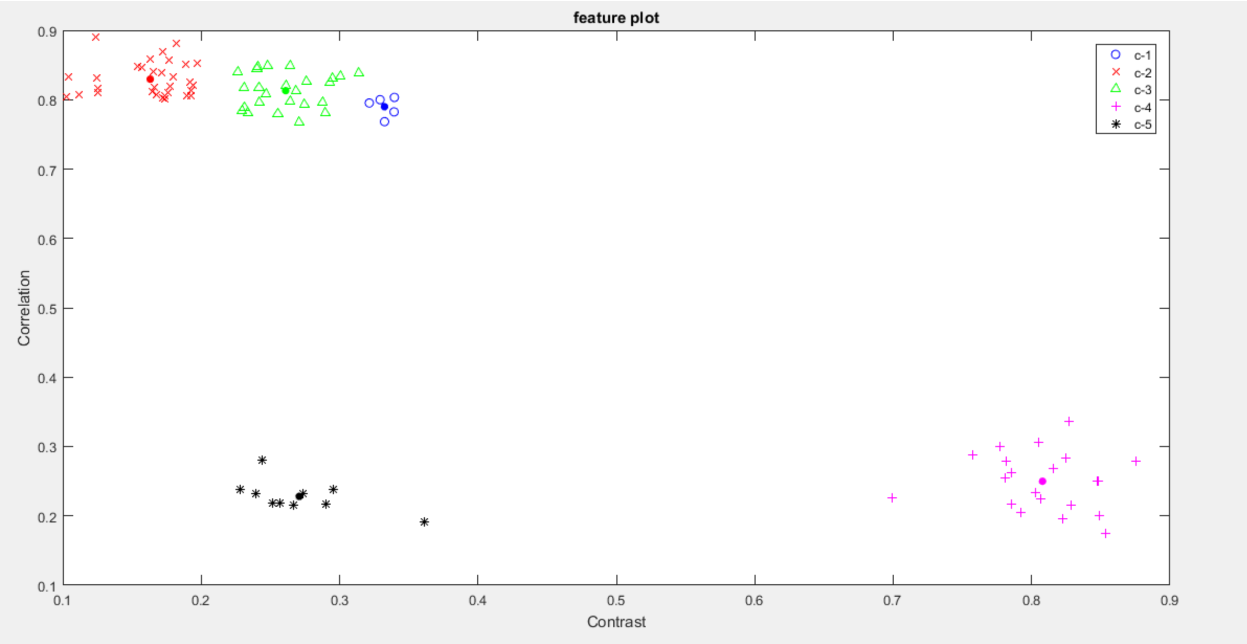


Figure16: feature plot of testing dataset 2 i.e, only Diseased leaves

As there is only Diseased Leaves in testing dataset 2, so it is the feature plot of all Diseased leaves.

Here Blue represent the Class 1 that is Mango, Red represents the Class 2 that is Arjun, Green represents the Class 3 that is Alstonia Scholaris, Light Blue represents the Class 4 that is Guava & Yellow represents the Class 5 that is Jamun.

Now, after feature extraction of all color features that is Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis and all texture features of GLCM, the tasting dataset 1 are classified using the K-nearest neighbors Classifier.

Therefore, all diseased leaves are Classified to their corresponding classes and stored in Classified disease.

Here is the confusion matrix for the Classification.

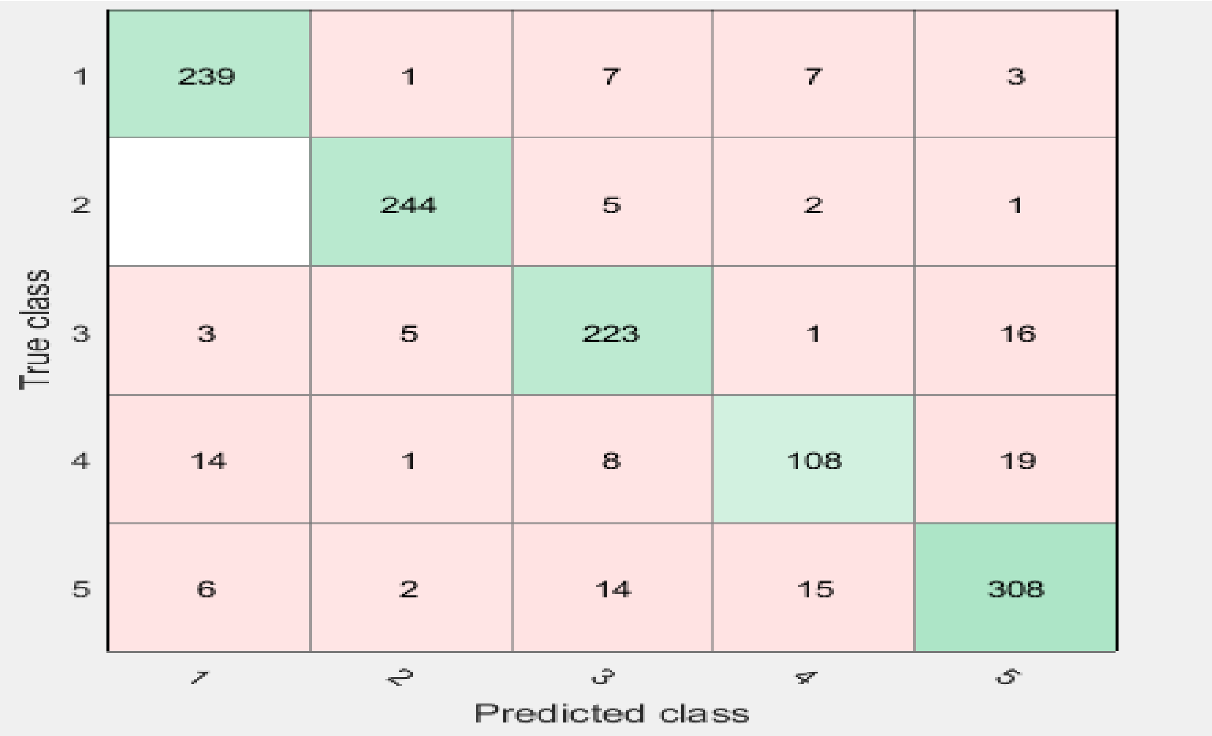


Figure17: Confusion matrix after classification

Now there is the true positive rate and false negative rate of classification which is shown below.

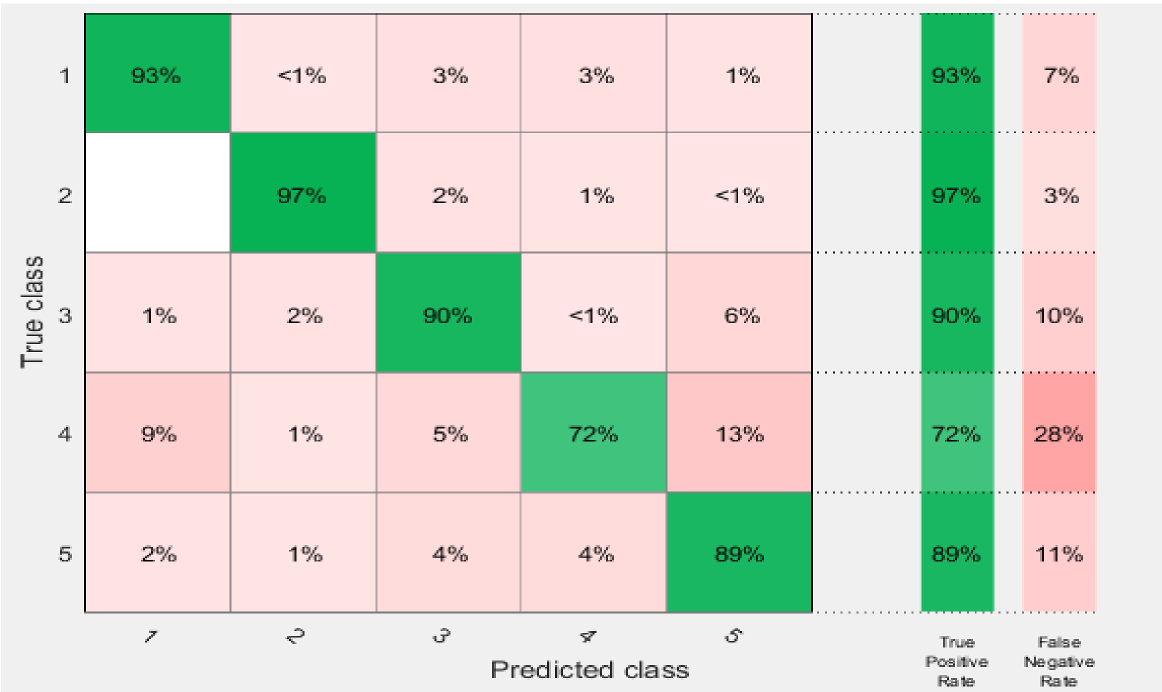


Figure18: Confusion matrix of true positive and false negative after Classification

So, it is clear that all the leaves are not classified accurately.

The 1^{st} accuracy that is 93% is shows that the accuracy to predict class 1,the 2^{nd} one shows the accuracy of 2^{nd} class that is 97%, the 3^{rd} one shows the accuracy of 90% which is belong to class 3, the 4^{th} one shows 72% accuracy for having class 4 & the last one shows 89% accuracy for class 5. That's why the overall accuracy is 89.60%.

So, there is problem of having two or more one-of-a-kind diseases that could have identical chances of having similar diseases which is also impacted in the overall accuracy.

To avoid the problems an extra inspection step is added to the algorithm at the final stage. Here, at the final stage the extra inspecting step checks for a newly added feature i.e; Speeded Up Robust Feature (SURF) which is calculated and Classified Perfectly.

4.1.5 Dataset for Final Comparison part:

Now to select the images which are having the problem a threshold of >=839 is selected. So, here are the leaf images which are having these problems.

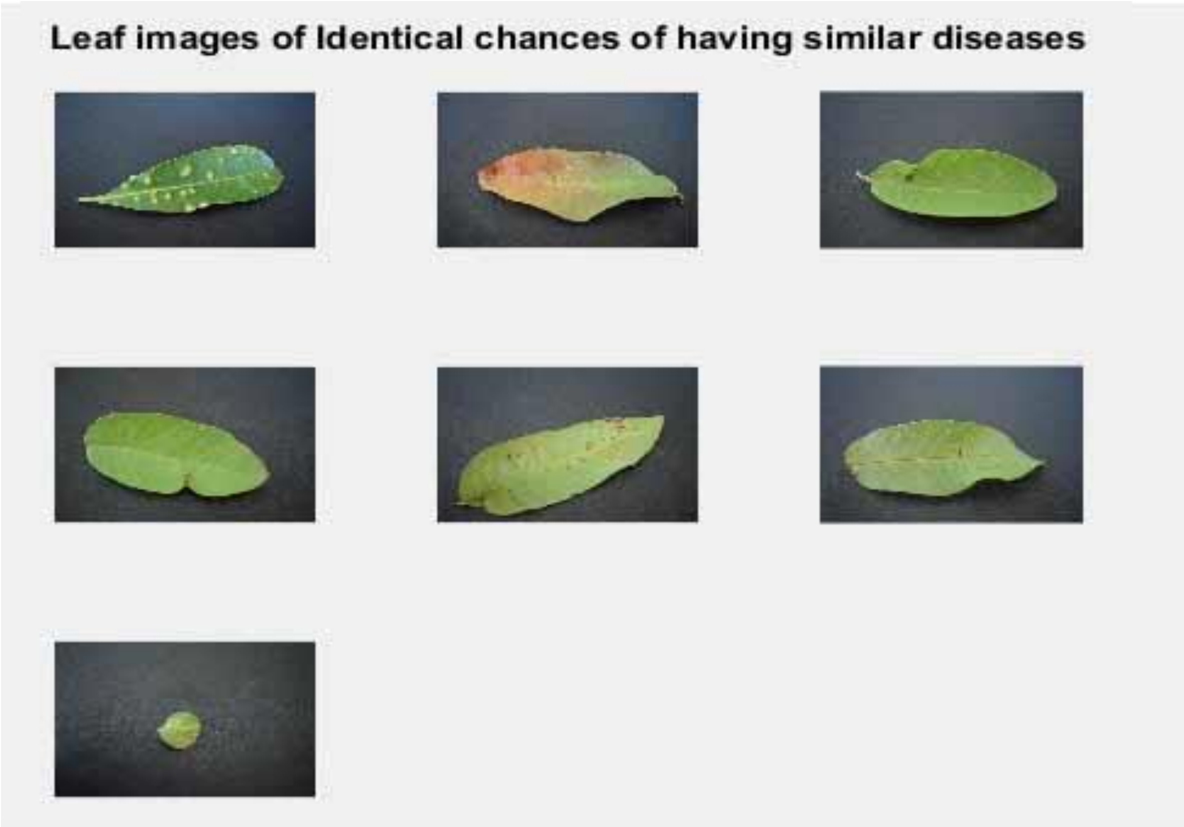


Figure19: Leaf Images of Identical Chances of having similar diseases

4.1.6 Experimentation:

TABLE 3: Tabulation of Results of Final Experiment

Total Dataset	Training Dataset-2 is dividing upon 5 classes					Identical Chances For Similar Diseases leaves	Feature Extraction	Comparison	Accuracy
	1	2	3	4	5				
2348	251	221	222	129	336	Those 7 leaves	Speeded Up Robust Features (SURF)	Based on Surf feature. Finding maximum accuracy to having one particular class.	98.91%

So, training Dataset 2 that is All diseased images are separated into 5 classes that is Mango, Arjun, Alstonia Scholaris, Guava, Jamun.

Then finding the all SURF features of these 5 classes and the diseased leaves which have an identical chance for similar diseases.

After that a comparison is done between those 5 classes and the leaves of having an identical chance for similar diseases.

And here is the Confusion matrix after comparing the similarities via SURF features.

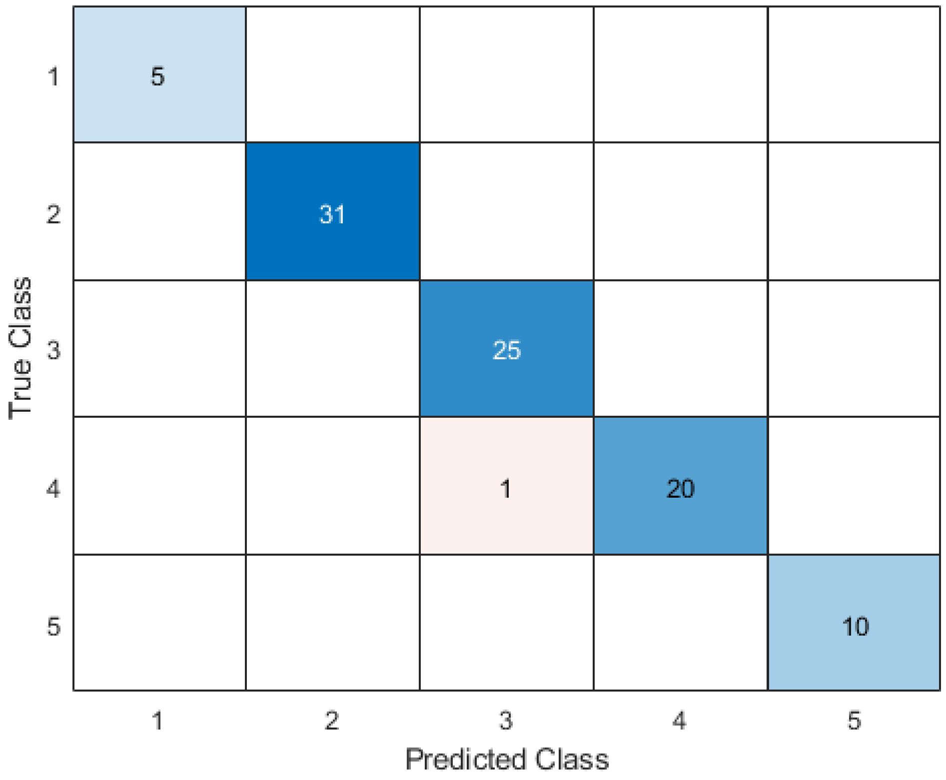


Figure20: Confusion matrix after comparison of SURF features

So, it is easily showing that accept one leaf image every image is correctly Classified to their own classes. That's why the Accuracy is now 98.91%.

The image is not Classified in any group is shown below.

Figure21: Not identified the class

4.1.7 Chapter Summary:

In summary of Plant leaf Disease Identification, the 1st step of experimentation is going on by the detection part. In this part the Diseases leaves are detected from the Healthy leaves by the accuracy of 100% using Knn classifier.

Then in the next step the detected Diseased leaves are classified into their several classes. In this Classification part Diseased leaves are classified by the help of Quadratic Support Vector Machine Classifier. Here, Classification is done between training dataset 2 that is all diseased leaves and the testing dataset 2 which is the result after Knn classifier.
After that there is problem of having two or more one-of-a-kind diseases that could have identical chances of having similar diseases which is also impacted in the overall accuracy. So, to solve this problem all diseased leaves are separated into 5 different classes and then Finding the Speeded Up Robust Features for all 5 classes.

Then the comparison is done between those 5 classes and the leaves of having an identical chance for similar diseases. Then the overall accuracy is stand to 98.91%.

5 Comparative Analysis

Comparative analysis refers to the comparison of two or more processes, documents, data sets or other objects. Pattern analysis, filtering and decision-tree analytics are forms of comparative analysis. It is the process of comparing items to one another and distinguishing their similarities and differences.

Here, comparative analysis refers to the comparison of features, Classifiers & accuracy to the previous work to the practiced work.

5.1 Comparative Analysis 1

Comparison of practice work and literature surveys.

TABLE 4: Tabulation of Results of Comparative analysis 1

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[1]	200	Tomato & Corn Disease(2-types)	HOG & SVM	65%-70%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	HOG & SVM	49.70%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis, QSVM & SURF.	98.91%

- **Analysis**

In Previous Approach of Literature Survey 1, the technique which is used as feature extraction is Histogram oriented gradients (HOG). The Main problem with using HOG is it is very sensitive to image rotation.

There are many images which are captured in different angles.

Therefore, HOG is not good choice for classification of textures or objects which can often be detected as rotated image.

That’s why in the proposed algorithm there is no Histogram oriented Gradient feature (HOG).

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

As classifier in the previous approach they used support vector machine to classify.

However, in the suggested technique, K-nearest is utilized to distinguish diseased leaves from healthy leaves, and quadratic Support vector machine is employed for classification because QSVM is capable of separating data non-linearly, a dual optimization form that is a quadratic decision function is used.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

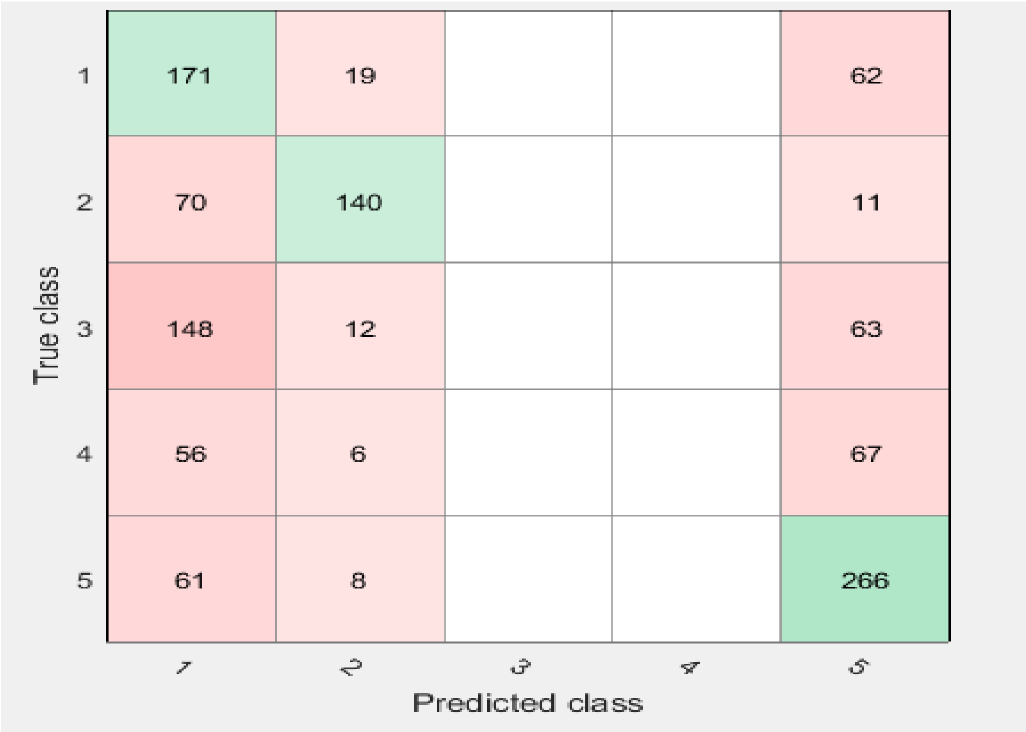


Figure22: Confusion matrix of Previous Approach

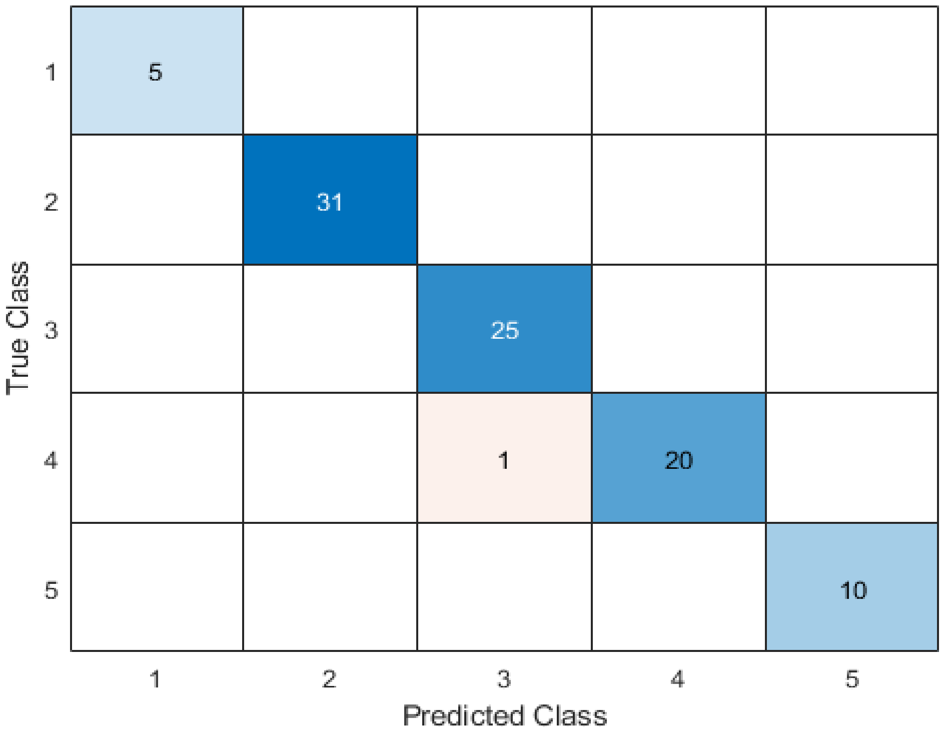


Figure23: Confusion matrix of Proposed Approach Using SURF comparison

5.2 Comparative Analysis 2

Comparison of practice work and literature surveys

TABLE 5: Tabulation of Results of Comparative analysis 2

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[2]	560	Bacterial Blight, Cercospora Leaf Spot, Powdery Mildew, and Rust are all typrs of plant diseases. (4-types)	GLCM, LBP & KNN	80.2%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	GLCM, LBP & KNN	54.5%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis, QSVM & SURF.	98.91%

- **Analysis**

In Previous Approach of Literature Survey 2, the technique which is used as feature extraction is Local Binary Pattern (LBP) and Gray Level Co-Occurrence Matrix (GLCM)

The Main problem with using LBP is they produce rather long histograms, which slow down the recognition speed especially on large-scale face database and Under some certain circumstance, they miss the local structure as they don't consider the effect of the center pixel. LBP is not good choice for classification of textures or objects.

That’s why in the proposed algorithm there is no Local Binary Pattern feature (LBP).

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

As classifier in the previous approach they used K-Nearest neighbor to classify. But the main problem with KNN doesn't work well with a large dataset and doesn't work well with a high number of dimensions.

But in the proposed approach there is K-nearest for extracting the Diseased leaves from healthy leaves and the used quadratic Support vector machine for classification because QSVM is a dual optimization form that is a quadratic decision function capable of segregating data non-linearly.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

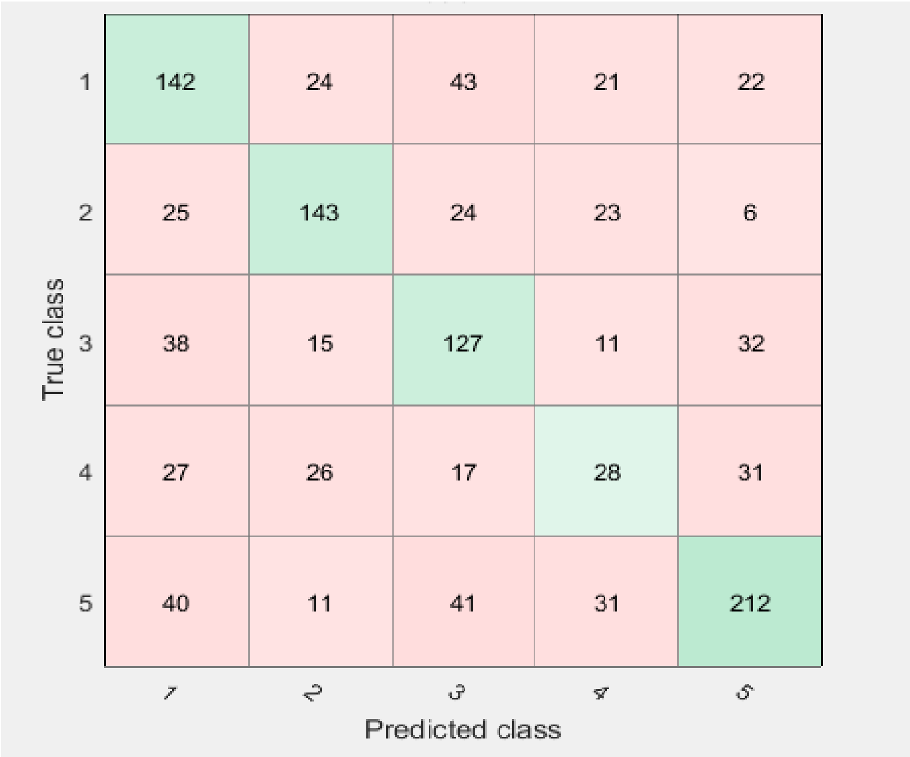


Figure24: Confusion matrix of Previous Approach

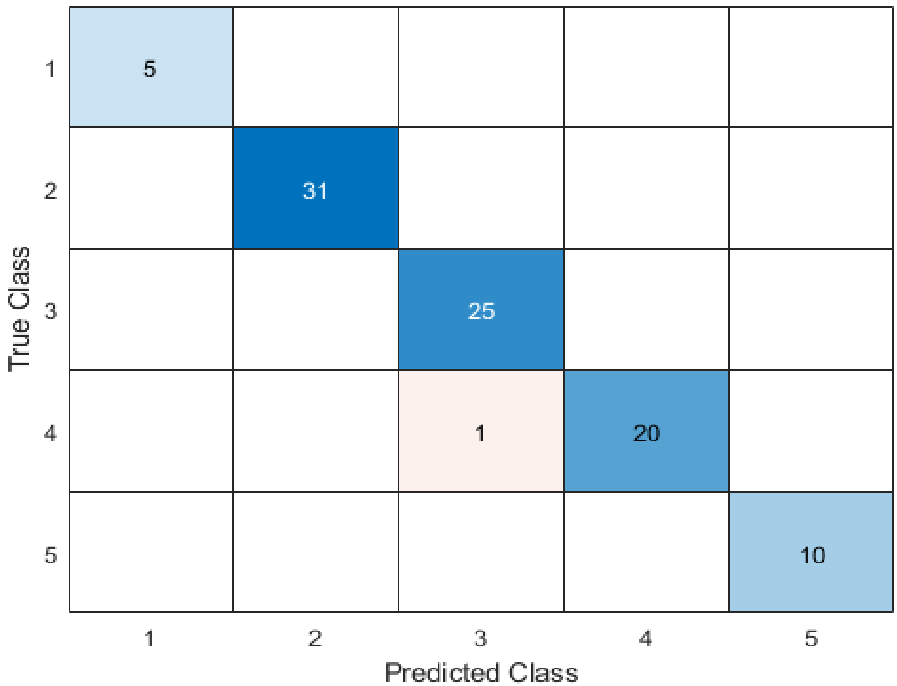


Figure25: Confusion matrix of Proposed Approach Using SURF comparison

5.3 Comparative Analysis 3

Comparison of practice work and literature surveys

TABLE 6: Tabulation of Results of Comparative analysis 3

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[3]	28	Dag, Golmachi, ShutiMold, and Red Moricha Disease (4-types)	GLCM & SVM	80%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	GLCM & SVM	60.1%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis, QSVM & SURF.	98.91%

• **Analysis**

In Previous Approach of Literature Survey 2, the technique which is used as feature extraction is Gray Level Co-Occurrence Matrix (GLCM)

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

As classifier in the previous approach they used Linear Support Vector Machine to classify. But the main problem with Linear Support Vector Machine ---

1. Linear SVM performs poorly when the data set has more noise, i.e. target classes overlap.

2. The Linear SVM will underperform when the number of features for each data point exceeds the number of training data samples.

3. There is no probabilistic justification for the classification because the support vector classifier operates by placing data points above and below the classifying hyperplane.

But in the proposed approach there is K-nearest for extracting the Diseased leaves from healthy leaves and the used quadratic Support vector machine for classification because QSVM is a dual optimization form that is a quadratic decision function that may separate data non-linearly.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

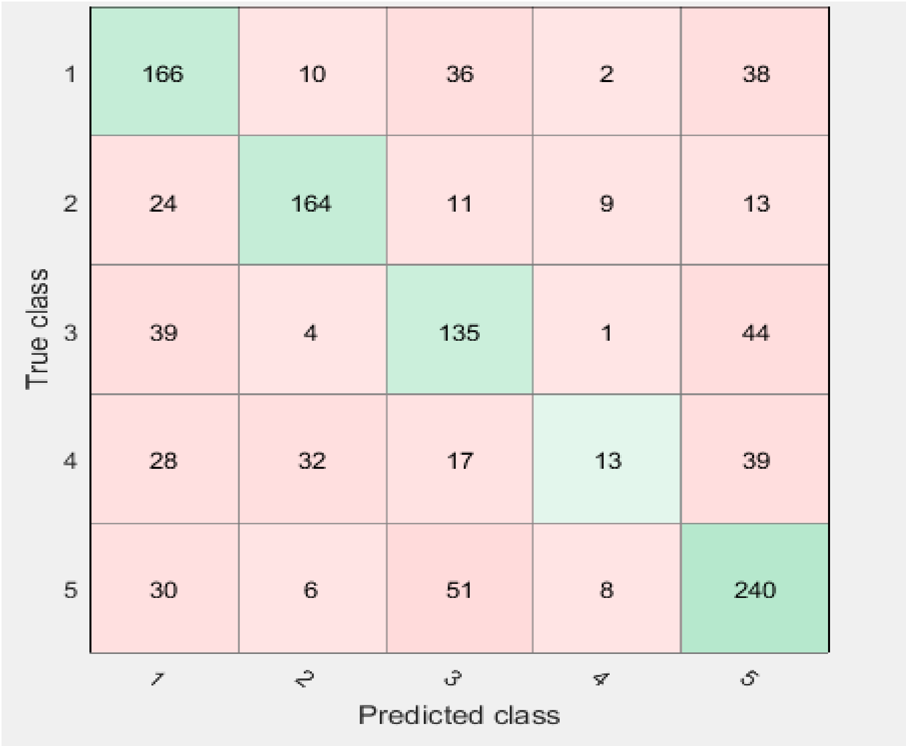


Figure26: Confusion matrix of Previous Approach

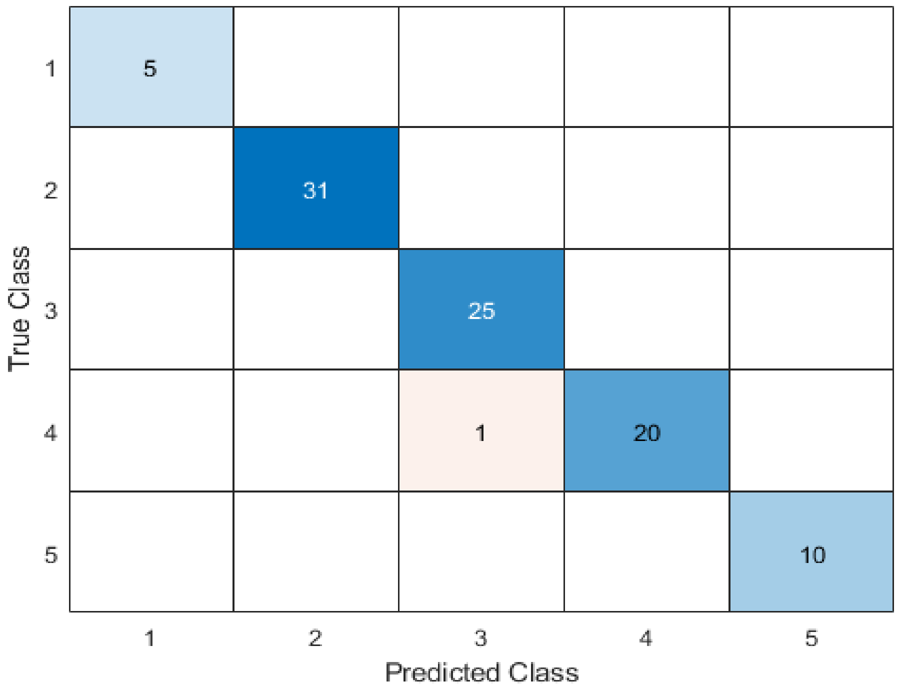


Figure27: Confusion matrix of Proposed Approach Using SURF comparison

5.4 Comparative Analysis 4

Comparison of practice work and literature surveys

TABLE 7: Tabulation of Results of Comparative analysis 4

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[4]	800	Alternaria Alternata, Anthracnose, Bacterial Blight, Disease (3-types)	GLCM, Mean, RMS, Variance, Smoothness, Standard Deviation, Entropy, Kurtosis, and Skewness, IDM & SVM	86%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	GLCM, Variance, Smoothness, Mean, Standard Deviation, Entropy, RMS, Kurtosis, and Skewness, IDM & SVM	71.10%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis, QSVM & SURF.	98.91%

• Analysis

In Previous Approach of Literature Survey 2, the technique which is used as feature extraction is Gray Level Co-Occurrence Matrix (GLCM) and all color features

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

Problem of Using IDM feature is some boundary values are required to stabilize the dynamic system.

As classifier in the previous approach they used Linear Support Vector Machine to classify. But the main problem with Linear Support Vector Machine ---

1. Linear SVM does not perform well when the data set has more noise, i.e. target classes overlap.

2. When the number of features for each data point exceeds the number of training data samples, the Linear SVM underperforms.

3. Because the support vector classifier operates by placing data points above and below the classifying hyperplane, there is no probabilistic justification for the classification.

But in the proposed approach there is K-nearest for extracting the Diseased leaves from healthy leaves and the used quadratic Support vector machine for classification because QSVM is a dual optimization form that is a quadratic decision function that may separate data non-linearly.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

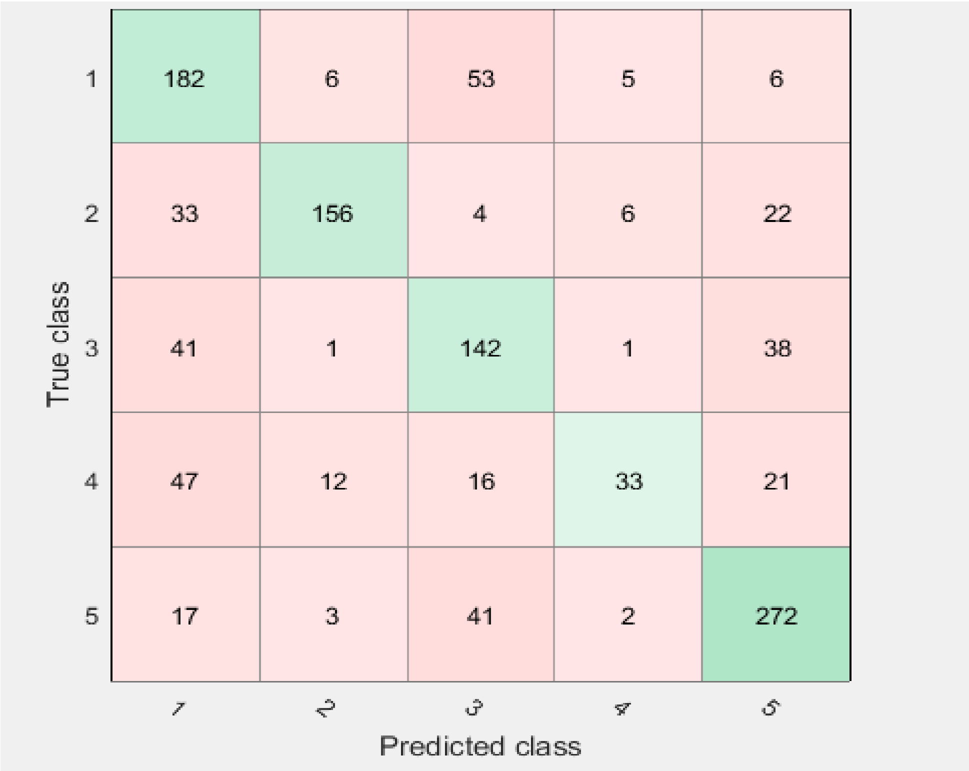


Figure28: Confusion matrix of Previous Approach

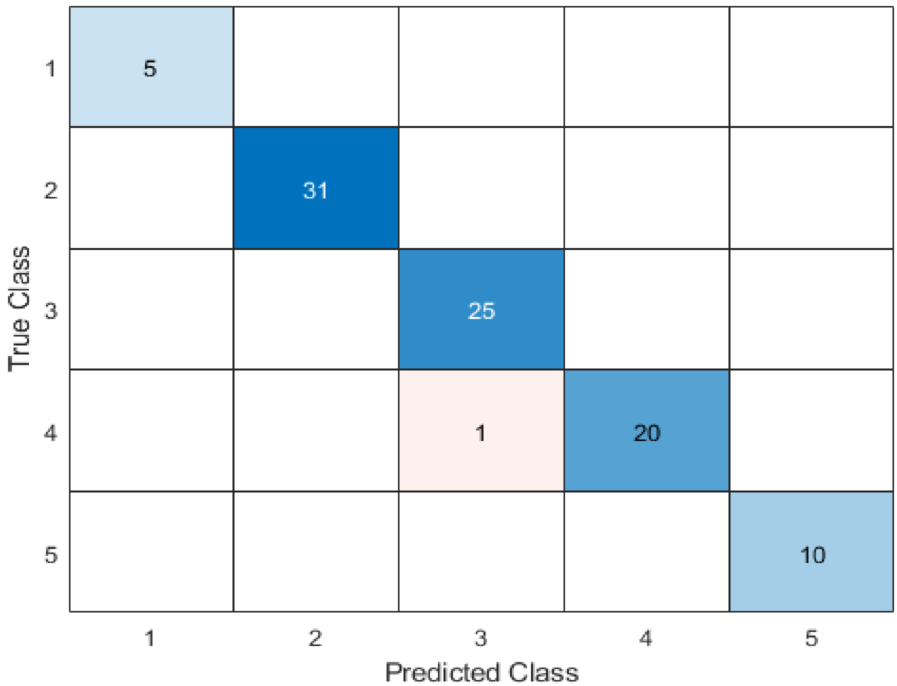


Figure29: Confusion matrix of Proposed Approach Using SURF comparison

5.5 Comparative Analysis 5

Comparison of practice work and literature surveys

TABLE 8: Tabulation of Results of Comparative analysis 5

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[6]	400	Early scorch, Ashen mould, Late scorch, Cottony mould and Ting whiteness are the types of diseases Disease (3-types)	Mean, Contrast, Correlation, Energy, SD, Entropy, RMS, Variance, Kurtosis, Skewness, IDM, and Homogeneity& SVM	89.61%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Mean, SD, Skewness, IDM, Contrast, Correlation, Entropy, RMS, Variance, Kurtosis, Energy and Homogeneity& SVM	64.40%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, , RMS, Skewness, Variance, Kurtosis, SD, Entropy, QSVM & SURF.	98.91%

- **Analysis**

In Previous Approach of Literature Survey 2, the technique which is used as feature extraction is Gray Level Co-Occurrence Matrix (GLCM) and all color features

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

As classifier in the previous approach they used Linear Support Vector Machine to classify. But the main problem with Linear Support Vector Machine ---

1. Linear SVM performs poorly when the data set has more noise, i.e. target classes overlap.
2. The Linear SVM will underperform when the number of features for each data point exceeds the number of training data samples.
3. There is no probabilistic justification for the classification because the support vector classifier operates by placing data points above and below the classifying hyperplane.

But in the proposed approach there is K-nearest for extracting the Diseased leaves from healthy leaves and the used quadratic Support vector machine for classification because QSVM is a dual optimization form that is a quadratic decision function capable of segregating data non-linearly.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

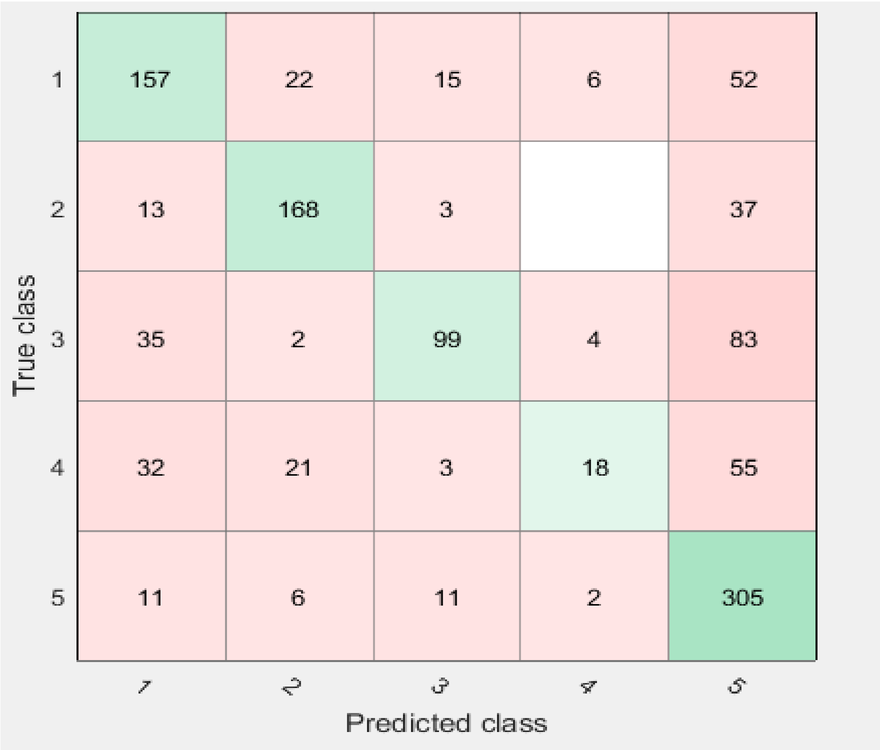


Figure30: Confusion matrix of Previous Approach

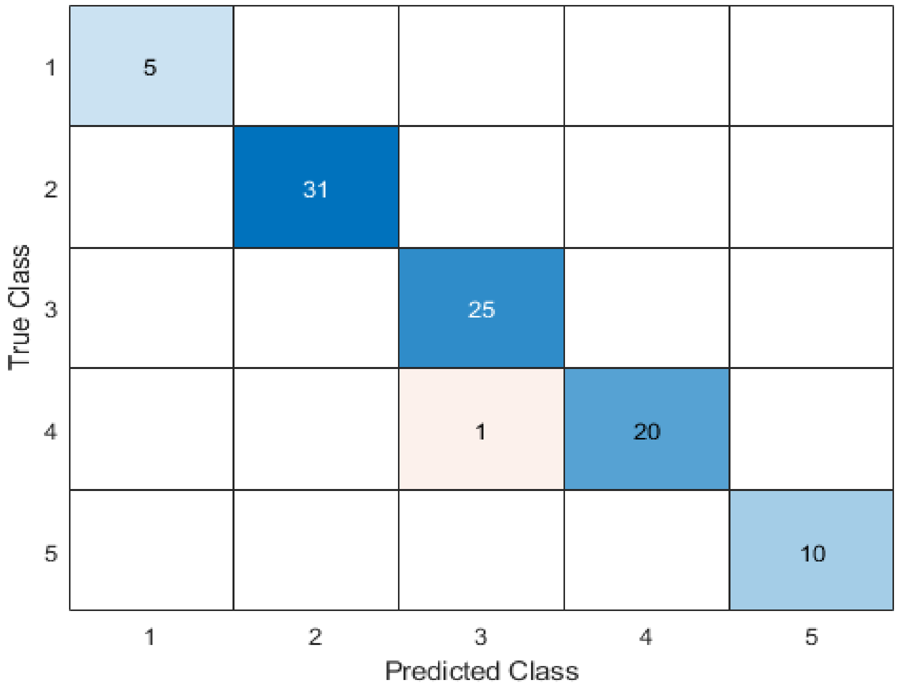


Figure31: Confusion matrix of Proposed Approach Using SURF comparison

5.6 Comparative Analysis 6

Comparison of practice work and literature surveys

TABLE 9: Tabulation of Results of Comparative analysis 6

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[7]	330	Paddy leaves burst & brown leaves diseases Disease (2-types)	GLCM & KNN	76.59%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	GLCM & KNN	52.4%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis, QSVM & SURF.	98.91%

- **Analysis**

In Previous Approach of Literature Survey 2, the technique which is used as feature extraction is Gray Level Co-Occurrence Matrix (GLCM).

As In the previous Approach there is no color feature extracted so there is a problem in finding disease color.

So, for this leaf disease recognition the color features of the diseased leaves need to be extracted.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

As classifier in the previous approach they used K-Nearest neighbor to classify. But the main problem with KNN doesn't work well with a large dataset and doesn't work well with a high number of dimensions.

But in the proposed approach there is K-nearest for extracting the Diseased leaves from healthy leaves and the used quadratic Support vector machine for classification because QSVM is a dual optimization form that is a quadratic decision function capable of segregating data non-linearly.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

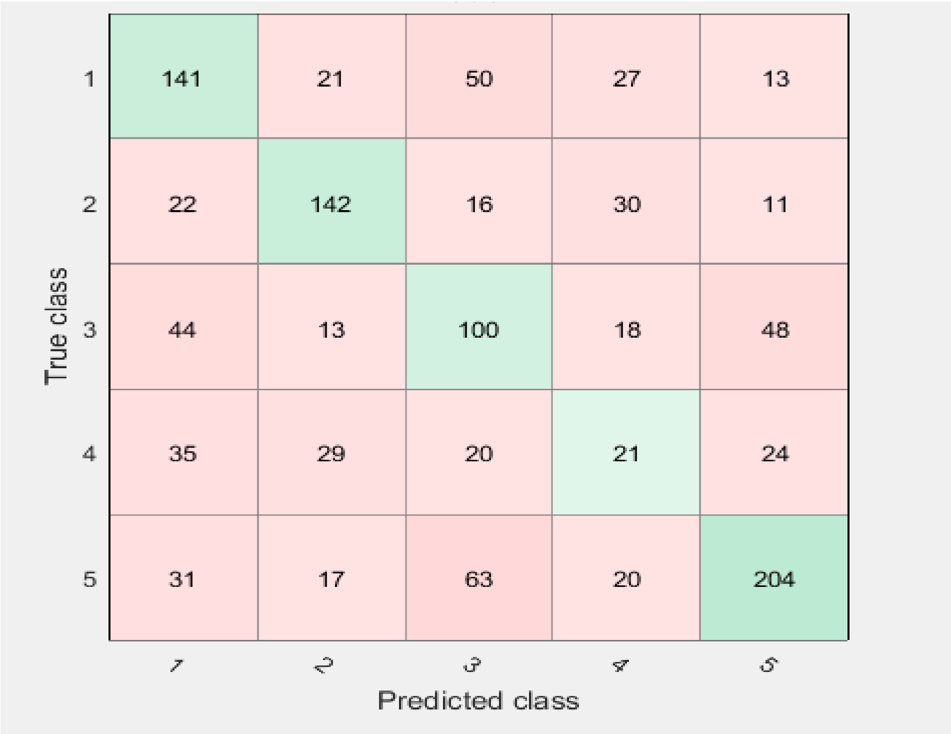


Figure32: Confusion matrix of Previous Approach

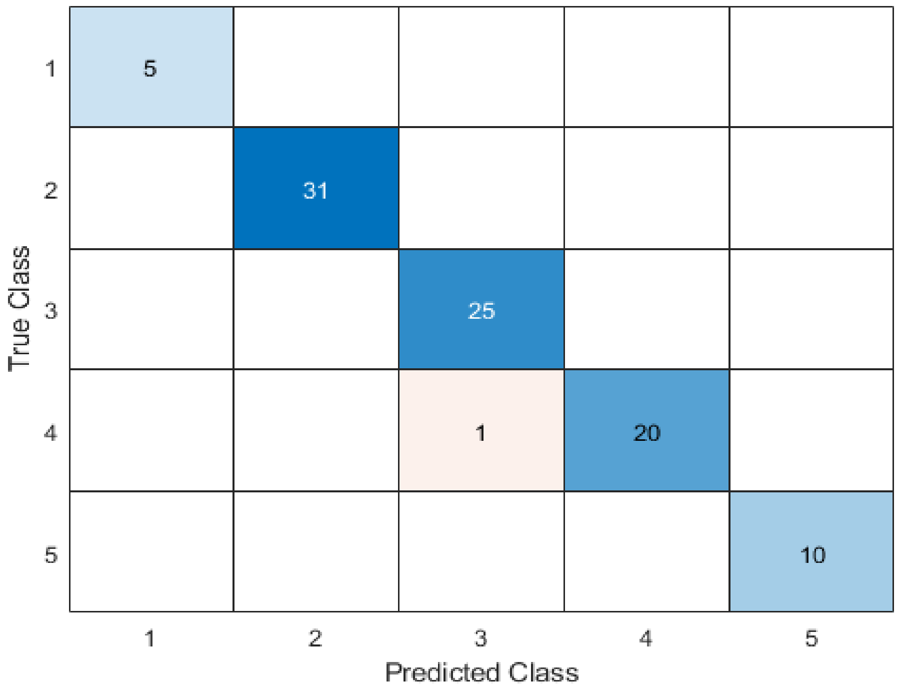


Figure33: Confusion matrix of Proposed Approach Using SURF comparison

5.7 Comparative Analysis 7

Comparison of practice work and literature surveys

TABLE 10: Tabulation of Results of Comparative analysis 7

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[8]	160	Papaya leaf Diseases	HOG & KNN	66.76%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	HOG & KNN	37.60%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis, QSVM & SURF.	98.91%

- ## Analysis

In Previous Approach of Literature Survey 1, the technique which is used as feature extraction is Histogram oriented gradients (HOG). The Main problem with using HOG is it is very sensitive to image rotation.

There are many images which are captured in different angles.

Therefore, HOG is not good choice for classification of textures or objects which can often be detected as rotated image. That's why in the proposed algorithm there is no Histogram oriented Gradient feature (HOG).

As classifier in the previous approach they used K-Nearest neighbor to classify. But the main problem with KNN doesn't work well with a large dataset and doesn't work well with a high number of dimensions.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

As classifier in the previous approach they used support vector machine to classify.

But in the proposed approach there is K-nearest for extracting the Diseased leaves from healthy leaves and the used quadratic Support vector machine for classification because QSVM is a dual optimization form that is a quadratic decision function capable of segregating data non-linearly.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

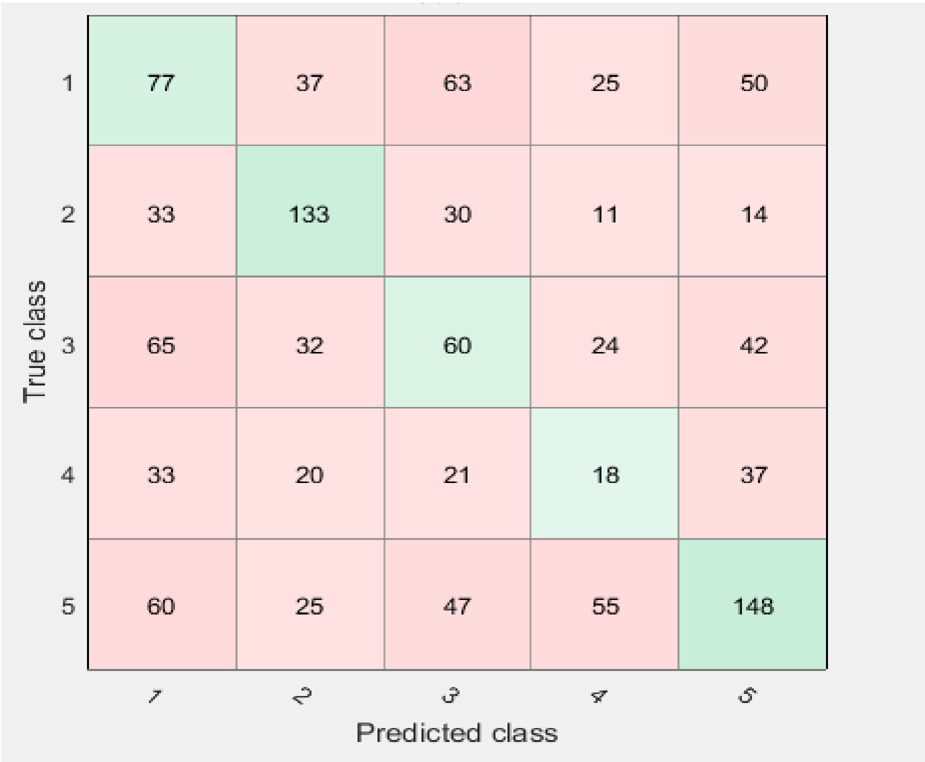


Figure34: Confusion matrix of Previous Approach

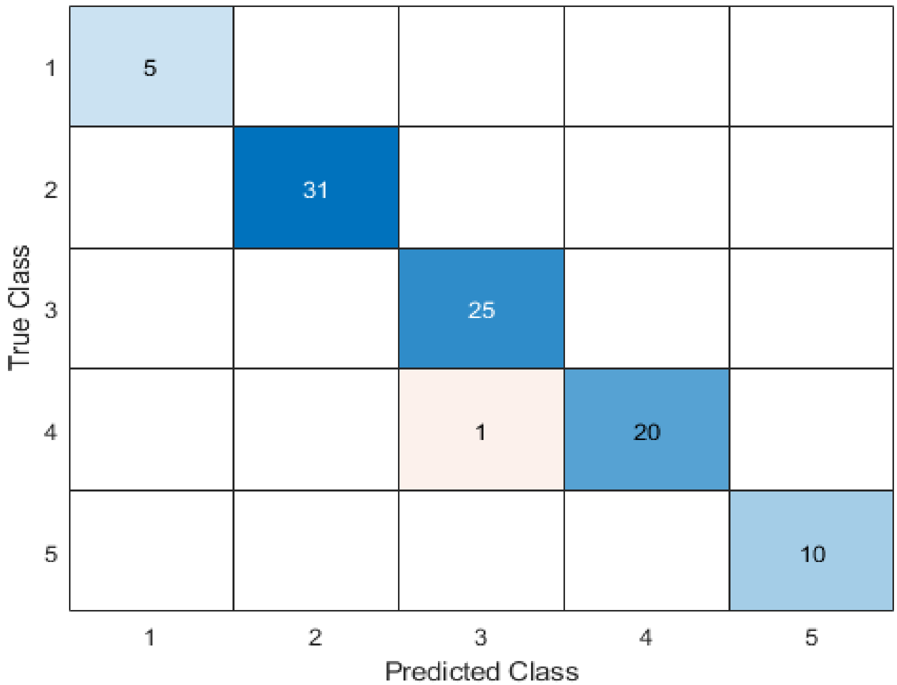


Figure35: Confusion matrix of Proposed Approach Using SURF comparison

5.8 Comparative Analysis 8

Comparison of practice work and literature surveys

TABLE 11: Tabulation of Results of Comparative analysis 8

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[11]	25	Fusarium Wilt, Leaf Curl, Leaf Spots, Ralstonia Bacterial Wilt, and Yellow Virus are the disease types of Chili leaves	GLCM & SVM	88%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	GLCM & SVM	60.1%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis, QSVM & SURF.	98.91%

- **Analysis**

In Previous Approach of Literature Survey 2, the technique which is used as feature extraction is Gray Level Co-Occurrence Matrix (GLCM)

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

As classifier in the previous approach they used Linear Support Vector Machine to classify. But the main problem with Linear Support Vector Machine ---

1. Linear SVM does not perform well when the data set has more noise, i.e. target classes overlap.

2. When the number of features for each data point exceeds the number of training data samples, the Linear SVM underperforms.

3. Because the support vector classifier operates by placing data points above and below the classifying hyperplane, there is no probabilistic justification for the classification.

But in the proposed approach there is K-nearest for extracting the Diseased leaves from healthy leaves and the used quadratic Support vector machine for classification because QSVM is a dual optimization form that is a quadratic decision function capable of segregating data non-linearly.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

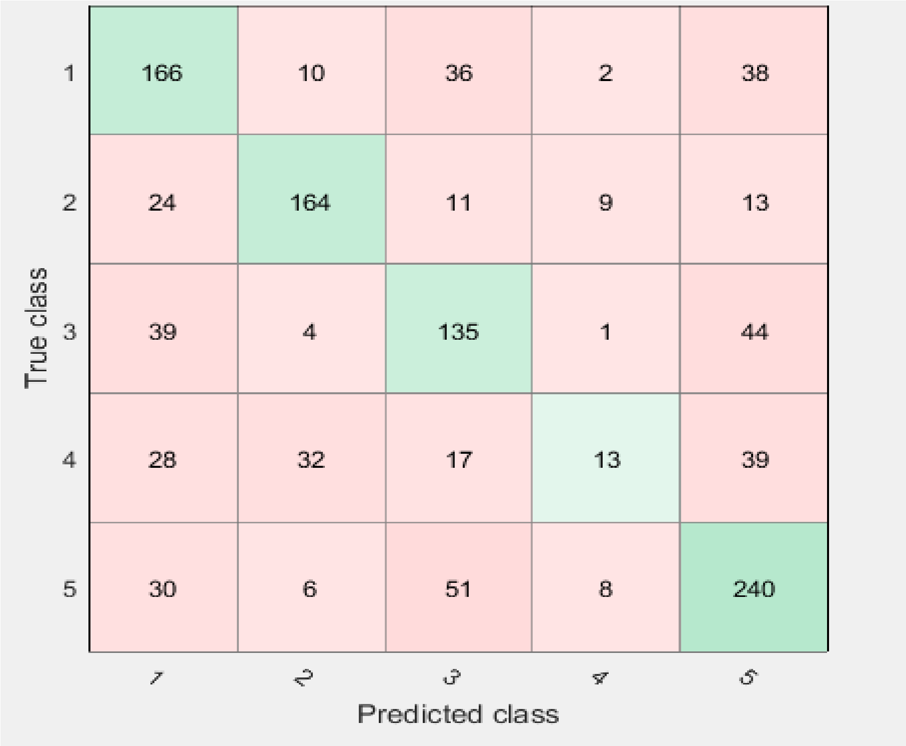


Figure36: Confusion matrix of Previous Approach

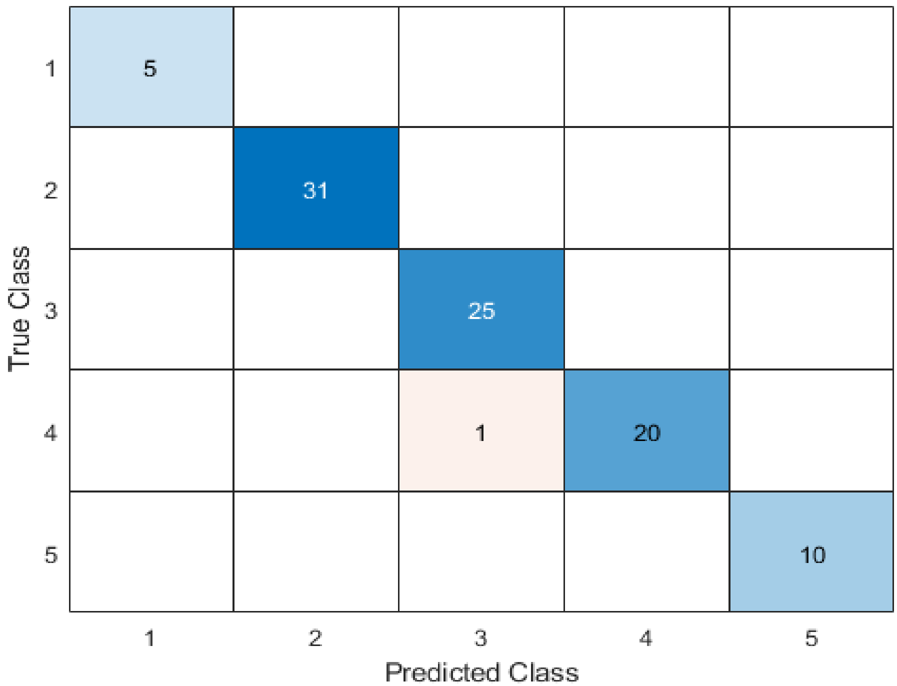


Figure37: Confusion matrix of Proposed Approach Using SURF comparison

5.9 Comparative Analysis 9

Comparison of practice work and literature surveys

TABLE 12: Tabulation of Results of Comparative analysis 9

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[12]	500	Black Rot affected leaves. Cedar Apple Rust affected leaves. Healthy apple leaves.	GLCM, Mean, SD, Entropy, RMS, IDM, Smoothness, Variance, Kurtosis & SVM	96%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	GLCM, Mean, SD, Entropy, RMS, IDM, Smoothness, Variance, Kurtosis & SVM	64.40%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis, QSVM & SURF.	98.91%

- **Analysis**

In Previous Approach of Literature Survey 2, the technique which is used as feature extraction is Gray Level Co-Occurrence Matrix (GLCM) and all color features

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

In previous approach they did not use skewness but skewness is important because it can be either positive or negative or it may even be undefined. They also turn up the data point of high skewness into skewed distribution.

As classifier in the previous approach they used Linear Support Vector Machine to classify. But the main problem with Linear Support Vector Machine ---

1. 1. Linear SVM does not perform well when the data set has more noise, i.e. target classes overlap.
2. 2. When the number of features for each data point exceeds the number of training data samples, the Linear SVM underperforms.
3. 3. Because the support vector classifier operates by placing data points above and below the classifying hyperplane, there is no probabilistic justification for the classification.

But in the proposed approach there is K-nearest for extracting the Diseased leaves from healthy leaves and the used quadratic Support vector machine for classification because QSVM is a dual optimization form that is a quadratic decision function that is capable of separating data non-linearly.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

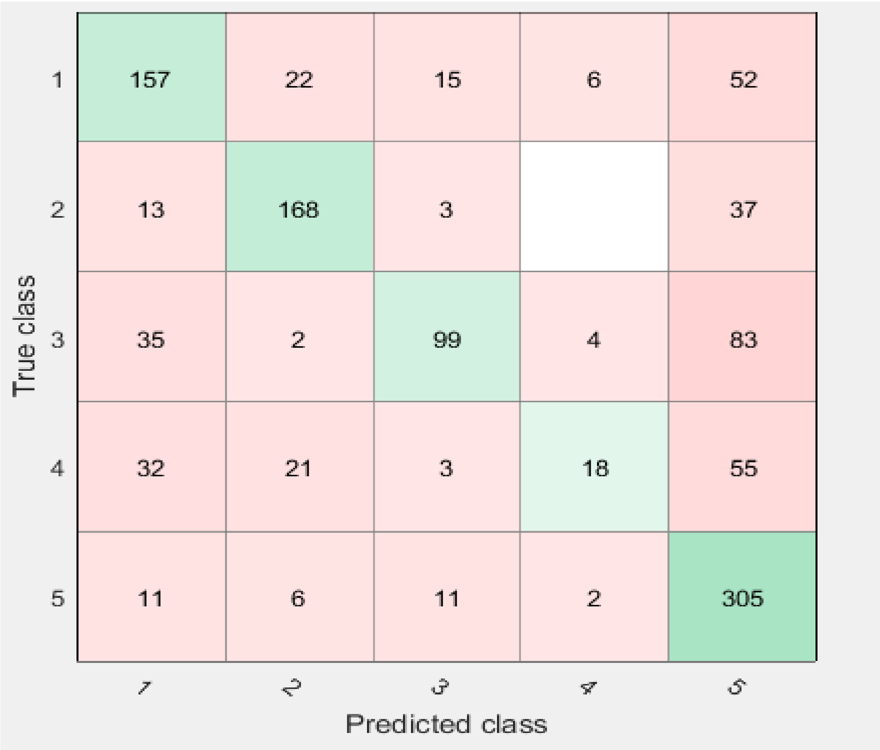


Figure38: Confusion matrix of Previous Approach

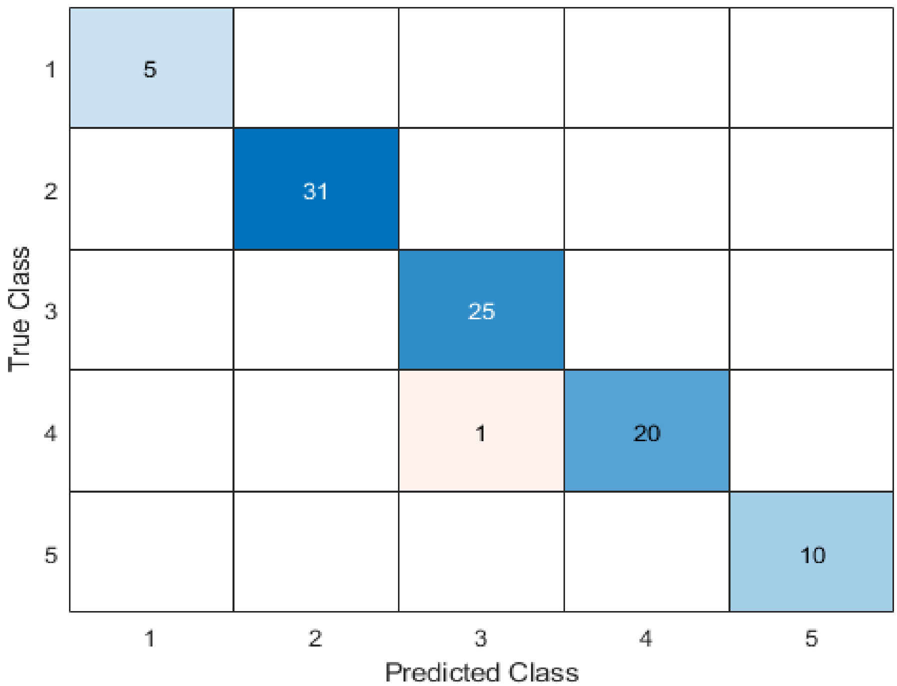


Figure39: Confusion matrix of Proposed Approach Using SURF comparison

5.10 Comparative Analysis 10

Comparison of practice work and literature surveys

TABLE 13: Tabulation of Results of Comparative analysis 10

	Dataset	Diseases	Feature Extraction & Classifier	Accuracy
[12]	300	Potato Leaf Diseases.	GLCM & KNN	86%
Previous Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	GLCM & KNN	52.4%
Proposed Approach	1160	Mango, Arjun, Alstonia Scholaris, Guava, Jamun Disease(5-types)	Knn, GLCM, Mean, SD, Entropy, RMS, Skewness, Variance, Kurtosis, QSVM & SURF.	98.91%

- ## Analysis

In Previous Approach of Literature Survey 2, the technique which is used as feature extraction is Gray Level Co-Occurrence Matrix (GLCM).

As In the previous Approach there is no color feature extracted so there is a problem in finding disease color.

So, for this leaf disease recognition the color features of the diseased leaves need to be extracted.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases using this Algorithm.

As classifier in the previous approach they used K-Nearest neighbor to classify. But the main problem with KNN doesn't work well with a large dataset and doesn't work well with a high number of dimensions.

But in the proposed approach there is K-nearest for extracting the Diseased leaves from healthy leaves and the used quadratic Support vector machine for classification because QSVM is a dual optimization form that is a quadratic decision function that is capable of separating data non-linearly.

And then to remove the problem of having the problem of different kind diseases that could have showing similar diseases another feature extraction is done that is Speeded up robust feature and then compare them.

By the confusion matrix we can easily saw that the Classifier images are not Classified Correctly.

There is having the problem of different kind diseases that could have showing similar diseases similar diseases.

So, here is the confusion matrix.

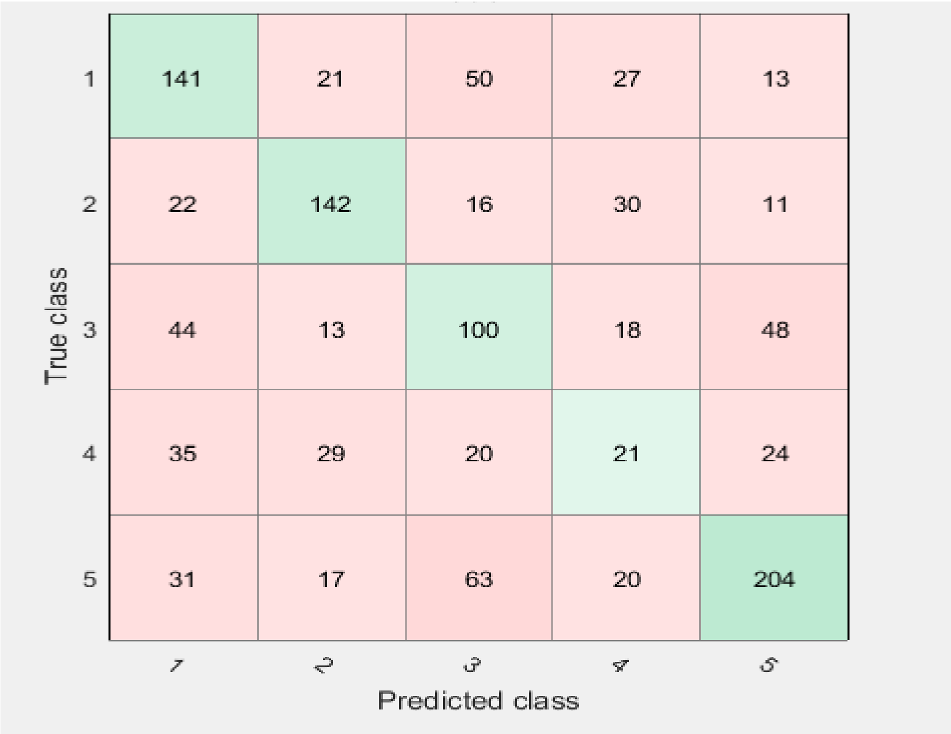


Figure40: Confusion matrix of Previous Approach

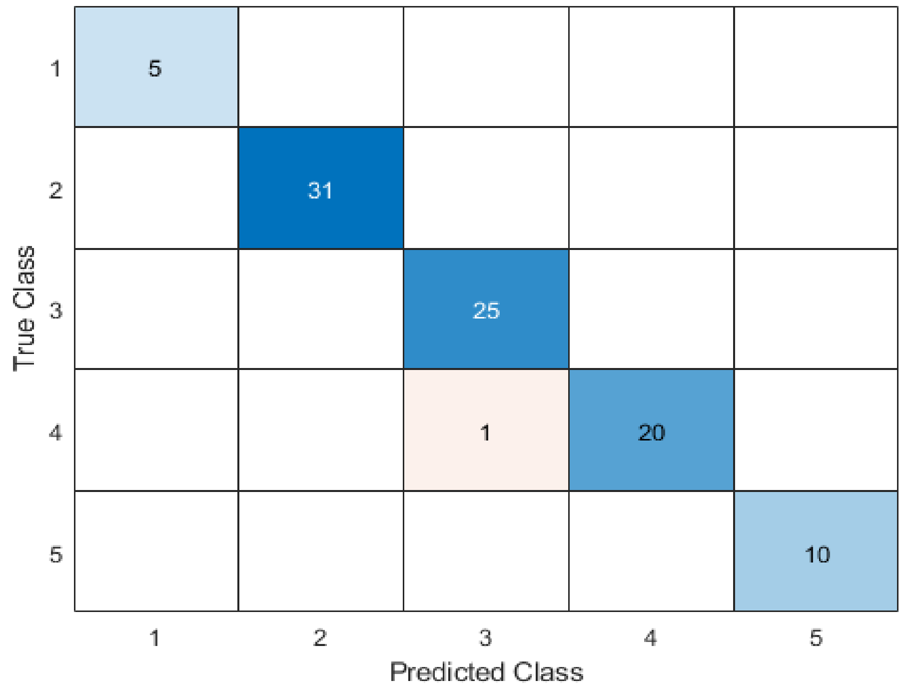


Figure41: Confusion matrix of Proposed Approach Using SURF comparison

5.11 Chapter Summary:

The conclusion drawn from analysis of musical instrument classification are as follows

1) In 1st case the previous work is done HOG features which has a problem of rotation invariant. And there is a problem of classification of different diseases as similar one. In the proposed approach this feature is not used and this problem is also solved. So, the proposed algorithm is pretty better than previous one.

2) In 2nd case GLCM and Local binary patter are used but Local binary pattern is not suitable for big dataset. And there is a problem of classification of different diseases as similar one. In the proposed approach this feature is not used and this problem is also solved. So, the proposed algorithm is pretty better than previous one.

3) In 3rd case GLCM for feature extraction and Linear SVM is used for classification. However, Linear SVM does not perform well when the data set has more noise, i.e. target classes overlap. Furthermore, there is an issue with classifying distinct disorders as comparable. This feature is not used in the suggested technique, and the problem is also solved. As a result, the suggested approach outperforms the existing one.

4) In the fourth scenario, GLCM and all Color features are utilized, but for classification, linear SVM is employed, which has the same issue of not performing well when the data set has more noise, i.e. target classes overlap. Furthermore, there is an issue with classifying distinct disorders as comparable. This feature is not used in the suggested technique, and the problem is also solved. As a result, the suggested approach outperforms the existing one.

5) In 5th case GLCM and all Color feature is used except smoothness but Smoothness has a big advantage Smoothing may be used in two important ways that can aid in data analysis by being able to extract more information from the data as long as the assumption of smoothing is reasonable. And for classification linear SVM is used So, there is a problem of classification of different diseases as similar one. In the proposed approach this feature is not used and this problem is also solved. So, the proposed algorithm is pretty better than previous one.

6) In 6th case for feature extraction GLCM is used and for classification KNN is used. But the main problem is KNN does not work well with large dataset and does not work well with the high dimension. And there is a problem of classification of different diseases as similar one. In the proposed approach this feature is not used and this problem is also solved. So, the proposed algorithm is pretty better than previous one.

7) In the seventh scenario, HOG is utilized for feature extraction, which has a rotation invariant issue, and KNN is employed for classification. However, the primary issue is that KNN does not function well with huge datasets and does not perform well with high dimension. Furthermore, there is an issue with classifying distinct disorders as comparable. This feature is not used in the suggested technique, and the problem is also solved. As a result, the suggested approach outperforms the existing one.

8) In the eighth scenario, GLCM is utilized for feature extraction. For classification, linear SVM is utilized. However, Linear SVM does not perform well when the data set has more noise, i.e. target classes overlap. Furthermore, there is an issue with classifying

distinct disorders as comparable. This feature is not used in the suggested technique, and the problem is also solved. As a result, the suggested method outperforms the old one, and the proposed strategy is enhanced.

9) In the ninth scenario, GLCM and all Color features are utilized, however for classification, linear SVM is employed, which has the same issue of not performing well when the data set has more noise, i.e. target classes overlap. Furthermore, there is an issue with classifying distinct disorders as comparable. This feature is not used in the suggested technique, and the problem is also solved. As a result, the suggested method outperforms the existing one, and the proposed technique improves accuracy.

10) In 10^{th} case for feature extraction GLCM is used and for classification KNN is used. But the main problem is KNN does not work well with large dataset and does not work well with the high dimension. And there is a problem of classification of different diseases as similar one. In the proposed approach this feature is not used and this problem is also solved. So, the proposed algorithm is pretty better than previous one.

6.CONCLUSIONS AND FUTURE SCOPES:

The two most important and popular field of research in the domain of sound recognition are dealt with, in this study. After observing the experimentations and in-depth analysis following conclusions can be drawn.

6.1 Conclusions:

1. Human life is entirely reliant on nature and plants. As a result, there should be specific strategies for saving plants against illness. Crop output declines have an impact on the country's economy. There is a need for an acceptable study approach that can identify plant leaf disease automatically.

2. The food inflation is a national challenge. Scientists, agriculturists work day and night to promote the yield of food grains. It is very difficult to infer the varieties of a leaf disease by simple visual observation.

3. It is very time consuming and can be accomplished by the trained botanists. Research outcome can help to recognition of leaf diseases quickly and easily by machine instead of manual system.

4. The primary goal of this system is to increase the accuracy of automated plant disease detection.

5. In this work ten comparative analysis was done. From all the comparative analysis, all the drawbacks and limitations were discussed. To overcome the discussed limitations, a reliable algorithm is proposed.

6. All the steps of the proposed algorithm are explained. Finally, the experimental results of the proposed model are analyzed and a good accuracy is obtained.

7. In this work, we used KNN to distinguish diseased leaves from healthy leaves and SVM classifiers to identify disease classes infected with mango, arjun, Alstonia scholaris, guava, and jamun leaves.

8. The approach was appropriately implemented, and performance tests using MATLAB software were performed. These five leaves are usually afflicted with diseases such as black mildew, powdery mildew, Pauropsylla Tuberculata, algal, and anthracnose, all of which cause yield loss.

9. As a result, these five plants' disease detection was discovered through the use of Machine Learning. Finally, depending on the accuracy rate, we determine which algorithm works best for which crop.

10. Experiment findings reveal that the suggested system can identify and categories plant diseases with a 98.91 percent accuracy.

As a result, the general conclusion and contributions have been presented above. The next debate will be on the future scope of study.

6.2 Future Scope:

1. In future development, the database will be expanded to include additional plant disease identification and a larger number of data points for classification training.

2. The system's accuracy will improve when the training data is increased. Then we may compare the system's accuracy rate and speed.

3. More studies on Plant Disease Recognition will aid in this process.

4. In future, a real time application based on proposed approaches is to be built.

5. To create a system that can identify pesticides for specific pests and diseases.

References:

[1] N. Kanaka Durga, G. Anuradha, “Plant Disease Identification Using SVM and ANN Algorithms”, International Journal of Recent Technology and Engineering (IJRTE), February 2019.

[2] Yin Min Oo, Nay Chi Htun, “Plant Leaf Disease Detection and Classification using Image Processing”, International Journal of Research and Engineering(IJRE), September-October 2018.

[3] Md. Rasel Mia, Sujit Roy, Subrata Kumar Das and Md. Atikur Rahman, “Mango Leaf Diseases Recognition Using Neural Network and Support Vector Machine” International Journal of Computer Science Engineering (IJCSE), Vol. 8 No.01 Jan-Feb 2019

[4] T. R. Ganesh Babu, S. Priya, J. Gopi chandru, M. Balamurugan, J. Gopika and R. Praveena, “Prediction and Analysis of Plant-Leaf Disease in Agricultural by using Image Processing and Machine Learning Techniques”, International Conference on Computational Performance Evaluation (ComPE), Dec 1–3, 2021.

[5] Indumathi.R, Saagari.N, Thejuswini.V and Swarnareka.R have presented a paper on “Leaf Disease Detection and Fertilizer Suggestion”, Proceeding of ICSCAN, 2019.

[6] Suresha M and Shreekanth K N, Thirumalesh B V have presented a paper on “Recognition of Diseases in Paddy Leaves Using kNN Classifier”, 2017 2nd International Conference for Convergence in Technology (I2CT)

[7] Shima Ramesh, Mr. Ramachandra Hebbar, Niveditha M, Pooja R, Prasad Bhat N, Shashank N, Mr. P V Vinod have presented a paper on “ Plant Disease Detection Using Machine Learning”, 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control.

[8] Yuslena Sari, Andreyan Rizky Baskara, Rika Wahyuni have presented a paper on “ Classification of Chili Leaf Disease Using the Gray Level Co-occurrence Matrix (GLCM) and the Support Vector Machine (SVM) Methods”, 2021 Sixth International Conference on Informatics and Computing (ICIC).

[9] Soarov Chakraborty, Shourav Paul, Md. Rahat-uz-Zaman have presented a paper on “Prediction of Apple Leaf Diseases Using Multiclass Support Vector Machine”, 2021 2nd International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST).

[10] Sakshi Sharma, Vatsala Anand, Swati Singh have presented a paper on ” Classification of Diseased Potato Leaves Using Machine Learning”, 10th IEEE International Conference on Communication Systems and Network Technologies.

[11] Jyoti, Prince Kumar, “A Brief Review on Plant Disease Detection Using Image Processing Techniques”, International Journal of Computer Science and Engineering, vol. 7, no. 9, September 2019. pp. 112-114.

[12] P. Chaitanya Reddy, M. Ayyappa Reddy, Rachakulla Mahesh Sarat Chandra, Mahesh T R, Vadiraj P, Sindhu Madhuri G, “Detection of Plant Leaf-based Diseases Using Machine Learning Approach”, IEEE International Conference on CSITSS, 2021.

[13] Ayswarya.R, Balaji.B, Balaji.R , Arun.S , Ramya.R, “Weed Detection in Agriculture Using Image Processing”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol. 6, no. 3, March 2017. pp. 1698-1703.

[14] Rima Herlina S. Siburian, Rahmi Karolina, Phong Thanh Nguyen, E. Laxmi Lydia, K. Shankar, “Leaf Disease Classification using Advanced SVM Algorithm”, International Journal of Engineering and Advanced Technology, vol. 8. No. 6S, August 2019. pp. 712-718.

[15] Mohammed A. Hussein, Amel H. Abbas, “Plant Leaf Disease Detection Using Support Vector Machine”, AlMustansiriyah Journal of Science, vol. 30, no. 1, August 2019. pp. 105-110.

[16] Vijai Singh, A.K. Misra, “Detection of plant leaf diseases using image segmentation and soft computing techniques”, Journal of Information processing in Agriculture, vol. 4, 2017. pp. 41-49.

[17] Prof. Sanjay B. Dhaygude, Mr.Nitin P.Kumbhar, “Agricultural plant Leaf Disease Detection Using Image Processing”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol. 2, no.2, January 2013. pp. 599-602.

[18] K. Vinoth Kumar, T. Jayasankar, “An Identification of Crop Disease Using Image Segmentation”, International Journal of Pharmaceutical Sciences and Research, vol. 10, no.3, March 209. pp. 1054-1064.

[19] Harshadkumar B Prajapati, Jitesh P. Shah, Vipul K. Dabhi, “Detection and Classification of Rice Plant Diseases”, Journal Intelligent Decision Technologies: IOS Press, vol. 11, no. 3, August 2017. pp. 357-373.

[20] Suhaili Beeran Kutty, Noor Ezan Abdullah, Dr. Hadzli Hashim, A’zraa Afhzan Ab Rahim, Aida Sulinda Kusim,Tuan Norjihan Tuan Yaakub, Puteri Nor Ashikin Megat Yunus, Mohd Fauzi Abd Rahman, “Classification of Watermelon Leaf Diseases Using Neural Network Analysis,” IEEE, Business Engineering and Industrial Applications Colloquium (BEIAC), Langkawi, pp 459 – 464, 2013.

[21] Sanjeev S. Sannaki, Vijay S. Rajpurohit, V B Nargund,Pallavi Kulkarni, “Diagnosis and Classification of Grape Leaf Diseases using Neural Network,” IEEE , Tiruchengode, pp 1 – 5, 2013.

[22] Akhtar, Asma, AasiaKhanum, Shoab Ahmed Khan, and ArslanShaukat. "Automated Plant Disease Analysis (APDA): Performance Comparison of Machine Learning Techniques." IEEE International Conference on Frontiers of Information Technology (FIT), pp. 60-65, 2013.

[23] Ms. Kiran R. Gavhale, Prof. Ujwalla Gawande, Mr. Kamal O. Hajari, “Unhealthy Region of Citrus Leaf Detection Using Image Processing Techniques,” IEEE, International Conference for Convergence of Technology, Pune, pp 1-6, 2014.

[24] Bhong, Vijay S., and B. V. Pawar. "Study and Analysis of Cotton Leaf Disease Detection Using Image Processing." International Journal of Advanced Research in Science, Engineering and Technology, 2016.

[25] Sachin D. Khirade, A. B. Patil, “Plant Disease Detection Using Image Processing,” IEEE, International Conference on Computing Communication Control and Automation, Pune, pp 768-771, 2015.

[26] Usama Mokhtar, Mona A. S. Alit, Aboul Ella Hassenian, Hesham Hefny “Tomato leaves diseases detection approach based on support vector machines” IEEE pp. 978-1-5090-0275-7/15, 2015.

[27] M. Ranjan, R. M. Weginwar, N. Joshi and A. B. Ingole, “Detection and classification of leaf disease using artificial neural network,” International Journal of Technical Research and Applications, vol. 3(3), pp. 331-333, 2015.

[28] R. R. Kajale, “Detection & Recognization of Plant Leaf Diseases Using Image Processing and Android O.S.,” International Journal of Engineering Research and General Science, vol. 3(2), pp. 6-9, 2015.

[29] H. Kulkarni, and K.R. PatilAshwin, “Leaf Disease Classification Using Artificial Neural Networks and Decision Tree Classifier,”Journal of Image Processing & Pattern Recognition Progress, vol. 5(1), 2018.

[30] A. Garg, V. Vijayaraghavan, S. S. Mahapatra, K. Tai, and H. C. Wong, “Performance evaluation of microbial fuel cell by artificial intelligence methods,” Expert Systems with Applications, vol. 41(4), pp. 1389–1399, 2014.