Early Detection and Control of Anthracnose Disease in Cashew Leaves to Improve Crop Yield using Image Processing and Machine Learning Techniques

Sudha P (viswasudha@gmail.com)
National Institute of Technology Puducherry
Kumaran P
National Institute of Technology Puducherry

Research Article

Keywords: Anthracnose Disease, Leaf Image Segmentation, Precision Agriculture, Machine Learning Techniques

Posted Date: January 23rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2490123/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Signal, Image and Video Processing on May 23rd, 2023. See the published version at https://doi.org/10.1007/s11760-023-02552-9.

Abstract

Agriculture is one of the primary pillars powering India's economy. It is alarming to note that India's agriculture rate is declining steeply. Climate change, environmental pollution, and soil erosion are well-known factors affecting crop productivity. The increasing prevalence of plant diseases is also a significant contributing factor affecting agriculture. Early disease detection and mitigation actions based on identified diseases in the plants are critical in increasing crop productivity. This study considers a machine-learning model for detecting disease in cashew leaves. This work concentrates on Anthracnose disease, which leads to severe yield loss when it affects the cashew plant. In this regard, cashew leaves are collected and used to train various machine learning classifiers to identify and classify the disease. This work focuses on the segmentation and classification of leaves using various Machine Learning models. For this, Basic segmentation approaches like Global threshold, Adaptive Gaussian, Adaptive Mean, Otsu, Canny, Sobel, and K-Means, and Machine Learning models like Random Forest, Decision Tree, KNN, Logistic Regression, Gaussian Naive Bayes Classifiers are employed. The final classification employs a Hard and Soft voting classifier in addition to the Decision Tree, KNN, Logistic Regression, and Gaussian Naive Bayes classifiers. Finally, we observe that K-Means segmentation with Random Forest outperforms other classifiers. The accuracy obtained from the Random Forest classifier is 96.7% for the CCDDB dataset, and the accuracy obtained from the Random Forest classifier is 99.7% for the PDDB dataset.

1. Introduction

Computer vision is a broad area where machines are trained to get, process, and analyze images automatically with the help of machine learning and artificial intelligence, just as a human would. A subset of artificial intelligence is machine learning, which includes computational techniques for predicting and analyzing the patterns present in given data, including images. Typically, Machine learning demands data pre-processing and classification. Image processing is the manipulation of an image into a preferred form. There are two types of image processing, analog and digital. A digital image is a grid of pixels represented in the form of a matrix. Comparing digital image processing to visual analysis, it is clear that digital image processing is a more successful approach. Features can be obtained by manipulating digital images. There are various techniques for selecting features from images and categorizing them through analysis.

Image processing in agriculture gives plenty of rewards, but various factors such as climatic variance impact crop productivity. As plants are affected by various diseases, it is difficult for humans to detect them early. Nowadays, novice farmers lacking experience require subject experts to help with disease identification and leaf classification. Intelligent systems are therefore expected to fill in the knowledge gaps and make disease detection seamless [1].

According to the government of India's annual report for 2021–2022, a net 197.3 million hectares out of 328.7 million hectares are used as cropped areas. The Department of Horticulture and Plantation Crops of the Government of Tamil Nadu reported that 91,058 hectares (hectors) are cultivated for plantations, among which 0.64 M. T. ha of production was attained. In the past, cashew disease was thought to be of minimal concern. Today several diseases have been shown to be severe enough to cause cashew trees to suffer major losses. This farm crop is reported to be attacked by several fungi, which reduces productivity [35]. Numerous biotic and abiotic restrictions are a hazard to cashew, causing severe output losses. The most detrimental biotic restrictions are diseases and pests, which reduce nut productivity.

In fact, more than 12 diseases have been identified as affecting cashew trees globally. In countries that produce cashews, anthracnose foliar blight, fruit rot, and gummosis of twigs and trunks are frequently regarded as the

diseases most likely to cause significant harm [36]. So, this work concentrates on identifying Anthracnose disease, which causes severe yield loss when it affects the cashew plant. Fig1 (a), (b), and (c) show the gummosis and anthracnose diseases that might occur in the cashew tree.

A routine monitoring system for plant disease identification becomes crucial in boosting productivity and production quality. The digital image processing system could be a powerful tool for diagnosing challenging symptoms far sooner than the naked eye [6]. It allows farmers to act appropriately and quickly to protect the crop and obtain the required quality and production of agricultural products [22,34].

Many researchers for the detection and classification of diseases have already put out numerous segmentation and classification approaches. These approaches have made it possible to eliminate the issues, but nowadays, the challenge is to make these results effective. This work has identified the best-performing image segmentation and classifier on the Cashew Crop Disease Data Base [CCDDB] that would perform the best detection and classification. This work measures different image segmentation techniques such as global threshold, Adaptive Mean, Adaptive gaussian, Otsu threshold, Canny edge detection, Sobel edge detection, and K means are applied to highlight the infected area. All these segmentations apply to the Random Forest model and ensemble classifiers of Decision Trees, Naive Bayes, KNN, and Logistic Regression. The same model was applied to the Image Database of Plant Disease Symptoms (PDDB) dataset for comparison which yielded a 99.7% result. The paper is structured as follows. Section 2 provides a literature review. Section 3 shows the methodology used for segmentation and classification. Section 4 presents the Experiment and Result Analysis. Section 5 conclusion.

2. Literature Review

Many researchers have previously worked to automatically and accurately diagnose illnesses using various classification approaches, We have surveyed such related papers to our work here. Youwen et al. (2008) detected downy mildew and powdery mildew diseases on cucumber plants. They segmented the leaf photos using statistical patterns and mathematical morphology after using a median filter to reduce the noise. They used SVM classification for disease detection. The downy mildew and powdery mildew diseases were detected using the features taken from the diseased cucumber leaf image and fed to the SVM classifier [7].

Arivazhagan et al. (2013) worked on identifying diseases such as late scorch, Leaf Lesion, bacterial spot, Ashen Mold, Fungal spot, Scorch, sunburn, Late blight, early Blight, and sooty mold, early scorch, brown spots, yellow spots, bacterial diseases, late scorch, and fungal diseases on plant leaves such as Banana, tomato, Beans, Potato, Mango, Lemon, and Jackfruit. A digital camera takes pictures of different leaves to determine the illnesses from the indications of the leaves. By masking and removing the green pixels from digital RGB photos of leaves, HSI color images were created. [22] using a specified threshold value, segmenting the image, and computing the texture surface statistics. They utilized an SVM which made use of textural information to categorize the illness. They claimed that while identifying disease infection in plant leaves, an accuracy of 94.74% was attainable [8].

H. Al-Hiary et al. (2011) employed the Otsu segmentation method and K-means clustering approach to locate the areas of plant and stem illnesses that infect a different plant leaf. For the sake of extracting the characteristics of diseased sections, they employed the color co-occurrence approach. They identified diseases such as Early Scorch, White Mold, Tiny Whiteness, Cottony Mold, Late Scorch, and Ashen Mold using an ANN classifier [9,24].

Sannakki et al. (2013) took pictures of Grape plant leaves and identified the illness using image processing and artificial intelligence techniques. They categorized grape leaf diseases as Downy and Powdery Mildew. To increase the accuracy. Masking was used to get rid of the backdrop.

Anisotropic diffusion was used to preserve the information from the damaged leaf section. The authors applied K-Means as a segmentation technique, and GLCM was used for feature extraction. In the end, a classifier known as the Feed Forward Back Propagation Network classifier performs classification. [10].

Hossain et al. (2019) used the KNN classification approach to identify the diseases such as Alternaria Alternate, Leaf Spot, Anthracnose, Canker, and Bacterial Blight. For the input, 237 leaf photos were obtained from the database of Arkansas plant disease. The segmentation, followed by the GLCM matrix, was used to extract features from photos of sick plants, and the training dataset was subjected to 5-fold cross-validation to avoid overfitting. This process offered 96.76% accuracy [11].

Samajpati et al. (2016) employed the Random Forest classifier to recognize illnesses of the apple fruit, such as Rot, Scab, and Blotch. K mean clustering was used for image segmentation to find contaminated areas, and the fusion approach was used to extract features. Feature level fusion increased the illness classification's accuracy [12].

Pang et al. (2011) endeavoured to detect the illnesses like Cercospora Leaf Spots, Leaf Sheaths, and Brown Spots that have been detected in maize crops. The method found that all pixels for which the green channel's (G) and red channel's (R) and grey levels were more significant than each other. The diseased region consists of 98% red pixels; corresponding regions were detected, appropriately named, and used as seeds in a region-growing approach to more reliably and effectively detect diseased regions to the level of 96% [13].

Ramesh et al. (2020) focus on GLCM feature extraction and k-means clustering for precise position detection in plant leaves. K-mean clustering is used to detect disease on real-time leaf images. Once the detection has been made, the GLCM filter extracts its features. Features-based matching is implemented using an ANN approach for classification, and it attains 96% accuracy [14].

Jayamoorthy et al. (2017) compared K-means and Fuzzy C-Means with the suggested Spatial Fuzzy C Mean (SFCM) to identify plant diseases (FCM) and, likewise, kernel-based FCM (KFCM). The characteristics like the image of the sick leaf, color, and texture [25] were retrieved. The disorders were categorized using a neural network method. The recommended approach produced better results and suggested pesticides to combat the illnesses [16].

Kaur, N. (2021), emphasizing feature extraction, has considered GLCM, LBP, Gabor features, and SIFT. KNN, SVM, ANN, Logistic Regression, and Naive Bayes as a classifier to create an effective ensemble classifier. It has been found that the ensemble classification using the law's mask's hybrid features produced the best results in terms of recall, accuracy, and precision [15].

The approach for identifying and categorizing crop illness using machine learning integrated with digital image processing is presented in this work. With several image processing techniques for pre-processing, segmentation, and feature extraction, Ganatra et al. (2020) have suggested a prediction model for identifying and categorizing various plant leaf diseases. The disease is categorized into one classification using a variety of categorization techniques. On two independent datasets, they tested the suggested model [5].

S.No	Authors & year of publication	Plant	Disease Identified	Segmentation and Classification techniques used	Limitations
1.	Youwen, T. et al. (2008)	Cucumber leaves	Powdery mildew and Downy mildew	Segmented the leaf photos using statistical patterns and mathematical morphology. SVM classifier is used. SVM linear kernel is good.	Limited training sets of samples are appropriate for SVM. The linear kernel is reasonable if many features are obtained and the distribution is good.
2.	Arivazhagan, S. et al. (2013)	Banana, tomato, Beans, Potato, Mango, Lemon, & Jackfruit	Early scorch, Brown spots, Yellow spots, A bacterial disease, Late scorch, and Fungal diseases	After color transformation, the infected leaf region is split into no equal patches, and the texture feature is extracted using the color co-occurrence approach. SVM classifier is used. The accuracy attained is 94.47%	The testing and training image size is only limited to 10 to 30. The computed value of texture properties is not specified.
3.	H. Al-Hiary, S. et al. (2011)	Plant stem and leaves	Early Scorch, White Mold, Tiny Whiteness, Cottony Mold, Ashen Scorch, and Late Mold	Otsu threshold is selected in order to define the variable threshold value. K means clustering is used to segment the leaf image where 3 or 4 clusters produce the best result. The Color Co-Occurrence Method is used for feature extraction. A neural network classifier performs classification. The accuracy obtained is 94%	The Dataset descript-tion needs to be mentioned. Number of categories used for training and testing is not.
4.	Sannakki, S. S. et al. (2013)	Grapes leaf image	Downy mildew and Powdery mildew.	Thresholding is applied for masking, and k means is used for segmenting images into 6 groups, and then extraction of feature is accomplished by GLCM. Feed forward back propagation neural network is used	As a whole, 33 images are used for training and testing.
5.	Hossain, E. et al. (2019)	Plants from Arkansas DB & redditplant leaf dataset	Alternaria Alternate, Leaf Spot, Anthracnose, Canker, and Bacterial Blight	Color segmentation is done using a k-nearest neighbor classifier with 3 neighbors. KNN classifier is used.	It is highly inefficient to choose the "right" value K. here the authors used only 237 images, for huge data

					prediction will consume more time.
6.	Samajpati, B. J. et al. (2016)	Apple fruit	Apple scab, apple blotch, and apple rot,	K means clustering is used for segmentation. LBP, CLBP, Local ternary pattern, and Gabor features approaches are used to extract a feature. A random forest classifier is used.	Training and testing data size is less. The Dataset description is missing. The local ternary pattern uses constant to threshold pixels, so the histogram will result in large range.
7.	Pang, J. et al. (2011)	Maize leaf	Cercospora Leaf Spot, Brown Spots and Leaf Sheaths,	LTSRG segmentation algorithm.	Parameter specifications is not given
8.	Ramesh, G. et al. (2020)	Four different categories from the plant image dataset	Alternaria, Bacterial Blight disease, anthracnose, and Cercospora Leaf Spot disease	K means clustering with value 5 is used. GLCM is for feature segmentation extraction. ANN is used for the classification	Parameters of the model must be used to control the underfit and overfit problems that frequently arise while training.
9.	Jayamoorthy, S. et al. (2017)	Several plant species	Bacterial blight, Footrot	They are comparing Spatial Fuzzy C Mean (SFCM) with other clustering methods to identify crop diseases Fuzzy C-Means (FCM), K-means, and Kernel-based FCM (KFCM), The Neural Network model is used.	In FCM, computation takes a long time. Due to its greater use of computational logic. Experimental results and accuracy need to be specified.
10.	Kaur, N. et al. (2021)	Bell pepper, Potato, and Tomato	Early Bight, Bacterial spot, Curl Virus, Target Spot, Yellow leaf, Mosaic Virus, Septoria Leaf Spot	K means segmentation. GLCM, LBP, Law's Texture mask, SIFT, and Gabor are used as feature extractors. Ensemble classifier with SVM, ANN, logistic, Naive Bayes, and KNN	The initial value of k needs to be specified. Parameters of the model must be used to control the underfit and overfit problems that frequently arise while training.
11.	Ganatra, N. et al. (2020)	General Leaves	Early Bight, Bacterial spot, Curl Virus etc.,	Otsu's technique does the segments of the image. Support Vector Machine,	Neural network models have the propensity to overfit on smaller datasets. They

				Artificial Neural Network, Random Forest and K-Nearest Neighbor are used. The accuracy achieved is-73.38%	must generalize successfully to new examples because they memorize the training data.
12.	Zamani, A. S. et al. (2022).	Rice	Brown Spot, Leaf Smut, and Leaf Blight	Background noise is removed via the mean filter. The image quality is improved via histogram equalization. K-Means is used for segmentation and PCA for feature extraction. Then, images are categorized using methods like ID3, RBF-SVM, random forest, and SVM.	An RBF kernel SVM's complexity increases with the training set's size since the RBF kernel is not a parametric model. Here authors used a limited number of training and testing images.
13.	Yang, X. et al. (2022).	15 spices	Leaf recognition	HSV color space segmentation method is used. SVM, BPNN, KNN, and BP-RBF are used for the classification	Backpropagation may be susceptible to erratic and inconsistent data. The training set of data significantly impacts how well backpropagation performs.
14.	Badiger, M. et al. (2022).	Skin or leaf	Different types of diseases in skin and leaf	K-Means segmentation technique, SVM for classification of leaf and skin image.	If the target class in SVM overlaps, the algorithm could not perform as well.
15.	Ansari, A. S. et al. (2022).	Grape	**Black rot,** **Anthracnose, Leaf blight**	Fuzzy c means the segmentation method used, HWT is used for feature extraction, PSO SVM, random forest, and BPNN algorithms are used for classification.	Limited number oftesting and training datasets. There is no certainty that an optimal solution will ever be found when using metaheuristics like PSO.

Through the literature survey, finally, the decision was to determine which basic segmentation would work best with which classifier on our CCDDB dataset. Basic segmentation like Global threshold, Adaptive Gaussian, Adaptive Mean, Otsu, Canny, Sobel, and K-Means is not applied to a single dataset. Basic Segmentation methods with classifiers have yet to be compared. The proposed paper used all of these segmentations and evaluated how well each classifier performed. In order to achieve better accuracy, this work determines which segmentation technique will work best with which classifier.

3. Methodology Of The Proposed System

This section describes the proposed model for cashew leaf disease detection. Determining what type of illness the cashew crop is experiencing is the primary goal of disease identification. The goal of disease management is to predict how an illness will progress. Accuracy is one of many aspects of identifying and categorizing plant leaf diseases.

Steps involved in image classification include Cashew Leaf image dataset collection, image pre-processing, segmentation, and classification. Fig (2) shows the proposed model for cashew leaf disease detection

3.1 Image Acquisition and Dataset Description

This part defines dataset description and image acquisition. The first step in plant leaf classification is data collection. Here Cashew crop leaf image data are collected from the cashew orchard, which is used as input for the classification model. Leaf inputs are taken from a digital camera.

The sample images are gathered from cashew orchards in Konnakavali, Varichikudi, and Karaikal, which is recognized as one of the regions with the highest cashew output in Karaikal. The orchards contain around 200 trees, where the data collected is 400 leaves, out of which 200 are healthy, and 200 are diseased. The samples comprise healthy and diseased leaves, including anthracnose, bacterial leaf spot, red rust, grey blight, minor, shooty, and vein necrosis of cashew leaf. Mainly, the focus is on anthracnose disease. These leaf samples were collected from 10 randomly selected trees that were located along the diagonal of the cashew plantation that was visited.

Each sample included a number, a name, a sampling date, a location, and coordinates. When the blob first appeared, the leaves were gathered, removed from the tree, and photographed the same day. The photos are captured with a digital camera at a resolution of 3020 x 3020 pixels and then resized to 100 x 100 pixels. The image background is white. The data has been collected around one month, and values are Maximum of 35.0 **°C -** 39.8 **°C** temperature, 50% - 56%humidity, and 18kmph - 21 kmph wind speed was recorded from 2.20 pm -3.35 pm, and a Minimum of- 33.1**°C -** 36.2**°C** temperature, 38%-51% humidity, and 13kmph-18kmph windspeed were recorded from 11.00 -12.10pm. Because the size of the leaves may vary, rescaling is necessary to ensure that the training and testing images have the same dimension. All experiments are performed in python 3 (Jupyter Notebook) over the DESKTOP_USVG06J computer with Intel i5-8400 CPU @ 2.80 GHz and 16 GB RAM, running Windows 10 Pro operating system.

3.2 Image Preprocessing

This section describes the image preprocessing used in this work. This step includes normalization and augmentation, and resizing. Digital cameras are used to take pictures. Cashew leaf images are normalized by employing a fixed size in a dataset for processing. [3].

In augmentation, flipping and rotation were applied to expand the CCDDB dataset. Flipping which includes horizontal & vertical flips. Rotations include rotating by 90,180,270 degrees clockwise. It is depicted in Fig 3.(b) and (c). After augmentation 850 diseased leaves and 850 healthy leaves are in our CCDDB dataset.

3.3 Image Segmentation

This section describes the segmentation used in our work. Image segmentation aims to know and identify what the image possesses at the pixel level. This proposal presents different image segmentation methods and the performance of different machine learning classifiers to predict automatically whether the leaf is diseased or healthy, where the Random Forest classifier is compared with Ensemble Classifiers. Segmenting involves dividing an image into its component pieces. While processing the image, the features and components for processing the image can be

obtained. There are several types of segmentation available. Here compared, the performance of different classifiers during which different segmentations are passed as input.

3.3.1 Threshold-based segmentation

Thresholding performs segmentation in an image by fixing all pixels whose intensity is more significant than a threshold value will group in the foreground, and the remaining are grouped in the background value.

3.3.1.1 Global Threshold

Thresholding is the primary option for segmenting images in digital image processing. Thresholding can be used to convert grayscale visuals into binary images. The simplest way of segmenting an image is using a single threshold value for the whole image. When a threshold value is reached, pixels below it are turned to black (bit value 0), and pixels above it are transformed to white (a bit value of one). Typically, thresholding is stated as dividing a picture into the foreground (black) and background (other colors) values (white).

A threshold image g(x,y) is defined as,

G(p,q)= $\left\{ \begin{array}{l} 1, iff\,(p,q) > ThValue \\ 0, iff\,(p,q) \le ThValue \end{array} \right\}$ **(1)**

Where one is represented as an object and 0 as a background, ThValue represents a threshold value [26].

3.3.1.2 Otsu threshold

The algorithm's most basic form yields a single intensity level that separates pixels into the background and foreground classes [30,33].

The following equation can be used to calculate the within-class variance at any threshold :

$$\vartheta \left(t||\phi bg\,(th)\,\vartheta^2 bg\,(th) + \phi fg\,(th)\,\vartheta^2 fg\,(th) \right.$$

2

Where, the probability of the number of pixels for each class reaching the threshold is represented by $\phi bg\,(th)$ and $\phi fg\,(th)$ the variance of color values is represented by ϑ^2.

The following formula can be used to compute the variance:

$$\vartheta^2\,(t) = \sum$$

3

The value of pixel I in the group at position PV_i, the group's average pixel values are represented by $\overline{PV}$ (bg or fg), the number of pixels is N.

3.3.1.3 Adaptive Thresholding

In Adaptive threshold, values of threshold change statically on the image. The smaller region will have a different threshold value. Adaptive thresholding, sometimes referred to as local thresholding, seeks to statistically assess the intensity values of the pixels surrounding a given pixel, p. there are distinctive methods in Adaptive Threshold.

Adaptive Mean, the threshold level is equal to the mean value of the surrounding area minus the constant Value C. Adaptive Gaussian, the threshold value for adaptive Gaussian thresholding is the summation of the neighborhood pixels, where the weights are a gaussian window.

3.3.2 Edge-Based Segmentation

Edge Detection refers to a group of mathematical techniques used to locate regions in digital images where the brightness abruptly changes. The objective is to identify object boundaries within photographs.

3.3.2.1 *Sobel Edge Detection*

The image is processed in the x and y direction, then the magnitude of both x and y are combined to make a new image. Usually, the edge of an object will have the highest variation in pixel intensity value; finding this pixel difference can also draw the edge of an object.

$$G = \sqrt{G_x^2 + G_y^2}$$

4

G_x and G_y Gradient factor for x and y orientation. Edges are discovered using the Sobel approximation to the derivative in the Sobel technique of edge identification for image processing. At the locations where the gradient is most excellent, it comes before the edges. The Sobel method applies a 2-D spatial gradient quantity to a picture, emphasizing edge-corresponding high spatial frequency regions. Typically, it determines the expected actual gradient magnitude at every position in n input grayscale images [27].

3.3.2.2 *Canny edge detection*

Algorithm 1: Canny edge detection

Step 1: Apply the gaussian smoothing function to reduce noise

Step 2: Calculate the intensity gradient

Step 3: Find the non-maximum suppression

Step 4: Use Hysteresis to trace the edge.

To obtain a smooth image, convolute the image with a Gaussian function and apply the difference gradient operator first to determine edge strength, then compute edge magnitude and direction as usual and use critical or non-maximal suppression on the gradient's magnitude. To the image of non-maximal suppression, apply a threshold [28].

3.3.3 Clustering-Based Segmentation

In K-Mean segmentation, multiple segmentations on a given image can be performed and bring it for a classifier and try to find the boundaries that describe the location of an object [21,23].

Algorithm2: K-Means segmentation

Step 1: Find out the number ofclusters required for processing

Step 2: Calculate the centroid value

Step 3: Find the pixel distance to the centroid cluster

Step 4: Find the pixel based on the nearest value.

Step 5: Recalculate the centroid to produce a new centroid by averaging the pixel

Step 6: Allocate each data into a cluster based on the new centroid.

The K-means approach is composed of two distinct steps[38]. Phase 1 involves calculating the k centroids, and phase 2 involves assigning each point to the cluster whose centroid is closest to it. The Euclidean distance measure is one of the most widely used methods for determining the distance to the nearest centroid. After clustering is finished, the centroid for each cluster is recalculated. Based on that centroid, a new Euclidean distance between each center and each data point is calculated, and the cluster points with the least Euclidean distance are given the highest ranking. Each cluster in the partition is identified by its centroid and members (member objects). the location from which all of the items' combined distances are measured.

3.4 Classifier Selection

This section details different classifiers which are used for this work. This work compares the ensemble model with a Decision Tree, KNN, Logistic Regression, and Gaussian Naive Bayes classifier with Random Forest Classifier.

An ensemble classifier can create a new classifier using ensemble learning that works better than any component classifiers by starting with various basic classifiers. To forecast the output character based on an immense majority of votes, it simply sums the results of each classification model that was provided into the voting classifier. This work compares the ensemble model with a Decision Tree, KNN, Logistic Regression, and Gaussian Naive Bayes classifier with Random Forest Classifier. The class with the most significant number of votes or the class with the highest likelihood of being forecasted from each of the classifiers is the actual output class in hard voting. In soft voting, the forecast for each output class is based on the overall probability assigned to that class.

A Decision tree, many traditional machines learning methods, including Random Forests, Bagging, and Boosted Decision Trees, are built based on decision trees. A collection of decision trees is called Random Forest[34]. A class is given to each tree, and each tree "votes" for that class, allowing a new object to be classed according to its characteristics. The forest selects the classification with the most votes (over all the trees in the forest). KNN is a simple algorithm that keeps all of the current cases after sorting new instances with the approval of a minimal ‘k’ of its neighbors. The task is then handed to the class that shares the most significant similarities with the case. A distance function is employed for calculation. K-Means is an Unsupervised learning technique, and it is used to classify unlabelled data or data without clearly defined categories or groups. The method finds groups in the data, with the variable ‘K’ indicating how many groups are found. Based on the supplied characteristics, it repeatedly assigns each data point to one of the K groups. The Naive Bayesian [31] model is simple to construct and effective for large datasets. The Naive Bayes classification approach operates under the premise that the availability of each feature in a class is unrelated to the existence of any other feature. The logistic regression algorithm [32], one of the most basic machine learning techniques, operates on predictions made for z=1 as a function of the input.

4. Experiment And Result Analysis

This section offers experimental findings based on the proposed framework model. This paper used 3 categories of segmentations, namely threshold-based segmentation, edge-based segmentation, and cluster-based segmentation, and 5 Different Classifiers such as Random Forest, KNN, Naive Bayes, Logistic Regression, Decision tree, and Ensemble Classifier with soft and hard voting using our CCDDB, and attempts to determine which segmentation would work best with which classifiers.

Table 1. Performance Evaluation of different segmentation methods with Ensemble Classifier and Random Forest classifier on the CCDDB dataset

Segmentation	Metrics in %	Decision Tree	Naive Bayes	KNN	Logistic regression	Ensemble classifier (Hard Voting)	Ensemble classifier (Soft Voting)	Random Forest
Global threshold	Accuracy	68.8	55.5	73.2	71.4	69.1	70.0	**77.3**
	Precision	68.9	61.3	73.3	71.5	72.3	71.3	**77.4**
	Recall	68.8	55.5	73.2	71.4	69.1	70.0	**77.3**
	F1 score	68.0	31.0	72.3	70.6	61.8	65.7	**77.8**
Otsu threshold	Accuracy	75.2	75.5	78.8	74.1	79.4	79.1	**81.1**
	Precision	75.3	78.5	79	74.3	80.9	80.4	**81.8**
	Recall	75.2	75.5	78.8	74.1	79.4	79.1	**81.1**
	F1 score	75.1	70.8	77.9	72.6	76.8	76.7	**80.0**
Adaptive mean	Accuracy	63.8	67.3	71.7	68.2	71.7	74.4	**75.5**
	Precision	63.8	69.0	72.2	68.2	72.7	74.4	**75.8**
	Recall	63.8	67.3	71.7	68.2	71.7	74.4	**75.5**
	F1 score	63.5	61.5	73.6	68.4	68.4	73.8	**76.4**
Adaptive Gaussian	Accuracy	74.4	74.1	78.2	71.1	80.2	80.5	**85.0**
	Precision	74.4	74.1	78.5	71.2	80.7	80.6	**85.2**
	Recall	74.4	74.1	78.2	71.1	80.2	80.5	**85.0**
	F1 score	74.0	73.4	79.3	70.6	79.0	80.3	**84.4**
Canny Edge	Accuracy	56.1	66.7	68.8	66.4	67.9	70.2	**77.0**
	Precision	56.1	67.7	69.2	66.4	68.4	70.4	**77.0**
	Recall	56.1	66.7	68.8	66.4	67.9	70.2	**77.0**
	F1 score	56.0	62.4	71.0	66.8	64.9	69.1	**76.7**
Sobel Edge	Accuracy	68.8	71.4	66.1	63.2	73.2	76.4	**89.7**

	Precision	68.8	71.6	68.7	63.2	74.0	76.4	**89.8**
	Recall	68.8	71.4	66.1	63.2	73.2	76.4	**89.7**
	F1 score	68.8	69.9	71.4	64.3	70.5	76.7	**89.4**
K-means	Accuracy	87.6	74.7	84.7	75.5	85.0	86.7	**96.7**
	Precision	87.7	74.8	84.9	75.5	85.5	86.7	**96.7**
	Recall	87.6	74.7	84.7	75.5	85.0	86.7	**96.7**
	F1 score	87.4	73.6	85.3	75.5	84.0	86.7	**96.7**

Table 2. Performance Evaluation of different segmentation methods with Ensemble Classifier and Random Forest classifier on the PDDB dataset

Segmentation	Metrics in %	Decision Tree	Naive Bayes	KNN	Logistic regression	Ensemble classifier (Hard Voting)	Ensemble classifier (Soft Voting)	Random Forest
Global threshold	Accuracy	92.3	91.4	96.4	96.4	95.5	94.9	**97.6**
	Precision	92.3	91.4	96.6	96.4	95.5	95.0	**97.6**
	Recall	92.3	91.4	96.4	96.4	95.5	94.9	**97.6**
	F1-Score	92.2	91.5	96.5	96.4	95.5	95.0	**97.6**
Otsu threshold	Accuracy	84.3	71.3	78.4	72.2	84.6	82.8	**88.2**
	Precision	85.2	72.1	82.3	72.5	84.7	84.6	**89.3**
	Recall	84.3	71.3	78.4	72.2	84.6	82.8	**88.2**
	F1-Score	85.4	73.8	81.7	73.7	84.9	84.6	**89.0**
Adaptive mean	Accuracy	82.8	96.4	94.6	94.9	96.1	97.3	**97.9**
	Precision	83.2	96.4	94.7	95.1	96.3	97.3	**98.0**
	Recall	82.8	96.4	94.6	94.9	96.1	97.3	**97.9**
	F1-Score	83.6	96.4	94.5	95.1	96.0	97.3	**97.9**
Adaptive Gaussian	Accuracy	86.4	96.1	94.1	98.5	97.0	97.9	**98.8**
	Precision	86.7	96.1	94.7	98.5	97.1	97.9	**98.8**
	Recall	86.4	96.1	94.1	98.5	97.0	97.9	**98.8**
	F1-Score	87.0	96.1	93.7	98.5	96.9	97.9	**98.8**
Canny Edge	Accuracy	86.1	80.5	51.0	83.7	84.3	84.9	**91.4**
	Precision	86.3	81.1	54.1	83.8	85.0	84.9	**92.2**
	Recall	86.1	80.5	51.0	83.7	84.3	84.9	**91.4**
	F1-Score	86.6	81.8	15.3	84.2	83.2	85.0	**91.9**
Sobel Edge	Accuracy	82.6	79.9	64.3	57.8	77.5	83.7	**99.4**

	Precision	82.7	80.1	74.4	57.8	82.8	85.2	**99.4**
	Recall	82.6	79.9	64.3	57.8	77.5	83.7	**99.4**
	F1-Score	83.0	80.7	47.6	56.8	72.0	81.9	**99.4**
K-means	Accuracy	98.8	97.9	99.4	99.4	98.8	99.4	**99.7**
	Precision	98.8	98.0	99.4	99.4	98.8	99.4	**99.7**
	Recall	98.8	97.9	99.4	99.4	98.8	99.4	**99.7**
	F1-Score	98.8	97.9	99.4	99.4	98.8	99.4	**99.7**

From Table 1 and Table 2, one thing that can be observed is that K-Means Segmentation with Random Classifier outperforms other classifiers for the CCDDB dataset; the accuracy obtained is 96.76%. The same model is compared and applied to the PDDB dataset where the k-means segmentation with the Random Forest classifier outperforms with an accuracy of 99.7%.

In Table 3. It has been shown that sample input and the segmented leaf of different segmentation techniques such as Global Threshold, Otsu, Adaptive Mean, Adaptive Gaussian, Canny Edge Detection, Sobel Edge Detection, and K-means clustering segmentation applied to the CCDDB and PDDB datasets.

In Table 4, it has been shown that the confusion matrix of K-Means Segmentation applied to different classifiers on CCDDB. Since K-Means segmentation outperforms all other segmentation, the confusion matrix possessed by K-Means is given here. In Fig 4.1 (g), according to the Random Forest classifier, the actual positive and true negative values are 48.8% and 47.9%, respectively.

In Table 5, it has been shown that the confusion matrix of K-Means Segmentation applied to different classifiers on PDDB. Since K-Means segmentation outperforms all other segmentation, the confusion matrix of K-Means is given here. In Fig 4.2 (g), according to the Random Forest classifier, the true positive and actual negative values are 49.56% and 50.15%, respectively.

Fig 4.3(a) shows the performance analysis of Global Threshold Segmentation with the random forest classifier outperforming other classifiers. In Global Threshold Segmentation, threshold value 110 is used. The Random Forest classifier's settings are adjusted to improve accuracy when the Global Threshold Segmentation image is given as input; for example, the minimum number of samples to split the internal node of our random forest is set to 12. The minimum number of samples in the leaf node is set to 12. The maximum number of features to be taken is set to log2 and is used to obtain a better accuracy value of 77.3.

Fig 4.3(b) shows the performance analysis of Adaptive Mean Segmentation with the random forest classifier outperforming other classifiers. In Adaptive Mean Segmentation, block size 5 represents the size of the pixel neighborhood, and the constant value is set to 5. The Random Forest classifier's settings are adjusted to improve accuracy when the Adaptive Mean Segmentation image is given as input, and the minimum number of samples to split the internal node of our random forest is set to 12. The minimum number of samples in the leaf node is set to 12. The maximum number of features The Random Forest classifier's settings are adjusted to improve accuracy when the Adaptive Mean Segmentation image is given as input, to be taken is set to log2 is used to obtain the better accuracy value of 75.5%.

Fig 4.3(c) shows the performance analysis of the Adaptive Gaussian Segmentation method with the random forest classifier outperforming other classifiers. In Adaptive Gaussian Segmentation, the block size is taken as 11, and the

constant value is set to 5. for example, the minimum number of samples to split the internal node of our random forest is set to 12. The minimum number of samples in the leaf node is set to 12. The maximum number of features to be taken is set to log2 and is used to obtain a better accuracy value of 85.0%.

Fig 4.3(d) shows the performance analysis of the Otsu Segmentation method with the random forest classifier outperforming other classifiers of Otsu Segmentation. The threshold value used to classify the pixel is 127, and the obtained threshold is 135. The Random Forest classifier's settings are adjusted to improve accuracy when the Otsu Segmentation image is given as input; for example, the minimum number of samples to split the internal node of our random forest is set to 12. The minimum number of samples in the leaf node is set to 12. The maximum number of features to be taken is set to log2 is used to obtain a better accuracy value of 81.1%.

Fig 4.3(e) shows the performance analysis of the Sobel Edge Detection method with the random forest classifier outperforming other classifiers. In the Sobel Edge Detection Segmentation method, the x derivative is taken as one, and the depth of the image is set to -1. The Random Forest classifier's settings are adjusted to improve accuracy when the Sobel Edge Detection image is given as input; for example, the minimum number of samples to split the internal node of our random forest is set to 12. The minimum number of samples in the leaf node is set to 12. The maximum number of features to be taken is set to log2 and is used to obtain a better accuracy value of 89.7%.

Fig 4.3(f) shows the performance analysis of the Canny Edge Detection method with the random forest classifier outperforming other classifiers. In Canny Edge Detection, this work is carried out by 10 and 200 for the high and low-intensity gradients. A gaussian kernel can make noise reduction. Here the width and height of the kernel are taken as 5 by 5. The sigma value is set to 0; thus, it calculates the standard deviation automatically. Here all the pixel values smaller than the lower threshold value of 10 are set to 0, and more significant than the higher threshold value of 200 are set to 1. The Random Forest classifier's settings are adjusted to improve accuracy when the Canny Edge Detection image is given as input; for example, the minimum number of samples to split the internal node of our random forest is set to 12. The minimum number of samples in the leaf node is set to 12. The maximum number of features to be taken is set to log2 and is used to obtain a better accuracy value of 77.0%.

Fig 4.3(g) shows the performance analysis of the K-Means Segmentation method with the random forest classifier outperforming other classifiers. In K-Means Segmentation, The value of k is taken as nine, and number of times the algorithm execution needs to be executed by a different initial value specified as 10. When the kmeans_PP_centers flag is enabled, the method iterates the entire image to find the (probable points) possible centers before beginning to converge. The Random Forest classifier's settings are adjusted to improve accuracy when the K-Means Segmented image is given as input; for example, the minimum number of samples to split the internal node of our random forest is set to 12. The minimum number of samples in the leaf node is set to 12. the maximum number of features to be taken is set to log2 to obtain a better accuracy value of 96.7%.

In Fig 4.3(h) and Fig 4.3(i), it has been shown that the overall accuracy of different classifiers based on segmentation methods such as Global Threshold, Otsu, Adaptive Mean, Adaptive Gaussian, Canny Edge Detection, Sobel Edge Detection, and K-means clustering segmentation. In which K-means with random forest classifier outperforms other segmentation methods and other classifiers for the CCDDB dataset. K-means with Random Forest outperforms other segmentation methods on the PDDB dataset. For CCDDB, K-Means with Random Forest has 96.7% accuracy, shown in Fig 4.3(h) Red color, and for PDDB, Random Forest obtained a better accuracy of 99.7% accuracy shown in red, blue, and green color.

5. Conclusion And Future Work

This work has analyzed different methods for cashew leaf image segmentation and classifications. The Machine Learning models such as Random Forest, Decision Tree, Gaussian Naive Bayes, KNN, Logistic Regression, and Ensemble Classifiers with Hard and Soft Voting models are being used. In the result analysis, It has been shown that the comparison chart for each of the classifiers and their performance measure. Based on the work carried out, the conclusion is that the K-Means segmentation with Random Forest Classifier outperforms other classifiers with an accuracy of 96.7% on the CCDDB dataset. K-means with Random Forest outperformed other segmentation methods on the PDDB dataset and obtained 99.7% accuracy. In the future, the plan is to incorporate multiclass classifiers and feature extraction techniques to classify cashew leaf plant diseases.

Declarations

Acknowledgment:

Dr. Jeyalakshmi C, Professor, Plant Pathology, Pandit Jawaharlal Nehru College of Agriculture and Research, Karaikal-609603, UT of Puducherry, India. Validated our Cashew Crop Diseased DataBase (CCDDB).

Dataset Availability Statement

The dataset used in this article is available online and the link is provided in the reference sections [37].

Conflict of Interest Statement

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version. This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue. The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

References

1. Kumar, M., Gupta, S., Gao, X. Z., & Singh, A. (2019). Plant species recognition using morphological features and adaptive boosting methodology. IEEE Access, 7, 163912-163918.
2. Sharma, P., Hans, P., & Gupta, S. C. (2020, January). Classification of plant leaf diseases using machine learning and image preprocessing techniques. In 2020 10th international conference on cloud computing, data science & engineering (Confluence) (pp. 480-484). IEEE.
3. Srivastava, A. R., & Venkatesan, M. (2020). Tea leaf disease prediction using texture-based image processing. In Emerging Research in Data Engineering Systems and Computer Communications (pp. 17-25). Springer, Singapore.
4. Mardhiyah, A., & Harjoko, A. (2011). Metode segmentasi paru-paru dan jantung pada citra X-ray thorax. IJEIS, 1(2), 35-44.
5. Ganatra, N., & Patel, A. (2020). A multiclass plant leaf disease detection using image processing and machine learning techniques. Int. J. Emerg. Technol, 11(2), 1082-1086.
6. Kartikeyan, P., & Shrivastava, G. (2021). Review on emerging trends in detection of plant diseases using image processing with machine learning. International Journal of Computer Application, 975, 8887.
7. Youwen, T., Tianlai, L., & Yan, N. (2008, May). The recognition of cucumber disease based on image processing and support vector machine. In 2008 congress on image and signal processing (Vol. 2, pp. 262-267). IEEE.

8. Arivazhagan, S., Shebiah, R. N., Ananthi, S., & Varthini, S. V. (2013). Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features. Agricultural Engineering International: CIGR Journal, 15(1), 211-217.
9. H. Al-Hiary, S. Bani-Ahmad, M. Reyalat, M. Braik and Z. AlRahamneh, “Fast and accurate detection and classification of plant diseases”, International journal of computer applications (0975-8887), 2011, vol.17- no.1, pp-31-38.
10. Sannakki, S. S., Rajpurohit, V. S., Nargund, V. B., & Kulkarni, P. (2013, July). Diagnosis and classification of grape leaf diseases using neural networks. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.
11. Hossain, E., Hossain, M. F., & Rahaman, M. A. (2019, February). A color and texture based approach for the detection and classification of plant leaf disease using KNN classifier. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE.
12. Samajpati, B. J., & Degadwala, S. D. (2016, April). Hybrid approach for apple fruit diseases detection and classification using random forest classifier. In 2016 International conference on communication and signal processing (ICCSP) (pp. 1015-1019). IEEE.
13. Pang, J., Bai, Z. Y., Lai, J. C., & Li, S. K. (2011, October). Automatic segmentation of crop leaf spot disease images by integrating local threshold and seeded region growing. In 2011 international conference on image analysis and signal processing (pp. 590-594). IEEE.
14. Ramesh, G., Albert, D. W., & Ramu, G. (2020). Detection of plant diseases by analyzing the texture of leaf using ANN classifier. Int J Adv Sci Technol, 29(8s), 1656-1664.
15. Kaur, N. (2021). Plant leaf disease detection using ensemble classification and feature extraction. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 2339-2352.
16. Jayamoorthy, S., & Palanivel, N. (2017). Identification of Leaf Disease Using Fuzzy C-MEAN and Kernal Fuzzy C-MEAN and Suggesting the Pesticides. International Journal of Advanced Research in Science, Engineering and Technology, 4(5), 3852-3855.
17. Zamani, A. S., Anand, L., Rane, K. P., Prabhu, P., Buttar, A. M., Pallathadka, H., ... & Dugbakie, B. N. (2022). Performance of Machine Learning and Image Processing in Plant Leaf Disease Detection. Journal of Food Quality, 2022.
18. Yang, X., Ni, H., Li, J., Lv, J., Mu, H., & Qi, D. (2022). Leaf recognition using BP-RBF hybrid neural network. Journal of Forestry Research, 33(2), 579-589.
19. Badiger, M., Kumara, V., Shetty, S. C., & Poojary, S. (2022). Leaf and Skin Disease Detection using Image Processing. Global Transitions Proceedings.
20. Ansari, A. S., Jawarneh, M., Ritonga, M., Jamwal, P., Mohammadi, M. S., Veluri, R. K., ... & Shah, M. A. (2022). Improved Support Vector Machine and Image Processing Enabled Methodology for Detection and Classification of Grape Leaf Disease. Journal of Food Quality, 2022.
21. Trivedi, V. K., Shukla, P. K., & Pandey, A. (2022). Automatic segmentation of plant leaves disease using min-max hue histogram and k-mean clustering. Multimedia Tools and Applications, 1-28.
22. Singh, V. (2019). Sunflower leaf diseases detection using image segmentation based on particle swarm optimization. Artificial Intelligence in Agriculture, 3, 62-68.
23. Archana, K. S., & Sahayadhas, A. (2018). Automatic rice leaf disease segmentation using image processing techniques. Int. J. Eng. Technol, 7(3.27), 182-185.

24. Ireri, D., Belal, E., Okinda, C., Makange, N., & Ji, C. (2019). A computer vision system for defect discrimination and grading in tomatoes using machine learning and image processing. Artificial Intelligence in Agriculture, 2, 28-37.
25. Singh, V., & Misra, A. K. (2017). Detection of plant leaf diseases using image segmentation and soft computing techniques. Information processing in Agriculture, 4(1), 41-49.
26. Senthilkumaran, N., & Vaithegi, S. (2016). Image segmentation by using thresholding techniques for medical images. Computer Science & Engineering: An International Journal, 6(1), 1-13.
27. Muthukrishnan, R., & Radha, M. (2011). Edge detection techniques for image segmentation. International Journal of Computer Science & Information Technology, 3(6), 259.
28. Al-Amri, S. S., Kalyankar, N. V., & Khamitkar, S. D. (2010). Image segmentation by using edge detection. International journal on computer science and engineering, 2(3), 804-807.
29. Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science, 54, 764-771.
30. Iqbal, Z., Khan, M. A., Sharif, M., Shah, J. H., ur Rehman, M. H., & Javed, K. (2018). An automated detection and classification of citrus plant diseases using image processing techniques: A review. Computers and electronics in agriculture, 153, 12-32.
31. Chaudhary, A., Thakur, R., Kolhe, S., & Kamal, R. (2020). A particle swarm optimization based ensemble for vegetable crop disease recognition. Computers and Electronics in Agriculture, 178, 105747.
32. Chaudhary, A., Kolhe, S., & Kamal, R. (2016). A hybrid ensemble for classification in multiclass datasets: An application to oilseed disease dataset. Computers and Electronics in Agriculture, 124, 65-72.
33. Prabu, M., & Chelliah, B. J. (2022). An intelligent approach using boosted support vector machine based arithmetic optimization algorithm for accurate detection of plant leaf disease. *Pattern Analysis and Applications*, 1-13.
34. Bendib, M. M., Merouani, H. F., & Diaba, F. (2015). Automatic segmentation of brain MRI through stationary wavelet transform and random forests. *Pattern Analysis and Applications*, *18*(4), 829-843.
35. Khatoon, A., Mohapatra, A., & Kunja, S. (2017). Major diseases of cashew (Anacardium Occidentale L.) Caused by fungi and their control in Odisha, India. International Journal of Biosciences, 11(1), 68-74.
36. Wonni, I., Sereme, D., Ouédraogo, I., Kassankagno, A., Dao, I., Ouedraogo, L., & Nacro, S. (2017). Diseases of cashew nut plants (Anacardium Occidentale L.) in Burkina Faso. Advances in Plants Agriculture Research, 6(3), 78-83.
37. Barbedo, J. G. A., Koenigkan, L. V., Halfeld-Vieira, B. A., Costa, R. V., Nechet, K. L., Godoy, C. V., & Oliveira, S. A. S. (2018). Annotated Plant Pathology Databases for Image-Based Detection and Recognition of Diseases. IEEE Latin America Transactions, 16(6), 1749-1757.
38. El Massi, I., Es-saady, Y., El Yassa, M., & Mammass, D. (2021). Combination of multiple classifiers for automatic recognition of diseases and damages on plant leaves. *Signal, Image and Video Processing*, *15*(4), 789-796.

Tables

Tables 3 to 5 are available in the Supplementary Files section.

Figures

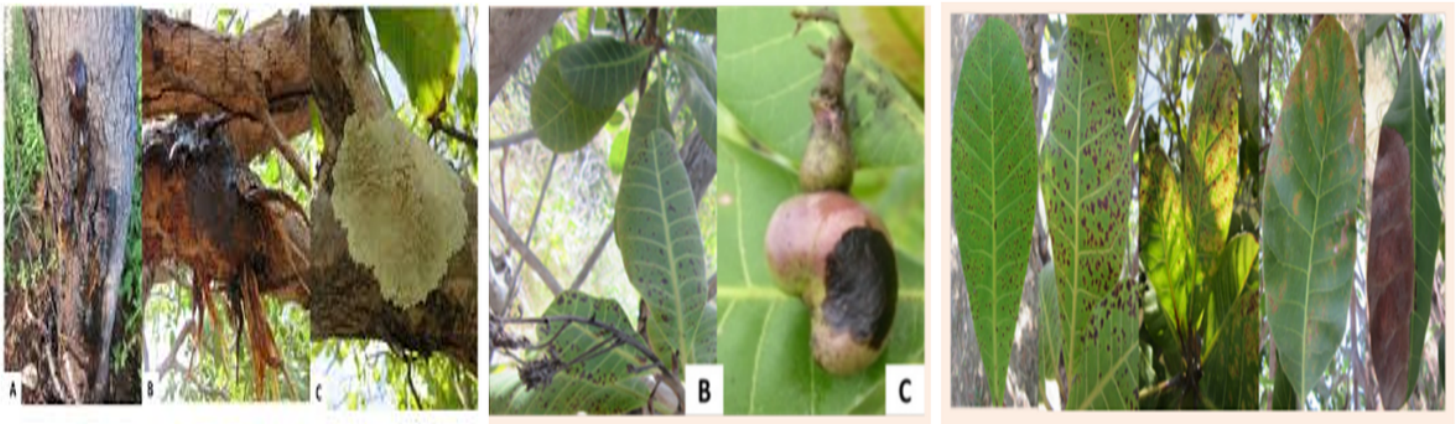


Figure 1

(a) Gummosis

(b) Symptoms of Anthracnose

(c) Foliar symptoms

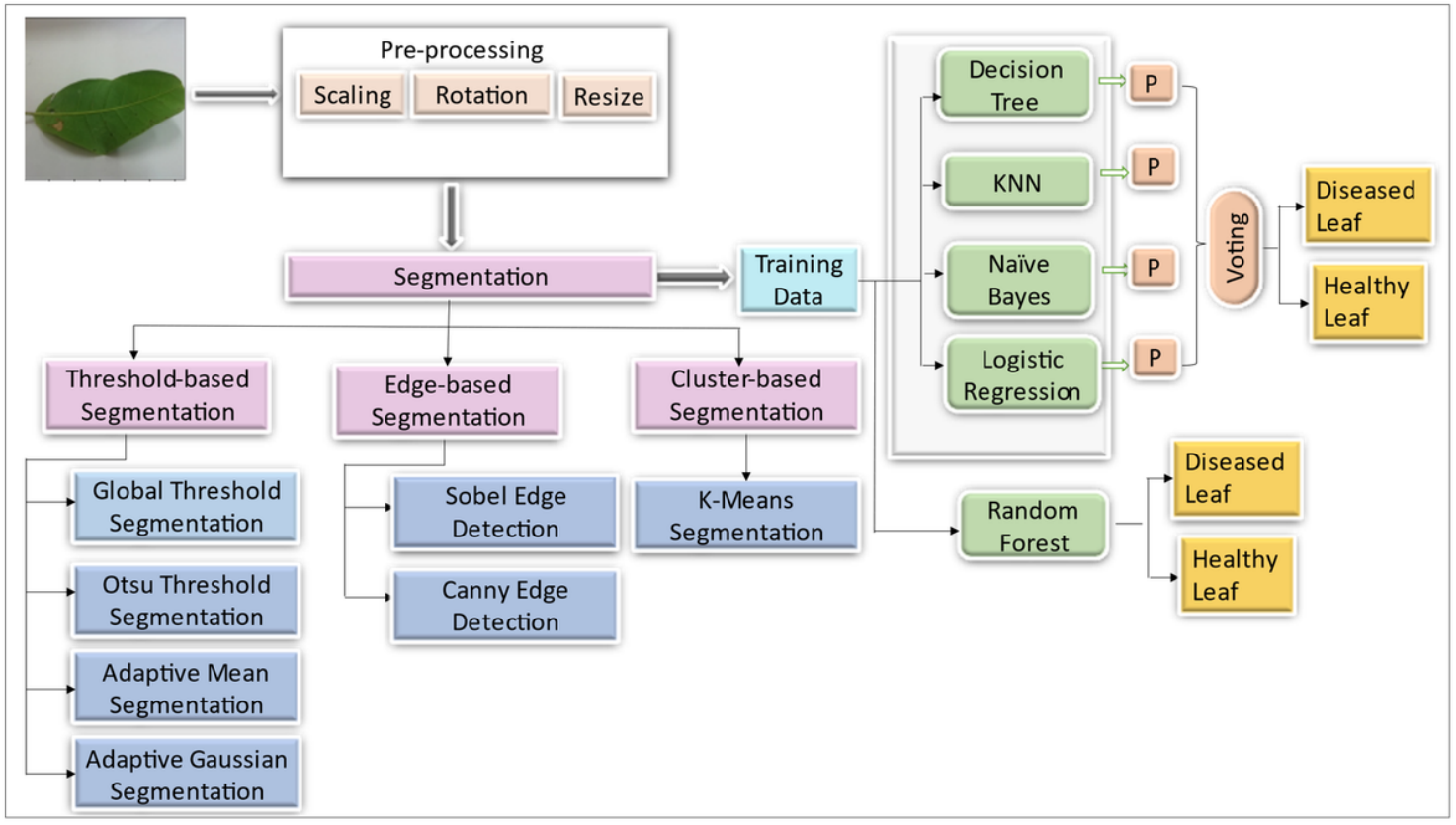


Figure 2

Proposed model for cashew leaf disease detection

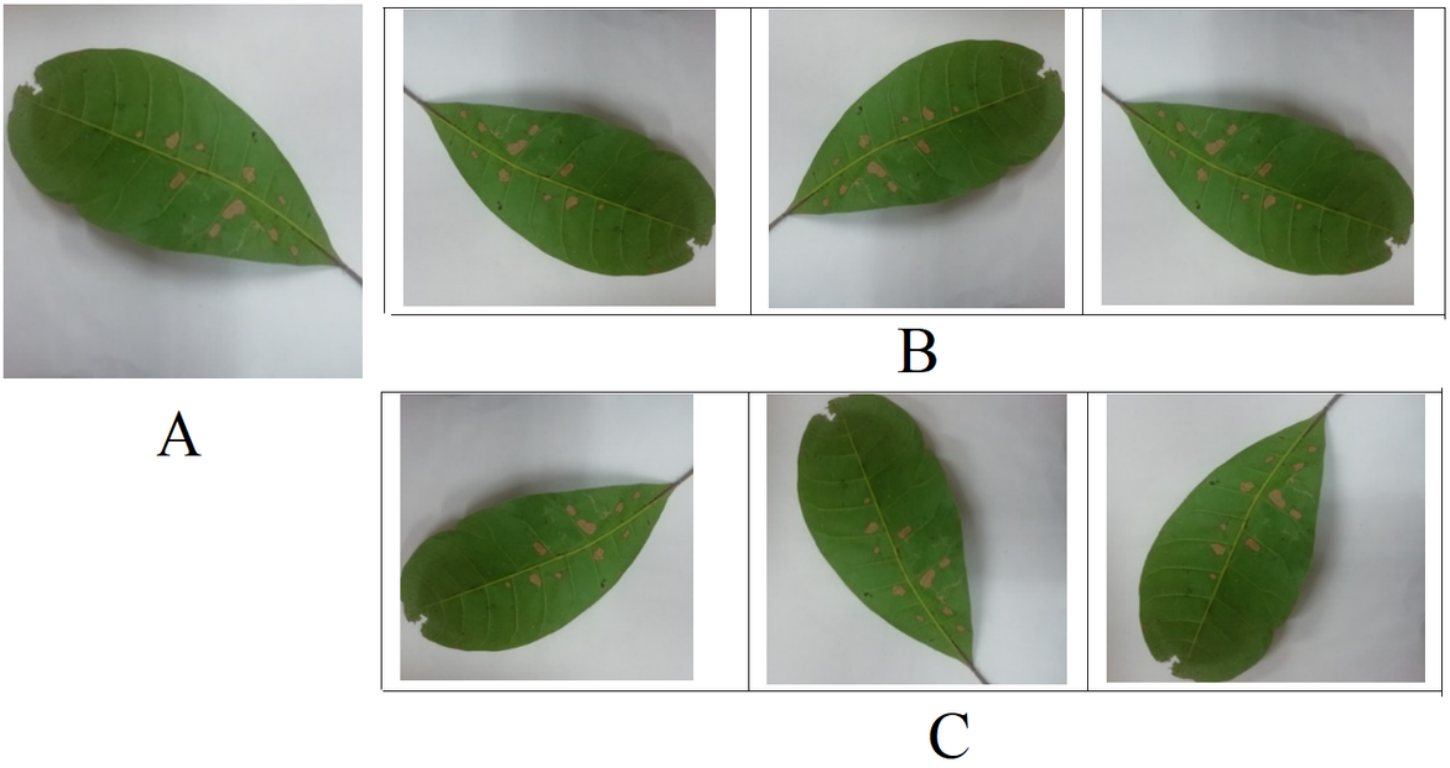


Figure 3

(a) Sample input image from the dataset

(b) Image after flipping

(c) Image after rotation

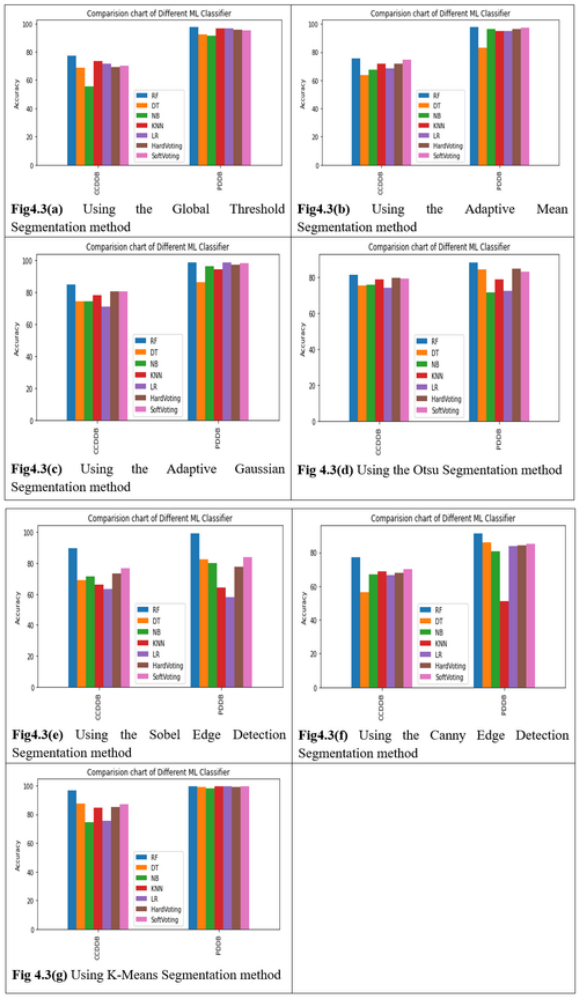


Fig4.3(a) Using the Global Threshold Segmentation method

Fig4.3(b) Using the Adaptive Mean Segmentation method

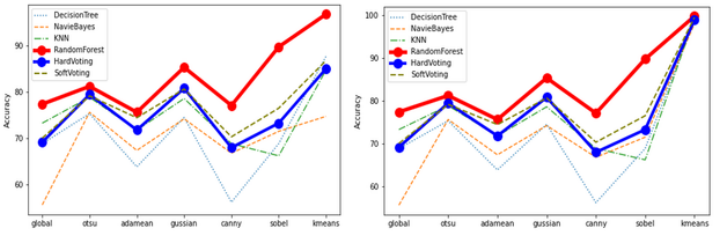
Fig4.3(c) Using the Adaptive Gaussian Segmentation method

Fig 4.3(d) Using the Otsu Segmentation method

Fig4.3(e) Using the Sobel Edge Detection Segmentation method

Fig4.3(f) Using the Canny Edge Detection Segmentation method

Fig 4.3(g) Using K-Means Segmentation method



ig 4.3(h). Accuracy Grapl different segmentation method CDDB

Fig4.3(i). Accuracy Graph for different segmentation methods on PDDB

Figure 4

Performance evaluation of different segmentation methods using an Ensemble Classifier and Random Forest model applied to CCDDB and PDDB datasets.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tab3.docx
- Tab4.docx

- Tab5.docx