RESEARCH ARTICLE OPEN ACCESS

Identification and Classification of Leaf Diseases in Turmeric Plants

Nandhini M, Pream Sudha V, Vijaya MS
M.Phil Research Scholar PSGR Krishnammal college for women Coimbatore, India
Assistant Professor Department of Computer Science PSGR Krishnammal College for women Coimbatore, India
Associate Professor Department of Computer Science PSGR Krishnammal College for women Coimbatore, India

Abstract
Plant disease identification is the most important sector in agriculture. Turmeric is one of the important rhizomatous crops grown in India. The turmeric leaf is highly exposed to diseases like rhizome rot, leaf spot, and leaf blotch. The identification of plant diseases requires close monitoring and hence this paper adopts technologies to manage turmeric plant diseases caused by fungi to enable production of high quality crop yields. Various image processing and machine learning techniques are used to identify and classify the diseases in turmeric leaf. The dataset with 800 leaf images of different categories were pre-processed and segmented to promote efficient feature extraction. Machine learning algorithms like support vector machine, decision tree and naïve bayes were applied to train the model. The performance of the model was evaluated using 10 fold cross validation and the results are reported.
Keywords: Machine Learning; Segmentation; Leaf disease; Turmeric.

I. INTRODUCTION

Agriculture is one of the emerging topics in data mining.Research in agriculture is aimed towards increase of productivity and food quality at reduced expenditure and with increased profit. In the past few years new trends have emerged in the Indian agricultural sector and technological advancements are gradually finding their place.Turmeric is the dried rhizome of Curcuma longa, aherbaceous perennial belonging to the family Zingiberaceae. Curcumin, the most biologically active phytochemical compound is available upto 3% in Turmeric. Indian turmeric is considered as the best in the world due to its high curcumin content. It is extracted and researched for its well-known range of therapeutic effects.

The idea of integrating Information and Communications Technology (ICT) with agriculture sector motivates the development of an automated system for turmeric disease classification. The various diseases that affect turmeric leaf are categorized as leaf spot, leaf blotch and rhizome rot. The cause and symptoms for each of these diseases are described below. Leaf spot of turmeric is the most important disease of turmeric. It has become a major constraint in successful cultivation of turmeric. Symptom appears as brown spots of various sizes on the upper surface of the young leaves.It is caused by the fungi Colletotrichumcapsici. Leaf blotch disease symptom appears as small, oval, rectangular or irregular brown spots on either side of the leaves which soon become dirty yellow or dark brown. The causal organism is the fungiTaphrinamaculans. Foliar symptoms due to rhizome rot appear as light yellowing of the tips of lower leaves which gradually spreads to the leaf blades. The fungiPythiumgraminicolum is the agent causing this disease.

Meenakshi M. Pawar et al., [1], approaches an automatic grading and sorting system for pomegranate. Color texture feature analysis was used for detection of surface defects on pomegranates. Best features were used as an input to Support Vector Machine (SVM) classifier and tests were performed to identify best classification model.K.R.Wang and Shaokun Li [2] created a model of cotton leaf chlorophyll determination based on using the machine vision technology for the colour features of cotton leaf. The research showed that the BIR values of ROB color system values of chromaticity coordinate and the S values of HIS color system were all significantly correlated with chlorophyll content of cotton leaf. These values could be used to determine the concentration of chlorophyll.

Yan Cheng Zhang, et al., [3] tries to identify and diagnose cotton disease using computer vision. They developed the fuzzy feature selection approach, fuzzy curves (FC) and surfaces (FS) to select features of cotton diseased leaves image. They showed that the effectiveness of features selected by the FC and FS method was much better than that selected by human randomly or other methods.NitinP.Kumbhar et al., [4], explained texture statistics for detecting the plant leaf disease by color transformation structure. RGB is converted into HSV space because HSV was a good color descriptor. Masking and removing of green pixels with pre-computed threshold level was done. Segmentation was performed and they obtained useful segments. These segments were used for texture analysis by color co-occurrence matrix. Finally SGDM texture parameters were compared to texture parameters of normal leaf.

PradnyaRavindraNarvekar et al., [5], proposed a system to discuss the effective way used in performing detection of grape diseases through leaf feature inspection. Leaf image was captured and proposed to determine the health status of each plant. The digital images were acquired from the environment using a digital camera. Then image-processing techniques were applied to the acquired images to extract useful features that were necessary for further analysis. After that, several analytical discriminating techniques were used to classify the images according to the specific problem.

From the above literature survey a clear outlook about the techniques and methodologies followed in the existing works for detection of plant diseases has been obtained. This assists developing a proposed work of disease classification in turmeric with better performance.

II. PROPOSED WORK

The proposed work focuses on classifying the diseases in the turmeric leaf. Initially, the leaf images are collected for three different diseases namely Leaf spot, Leaf blotch, Rhizome rot and stored in a database for further processing. The database consists of 800 leaf images with 200 diseased leaf images for each disease category and 200 images for normal category. The proposed system includes different phases as follows.In the pre-processing phase, the given image is resized and converted into HSI images. The hue component is alone taken for further processing. In segmentation phase, K-means segmentation technique is used to segment the diseased portion from the original image. The segmentation result is processed and feature vectors are generated including color, texture and shape features. A subset of best features is selected from the feature set for accurate classification using ranking method. The overview of proposed framework is shown in Fig.1.

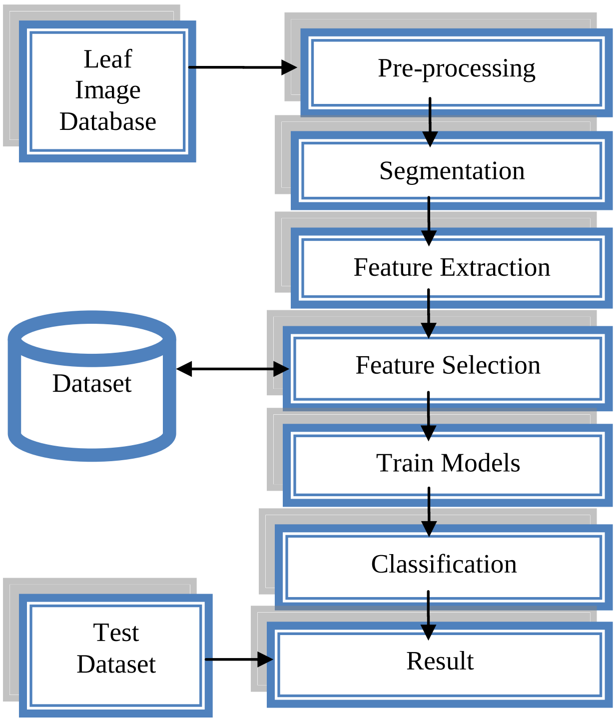

Fig. 1. *Overview of proppoosed system*.

A. Pre-Processing

The Turmeric leaf images are collected manually from Annur, Thondamuthur and Sulur. The images are taken using high resolution camera of 12 mega pixels. The collected images are of different dimensions hence it is essential to convert them to uniform size for efficient pre-processing. The images are resized to 256*256 dimensions using interpolation method.

- *RGB to HSI*

RGB images are converted into Hue Saturation Intensity (HSI) color` space representation. RGB is ideal for color generation. But HSI model is an ideal tool for color perception. Hue is a color attribute that describes pure color as perceived by an observer. Saturation refers to the relative purity or the amount of white light added to hue and intensity means amplitude of light. After the transformation process, saturation and intensity are dropped since it does not give extra information the hue component is taken for further analysis.The output of HSI conversion is shown in Fig.2.

Fig. 2. *Results of HIS conversion*.

B. Segmentation

Image segmentation is an important aspect of digital image processing. It may be defined as a process of assigning pixels to homogenous and disjoint regions which form a partition of the image that share certain visual characteristics. The major goal of segmentation is to simplify or change the representation of an image into meaningful image that is more proper and easier to explore. Segmentation is essentially a collection of methods that allows spatial partitioning to the close parts of the image as objects. In this proposed work segmentation process is done using kmeans segmentation algorithm.K-Means segmentation algorithm classifies the input data points into multiple classes based on their inherent distance from each other. The algorithm assumes that the data features form a vector space and tries to find natural grouping in them. The segmented image is shown in Fig.3.

Fig. 3. *Results of segmented image*.

C. Feature Extraction

Drawing out specific features from the pre-processed images is called feature extraction.Feature extraction plays a vital role in data mining. It can be used to improve the classification effectiveness and computational efficiency. Feature extraction is carried out with all the pre-processed leaf images. Three types of features such as color, texture and shape features are extracted in this work in order to extract the prominent features of an image.

- *Color features*

Color features play a decisive role in the classification ofleaf image. A color image is a combination of some basic colours. Each individual pixel of a color image is broken down into red, green and blue values. The entire image is represented in three matrices, each corresponding to red, green and blue. The three matrices are arranged in sequential order, to create a three m by n by 3 matrixes. For an image which has a height of 5 pixels and width of 10 pixels the result is 5 by 10 by 3 matrixes for a true colorimages.The RGB color features such as meanR, meanG, meanB are extracted from the leaf images.

- *Texture features*

The texture features are extracted using Grey Level Co-occurrence Matrix (GLCM). A GLCM is a matrix where number of rows and columns is equal to number grey levels G in an image. It is defined over an image to be the distribution of co-occurring values in the given offset. It is a way of extracting second order statistical features.

1) Energy

$$energy(ene) = \sum_i \sum_j g_{ij}^{\;2}$$

This statistic is also called uniformity or angular second moment. It measures the textural uniformity that is pixel pair repetitions. It detects disorders in textures. Energy reaches a maximum value equal to one. High energy values occur when the gray level distribution has a constant or periodic form. Energy has a normalized range.

2) Entropy

$$entropy(ent) = -\sum_i \sum_j g_{ij} \log_2 g_{ij}$$

This statistic measures the disorder or complexity of an image. The entropy is larger when the image is not texturally uniform and many GLCM elements have very small values. Complex textures tend to have high entropy. Entropy is strongly, but inversely correlated to entropy.

3) Contrast

$$contrast(con) = \sum_i \sum_j (i-j)^2 g_{ij}$$

This statistic measures the spatial frequency of an image and difference moment of GLCM. It is the difference between the highest and the lowest values of a contiguous set of pixels. It measures the amount of local variations present in the image. A low contrast image presents GLCM concentration term around the principal diagonal and features low spatial frequencies.

4) Variance

$$Variance\ (var) = \sum_i \sum_j (i - \mu)^2 g_{ij} \text{ where } \mu \text{ is the mean of } g_{ij}$$

This statistic is a measure of heterogeneity and is strongly correlated to first order statistical variable such as standard deviation. Variance increases when the gray level values differ from their mean.

5) Homogeneity

$$homogeneity(hom) = \sum_i \sum_j \frac{1}{1+(i-j)^2} g_{ij}$$

Homogeneity weights values by the inverse of the Contrast, weight, with weights decreasing exponentially away from the diagonal.

6) Correlation

The correlation feature is a measure of gray tone linear dependencies in the image. GLCM Correlation is quite a different calculation from the other texture measures. It is independent of them and can often be used profitably in combination with another texture measure. It also has a more intuitive meaning to the actual calculated values: 0 is uncorrelated, 1 is perfectly correlated.

$$correlation(cor) = \frac{\sum_j \sum_j (ij) g_{ij} - \mu_x \mu_y}{\sigma_x \sigma_y} g_{ij}$$

whereμ_x,μ_y,σ_x and σ_y are the means and standard deviations of g_x and g

7) Sum Average

$$sumaverage(sa) = \sum_{i=2}^{2N_g} i g_{x+y}(i)$$

8) Maximum Probability

Maximum Probability: max(Pij)

9) Sum Entropy

$$sumentropy(se) = -\sum_{i=2}^{2N_g} i g_{x+y}(i) \log\{g_{x+y}(i)\}$$

10) Sum Variance

$$sumvariance(sv) = \sum_{i=2}^{2N_g} (i - sa)^2 g_{x+y}(i)$$

11) Difference variance

$$differencevariance = varianceof\, g_{x-y}$$

12) Difference Entropy

$$differenceentropy(se) = -\sum_{i=0}^{N_g-1} g_{x-y}(i) \log\{g_{x-y}(i)\}$$

13) Information Measures of Correlation 1 (IMC1)

$$IMC1 = \frac{HXY - HXY1}{\max\{HX, HY\}}$$

14) Information Measures of Correlation 2 (IMC2)

$$IMC2 = \sqrt{(1 - \exp[-2.0(HXY2 - HXY)]}$$

15) Cluster shade

$$Shade = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \{i + j - \mu_x - \mu_y\}^3 * P(i,j)$$

16) Cluster Prominence

$$Prom = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \{i + j - \mu_x - \mu_y\}^4 * P(i,j)$$

A GLCM is a matrix where the number of rows and columns is equal to the number of gray level G in the image, where

$$\mu_x = \sum_{i=0}^{G-1} i P_x(i)$$

$$\mu_y = \sum_{j=0}^{G-1} i P_y(j)$$

17) Dissimilarity

$$Diss = \sum P_{i,j} * |i - j|$$

Dissimilarity is a measure that defines the variation of grey level pairs in an image. It is the closest to Contrast with a difference in the weight. Contrast unlike dissimilarity grows quadratically. It is expected that these two measures behave in the same way for the same texture because they calculate the same parameter with different weights. Contrast will always give slightly higher values than dissimilarity. Dissimilarity ranges from [0,1] and obtain maximum when the grey level of the reference and neighbour pixel is at the extremes of the possible grey levels in the texture sample.

- ***Shape Features***

The digital morphology features (DMF) generally include geometrical features (GF) and invariable moment features (MF). The geometrical features consist of perimeter, solidity, circularity, eccentricity, extent, diameter, orientation etc.

1) Solidity

Solidity describes the extent to which the shape is convex or concave and it is defined by

$$S = \frac{r_i}{r_c}$$

Where r_i represents the radius of in circle of the ROI, and r_c the radius of the ex-circle of the ROI.

2) Eccentricity

Eccentricity is the measure of aspect ratio. It is the ratio of the length of major axis to the length of minor axis. It can be calculated by principal axes method or minimum boundingrectangle method. The eccentricity is defined as the ratio of the length of main inertia axis of the ROI E_A to the length of minor inertia axis of the ROI E_B

$$E = \frac{E_A}{E_B}$$

3) Perimeter

Perimeter gives the number of pixels on the contour of the suspicious region. This perimeter is used for the parametric boundary representation. If x1,x2,……..,xn is a boundary coordinate list, N is the number of pixels on the boundary, the region perimeter

4) Extent

Extent refers to the proportion of pixels in the bounding rectangular box of a region.

5) Centroid

Centroid is also called as the centre of mass; h is a mask of cluster c over image S(x, y). The co-ordinates (xc, yc) of the
Centroid are defined as

$$x_c = \frac{\sum_{xy} x * h(x,y)}{\text{Mass}}$$

$$y_c = \frac{\sum_{xy} y * h(x,y)}{\text{Mass}}$$

D. Feature Selection

Feature Selection or attribute selection is a process which enables the automatic search for the best subset of attributes in the dataset. Data sets for analysis may contain many attributes, which may be irrelevant to the mining task, or redundant.In this work information gain method is used to rank the features.

- Information gain

Info Gain Attribute Eval evaluates the worth of an attribute by measuring the information gain with respect to the class.

Info Gain (Class, Attribute) = H(Class) – H(Class | Attribute)
where H is the information entropy.
Based on the ranking result out of 33 features 25 features are selected as best features for classification of leaf disease.

III. MACHINE LEARNIING ALGORITHMS

Machine learning algorithms decision tree, Support vector machine and Naïve bayes were used for learning the classification model.

A. Support vector machine

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. A support vector machine constructs a hyper plane or set of hyper planes in a high-dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class called functional margin, since in general the larger the margin the lower the generalization error of the classifier. Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a *new* data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector of a list of p numbers, and one wants to know whether one can separate such points with a $p - 1$-dimensional hyper plane. This is called a linear classifier [13].

B. Decision tree

Decision tree learning is the process of learning decision trees from the labelled training examples. Decision tree classification algorithm generates the output as a binary tree like structure called a decision tree, where each non leaf node i.e., internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node or terminal node holds a class label. The topmost node in a tree is the root node. A decision tree model contains rules which are used to predict the target variable. The class label of a new instance is predicted by testing the attribute values of the instance against the decision tree. A path is traversed from the root to a leaf node, which gives the class label of that data. Decision trees can be easily converted into classification rules.

C. Naive bayes

The Naive Bayes Classifier technique is based on Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Naïve Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. The Naive-Bayes inducers compute conditional probabilities of the classes given the instance and pick the class with the highest posterior. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. A Naive Bayes classifier is a simple probabilistic classifier based on Bayes theorem with strong independence assumptions. General formulation is given by

$$Given\, classes\ \omega_j\ and\ dataset\ x$$

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

$$where$$

$$p(x) = \sum_j p(x \mid \omega_j) P(\omega_j)$$

IV. Experiments And Results

The classification models can be evaluated using various criteria such as accuracy, precision, recall and F-measure.

The formula for calculating accuracy is,

Accuracy= $\dfrac{TP + TN}{TP + TN + FP + FN}$

The formula for calculating Precision is,

Precision = $\dfrac{TP}{TP+FP}$

The formula for calculating Recall is,

Recall = $\dfrac{TP}{TP+FN}$

The formula for calculating F-measure is,

F-measure=2*(precision*recall)/ (precision+recall)

Where TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The comparative results indicate that Support Vector Machine based classification model yields a better performance when compared to other models. The performance of the classifiers are assessed using measures such as accuracy, precision, recall, f-measure and time taken to build the model are shown in Table I and Fig. 4.

TABLE.1 PREDICTIVE PERFORMANCE OF CLASSIFIERS

Classifier	DT	SVM	Naive bayes
Precision	0.8817	0.9395	0.9354
Recall	0.875	0.9437	0.725
F-measure	0.8783	0.9415	0.8168
Accuracy	87.50%	93.75%	90%

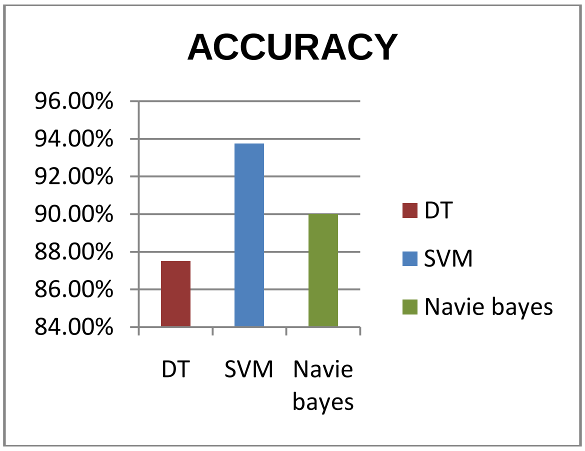

Fig. 4. *Comparison of accuracy*.

V. Conclusion

The research has modeled disease detection and classification for turmeric leaves. The work involves various image processing and machine learning techniques. The machine learning algorithms such as Support Vector Machine (SVM), Decision Tree (DT) and NB (Naive Bayes) were implemented using MATLAB platform. This analysis was verified by the result of 10-fold cross validation. The study shows that classification of turmeric leaf diseases using Support Vector Machine (SVM) gives better accuracy of 93.75% when compared to other algorithms. In future the research can be extended using different techniques for segmenting the diseased portion of the original leaf.

References

[1] Meenakshi M. Pawer, Sanman Bhusari, Akshay Gungewar (2012), "*Indentification of infected pomegranates using color texture feature analysis*", International journal of computer application(0975-8887).

[2] K.R.Wang, S.K.Li, CT.wang (2006), "Acquired chlorophyll concentration of cotton leaves with technology of machine vision", ActaAgronvo1.1, pp. 34-40, sin.

[3] Yan Cheng Zhang, et al.," *Features selection of cotton disease leaves image based on fuzzy feature selection techniques*", IEEE Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition, China, 2-4 Nov 2007.

[4] *Mr.Nitin, P.Kumbhar, Prof. Sanjay, B. Dhaygude. (2013), "Crop Disease Detection Using CBIR", International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS).Prof.*

[5] Pradnya Ravindra Narvekar, Mahesh Manik Kumbhar, S. N. Patil. (2014), *"Grape leaf disease detection & Analysis using SGDM Matix Method",* International Journal of

Innovative Research in Computer and Communication Engineering, Vol.2, Issue 3.

[6] Yeo Zhang, M.Z. Li (2005), "Nutrition information extraction of the cucumber leaves in the greenhouse based on computer vision technology",Transactions of the CSAE,vo1.21, pp.102-IOS.

[7] Sanjay, B. Dhaygude, Mr.Nitin, P.Kumbhar (2013), *"Agricultural plant Leaf Disease Detection using Image proessing"*, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering , Vol. 2, Issue 1.

[8] Patil J.K, Raj Kumar(2012), "*Feature Extraction of Diseased Leaf image",* Journal of Signal and Image Processing Volume 3, Issue 1.

[9] P.Revathi, M. Hemalatha. (2012), " Homogenous segmentation based edge detection techniques for proficient Identification of the Cotton leaf Spot Disease", International Journal of Computer Application Vol.47-No 2.

[10] Mokhled S, Al-Tarawneh (2013), "An Empirical Investigation of Olive Leave Spot Disease Using Auto-Cropping Segmentation and Fuzzy C-Means Classification", World Applied Science Journal 23 (9): 1207-1211.

[11] Anthony.G, wickramarachchf.N, "*An image recognition system for crop disease Identification of paddy fields in Sri Lanka*", Fourth International Conference on Industrial and Information System, ICIIS 2009,Sri Lanka.P 403-407, 2009.

[12] Moshou et al., "Automatic detection of yellow rust in wheat using reflectance measurements and neural networks", Computer and Electonics in agriculture, v.44, p.173-188, 2004.

[13] Soman K.P, Loganathan R, Ajay V (2009), "*Machine Learning with SVM and other Kernel Methods*", PHI, India.

[14] V.Premalatha, Dr. M. G. Sumithra, S. Deepak, P. Rajeswari (2014), *"Implementation of Spatial FCM for Leaf Image Segmentation in Pest Detection"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 10, ISSN: 2277 128X.

[15] Manimekalai.K, Vijaya.MS(2014), "*Support Vector Machine Based Tool For Plant Species Taxonomic Classification",* Journal of Asian Scientific Research 4(4):159-173**.**

[16] K. Muthukannan et al., (2015) "*Classification of diseased plant leaves using neural network algorithms",* ARPN Journal of Engineering and Applied Sciences vol. 10, no. 4.

[17] Zulkifli Bin Husin et al.,(2012), "*Feasibility Study on Plant Chili Disease Detection Using Image Processing Techniques"*, IEEE International Conference on Intelligent Systems Modelling and Simulation.