Feature Selection Techniques for the Classification of Leaf Diseases in Turmeric

Pream Sudha V[#]

[#]*Assistant Professor, Department of Computer Science, PSGR Krishnammal College for Women Coimbatore , India*

Abstract **-** Crop maintenance is one of the crucial factors that determine the quantity and quality of the agricultural products. Protecting crops from plant diseases is an important aspect that increases the profit of the farmer. This study aims at developing a computational model that will facilitate crop production by accurately identifying diseases that affect productivity of turmeric plants. The turmeric leaf is highly exposed to diseases like rhizome rot, leaf spot, and leaf blotch. This system uses technologies such as feature selection and machine learning techniques for the identification and classification of diseases in turmeric leaf. Principal component analysis, Information gain and Relief-f attribute evaluator methods were investigated in combination with machine learning algorithms like Support Vector Machine, Decision Tree and Naïve Bayes. The performance of the models were evaluated using 10 fold cross validation and the results were reported. Comparatively, the model using SVM applied to features selected using Information gain performed well with an accuracy of 93.75.

Keywords — *Leaf disease, Turmeric, Machine Learning, Feature Selection.*

I. INTRODUCTION

Plant disease is one of the important factor which causes significant reduction in the quality and quantity of plant production. Detection and classification of plant diseases are important tasks to increase plant productivity and economic growth. Various plant diseases pose a great threat to the agricultural sector by reducing the life of the plants and relying on pure naked-eye observation to detect and classify diseases can be expensive. The diagnosis of leaf diseases may cause confusion due to the similarities in the shape, size and colour, which require an expert assessment. The first step in fighting against leaf diseases is the adequate recognition of their presence. The use of computers in agriculture has been the subject of several scientific works, many of them focusing on the identification of diseases through foliar symptoms in various cultivars such as, wheat, cotton, rice, cucumber, rose, rubber tree and grape. The idea of integrating Information and Communications Technology with agriculture sector motivates the development of an automated system for turmeric disease classification.

Turmeric is one of the important rhizomatous crops grown in India. The management of perennial crops like turmeric requires close monitoring especially for the management of diseases that can affect production considerably. Turmeric is the dried rhizome of Curcuma longa L., a herbaceous 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. India is the leading producer of Turmeric with 78% of world production and world's largest turmeric exporter. The fall in the production and yield of turmeric in India is attributed to several reasons, among which plant diseases are the foremost threat. The various foliar diseases that affect turmeric leaf are categorized as leaf spot, leaf blotch and rhizome rot.

Leaf spot of turmeric is the most important disease of turmeric. It has become a major constraint in successful cultivation of turmeric. The disease has resulted in drastic reduction in rhizome yield. Symptom appears as brown spots of various sizes on the upper surface of the young leaves. The spots are irregular in shape and white or grey in the centre. Later, spots may coalesce and form an irregular patch covering almost the whole leaf. 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 leaves also turn yellow. In severe cases the plants present a scorched appearance and the rhizome yield is reduced. Foliar symptoms of rhizome rot appears as light yellowing of the tips of lower leaves which gradually spreads to the leaf blades. In early stages of the disease, the middle portion of the leaves remain green while the margins become yellow.

The present work is aimed to develop a feature selection mechanism for identifying three types of leaf diseases in turmeric. Feature selection methods aid in creating an accurate predictive model by choosing features that will give good accuracy while requiring less data. Irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model can be identified and

removed as they decrease the accuracy of the model. The complexity of the model is reduced with fewer attributes. This research work is then carried out using machine learning methods for turmeric leaf disease identification.

II. LITERATURE SURVEY

Revathi, et al.,[1] proposed enhanced PSO feature selection method that adopted skew divergence method and used features like edge, color and texture variances. The extracted feature was input to the SVM, Back propagation neural network (BPN), Fuzzy with edge CYMK color feature and GA feature selection. Tests were performed to identify the best classification model to evaluate six types of diseases like Bacterial Blight, Fusariumwilt, Leaf Blight, Root rot, Micro Nurient, Verticilium Wilt. The Experimental results obtained show that the robust feature vector set which is an enhancement of a feature extraction method (EPSO) performed well.

Yan Cheng Zhang, et al., [2] tried to identify and diagnose cotton disease using computer vision. He proposed the fuzzy feature selection approach, fuzzy curves (FC) and surfaces (FS) to select features of cotton diseased leaves image. A subset of independent significant features was identified exploiting the fuzzy feature selection approach in order to get best information for diagnosing and identifying, This approach reduced the dimensionality of the feature space which lead to a simplified classification scheme appropriate for practical classification applications. The authors showed that the effectiveness of features selected by the FC and FS method is much better than that selected by human randomly or other methods.

Bernardes A. A. et al., [3] proposed a method for automatic classification of cotton diseases through feature extraction of leaf symptoms from digital images. Wavelet transform energy was used for feature extraction while SVM was used for classification. The image set of supposedly adulterated leaves was classified within one of the four other sub - classes, namely: MA, RA, AS, and NONE. The system performed well with 96.2% accuracy for the SA class, 97.1% accuracy for the MA class, 80% accuracy for the RA class, and 71.4% accuracy for the AS class. Mr. Hrishikesh P. et al.,[4] in their paper presented some important features of diseased leaves which will help to find exact disease of plant. P.Revathi et al.,[5] iused three features namely color feature variance, shape and texture feature to identify cotton leaf spot diseases. Skew divergence color variance feature was calculated by color histogram and color descriptor. The shape Skew divergence feature was calculated by Sobel and Canny through the edge variance and edge location using Edge detection method. The skew divergence texture feature was calculated by Gober filter and texture descriptor. This investigation was based on six types of diseases and utilized three features combining the classifier of proposed Cross Information Gain Deep forward Neural Network (CIGDFNN) with 95 % accuracy.

Xinhong Zhang, et al.,[6] recommended a machine vision technique based system for the automatic inspection of flue-cured tobacco leaves. Machine vision techniques were used in this system to solve problems of features extraction and analysis of tobacco leaves, which included features of color, size, shape and surface texture. The experimental results showed that this system was viable for the features extraction of tobacco leaves, and can be used for the automatic classification of tobacco leaves.

III. PROPOSED METHODOLOGY

The goal of this work is adoption of feature selection techniques with machine learning for detection of turmeric diseases caused by fungi. This system uses technologies such as image processing, support vector machines and machine learning techniques for the timely detection of foliar diseases in turmeric. This work focuses on combining classification approach with feature selection techniques to identify three types of diseases. The leaf images of turmeric plants from turmeric fields in Coimbatore district were collected. The dataset with leaf images of different categories were pre-processed and segmented to promote efficient feature extraction. Colour, shape and texture features from the diseased portion of the leaves were extracted.

There are four different set of images collected based on the diseases like leaf blotch, leaf spot, rhizome rot and normal leaf comprising of 200 images in each class. Image-processing techniques are then applied to the acquired images to extract useful features that are necessary for further analysis. The collected images are of different dimensions and hence it is essential to convert them to uniform size for efficient preprocessing. Initially the RGB images are resized and converted into Hue Saturation Intensity representation. HSI color space representation is an ideal tool for color perception. The green pixels are then removed using masking. Masking means setting the pixel value in an image to zero or some other background value. This step identifies the green colored pixels as they mostly represent the healthy areas of the leaf.

K-means segmentation technique is then used to segment the diseased portion from the original image. Segmentation aims to change the representation of an image into meaningful image that is more easier to explore. The input data points are classified 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 clustering in them. The overall architecture of the proposed system is illustrated in Fig.3.1.

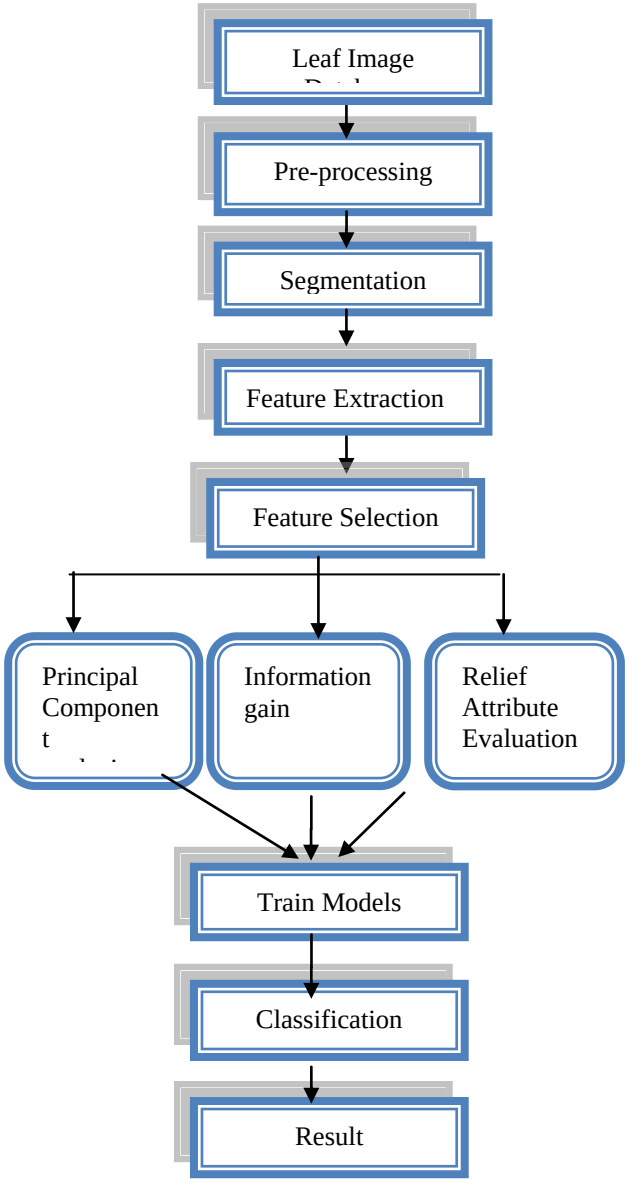

Fig.1 Overview of Proposed System

33 different features related to color, texture and shape are then extracted from the segmented images. The texture features investigated are energy, entropy, contrast, variance, homogeneity, correlation, maximum probability, sum average, sum entropy, sum variance, difference variance, difference entropy, information measures of correlation, cluster shade, cluser performance and dissimilarity. The shape features like Solidity, Eccentricity, Perimeter and the color features such as meanR, meanG, meanB are extracted from the leaf images. Feature selection is then applied to identify the best features from this dataset for accurate classification.

This paper experiments with three methods of feature selection namely Principal Component analysis, Information Gain and Relief-f Attribute Evaluator.

Feature Selection

Feature selection plays an important role in image processing and data mining. It computes an optimal subset of predictive features measured in the original data. It enables to achieve maximum classification performance by reducing the number of features used in classification while maintaining acceptable classification accuracy. Subset of the original features which retain adequate information to discriminate well among classes are selected. Several search algorithms have been used for feature selection. This work implements Principal Component Analysis, Information gain and Relief-f Attribute Evaluator.

A. Principal Component Analysis (PCA)

Principal component analysis, performs a linear mapping of the data to a lower dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. PCA, a non-parametric method builds a set of features by selecting those axes which maximize data variance. PCA can be used to reduce a complex data set to a lower dimensionality, to reveal the structures or the dominant types of variations in both the observations and the variables. It is a quantitatively rigorous method that generates a new set of variables, called *principal components*. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data.

B. Information gain

The information gain of an attribute tells the amount of information an attribute provides with respect to the classification target. Information gain (IG) measures the amount of information about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. In machine learning, information gain can be used to help ranking the features. A feature with high information gain should be ranked higher than other features because it has stronger power in classifying the data. Shannon entropy is the common measure for the information. Information gain is the reduction in the entropy that is archived by learning a variable. Concretely, it measures the expected reduction in entropy.

$IG= H(Y)- H(Y/X)$
$H(Y)= -\sum P(Y)LOG(P(Y))$
$H(Y/X)= -\sum P(X)\sum P(Y/X)LOG(P(Y/X))$

where P(Y) is the marginal probability density function for the random variable Y and P(Y |X) is the conditional probability of Y given X.

C. Relief-F Attribute Evaluator

A key idea of the original Relief algorithm (Kira & Rendell, 1992b), is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other. Given a randomly selected instance R_i, Relief searches for its two nearest neighbors, one from the same class, called nearest hit H, and the other from the different class, called nearest miss M . It updates the quality estimation for all attributes.

The ReliefF(Relief-F) algorithm(Kononenko, 1994) is not limited to two class problems, is more robust and can deal with incomplete and noisy data. Similarly to Relief, ReliefF randomly selects an instance R_i, but then searches for k of its nearest neighbors from the same class, called nearest hits H_j, and also k nearest neighbors from each of the different classes, called nearest misses Mj (C) . It updates the quality estimation WA for all attributes A depending on their values for R_i, hits H.

The selected features from each technique are used for training three classifiers SVM, Naive Bayes and Decision trees. The machine learning algorithms are tested using 10-fold cross validation to obtain better classification result.

IV. EXPERIMENTS AND RESULTS

Figure 2 shows the performance evaluation of the classifiers utilizing all 33 features before performing feature selection.

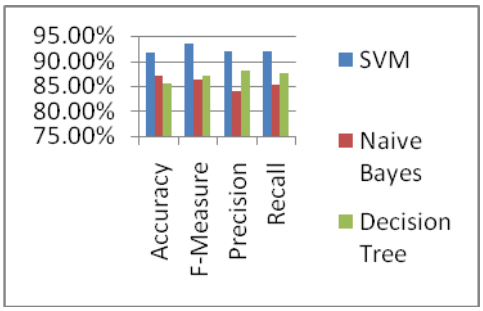

Fig.2 Comparison of Classifier Performance

SVM outperforms the rest of the classifiers in terms of accuracy (91.75%), precision(0.9195), recall(0.92) and F-measure(0.9350) as depicted in Table 1.

TABLE I
Predictive Performance of Classifiers

Evaluation Measure	SVM	Naive Bayes	Decision Tree
Accuracy	91.75%	87%	85.50%
F-Measure	0.9350	0.8620	0.87
Precision	0.9195	0.8395	0.88
Recall	0.92	0.8525	0.875

Method	Evaluation Measure	SVM	NB	DT
Principal Component Analysis	Accuracy	90.5%	87%	88.75%
	F-Measure	0.8820	0.8875	0.8515
Information Gain	Accuracy	93.75%	89%	87.50%
	F-Measure	0.9415	0.8168	0.8780
Relief Attribute Evaluation	Accuracy	83.25%	78.5%	80.75%
	F-Measure	0.8400	0.7800	0.8150

After each feature selection technique is completed, the resulting features from the respective methods were fed as input to each classifier. The classifiers in turn were trained and cross validated. There was a notable increase in the accuracy of the classifiers as shown in Table II.

TABLE II
Predictive Performance Of Classifiers With Feature Selection

SVM and Naive Bayes showed better accuracy with Information gain whereas Decision trees increased its accuracy with PCA. The comparative results indicate that the combination of SVM with Information gain results in a better performance when compared to other models (Figure 3).

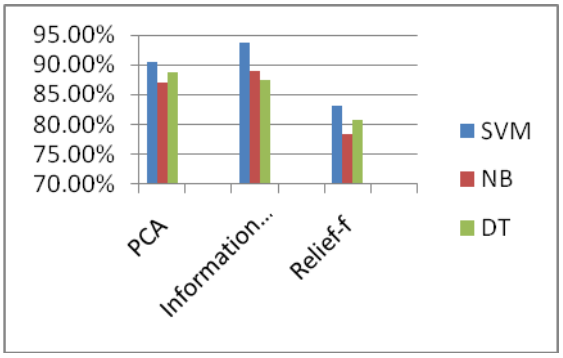

Fig.3 Comparison of Accuracy

V. CONCLUSION

The proposed work involves feature selection and machine learning techniques to investigate three types of diseases in turmeric namely leaf spot, leaf blight and rhizome rot. The work compares three feature selection techniques like PCA, Information gain and Relief-f attribute evaluator combined with the classifiers SVM, Naive Bayes and Decision trees. Performance evaluation of the proposed system shows that classification of turmeric leaf diseases using Support Vector Machine pooled with information gain gives better accuracy of 93.75% when compared to other algorithms.

REFERENCES

[1] P. Revathi, M. Hemalatha, Cotton Leaf Spot Diseases Detection Utilizing Feature Selection with Skew Divergence Method, International Journal of Scientific Engineering and Technology, Volume No.3, Issue No.1 , ISSN : 2277-1581

[2] Yan Cheng Zhang, Han Ping Mao, Bo Hu, Ming Xili,"Features selection of Cotton disease leaves image based on fuzzy featureselection techniques", IEEE Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, 2-4.2007

[3] Bernardes.A.A, J.G.Rogeri, N.Marranghello, A. S. Pereira, A.F. Araujo and João Manuel R. S. Tavares. "Identification of Foliar Diseases in Cotton Crop". SP, Brazil

[4] Mr. Hrishikesh P. Kanjalkar, Prof. S.S.Lokhande, "Feature Extraction of Leaf Diseases", International Journal of Advanced Research in Computer Engineering & Technology, Volume 3, Issue 1, January 2014

[5] P.Revathi, M.Hemalatha, "Identification of Cotton Diseases Based on Cross Information Gain Deep Forward Neural Network Classifier with PSO Feature Selection", International Journal of Engineering and Technology, Vol 5 No 6 Dec 2013-Jan 2014, ISSN : 0975-4024

[6] Xinhong Zhang , Fan Zhang, "Images Features Extraction of Tobacco Leaves", CISP '08 Proceedings of the 2008 Congress on Image and Signal Processing, Vol. 2 - Volume 02, p 773-776

[7] Gulhane.V. A & A. A. Gurjar, "Detection of Diseases on Cotton Leaves and Its Possible Diagnosis" (IJIP), 5 (5): 591-598. 2011

[8] Ajay A. Gurjar and Viraj A. Gulhane, "Disease Detection On Cotton Leaves by Eigenfeature Regularization and Extraction Technique". IJECSCSE.1 (1): 1-4.2012.

[9] Mrunalini R. Badnakhe and Prashant R. Deshmukh.."Infected Leaf Analysis and Comparison by Otsu Threshold and k-Means Clustering".2(3): 449-452. 2012

[10] Arivazhagan.S , Newlin Shebiah.R, Ananthi.S. Vishnu Varthini.S, 2013, "Detection of unhealthy regions of plant leaves and classification of plant leaf diseases using texture features", Agric Eng Int: CIGR Journal, 15 (1):211-217

[11] F.J. Ferri, P. Pudil, M. Hatef, and J. Kittler, 1994. "Comparative study of techniques for largescale feature selection," in: E. S. Gelsema and L. S.Kanal, eds., Pattern Recognition in Practice IV, Multiple Paradigms,Comparative Studies and Hybrid System (Elsevier, Amsterdam,) : 403–413.

[12] Meunkaewjinda. A, P.Kumsawat, K.Attakitmongcol and A.Sirikaew."Grape leaf disease Detection n from color imaginary using Hybrid intelligent system", Proceedings of ECTI-CON. 2008