Progressive Research – An International Journal
Vol. 5 (Special Issue) : 460-463 (2010), ISSN : 0973-6417

Society for Scientific Development in Agriculture
and Technology, Meerut (U.P.) INDIA

IDENTIFICATION OF TOMATO LEAF DISEASES USING FEATURE SELECTION TECHNIQUES

Kumar Sanjeev[1] and Suneeta Paswan[2]
[1]Rajendra Agricultural University, Pusa, Samastipur, Bihar
[2]Krishi Vigyan Kendra, Agwanpur, Saharsa, Bihar

ABSTRACT

The art of growing food is fundamental for the human subsistence. The manifestation of diseases causes many damages, either ?nancial or in terms of the quality of the crops, causing considerable losses if the degree of infestation is high. 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 tomato plants. The tomato leaf is highly exposed to diseases like early blight, late blight, bacterial leaf spot, etc. This system uses technologies such as feature selection and machine learning techniques for the identification and classification of diseases in tomato 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 performances 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 PCA performed well with an accuracy of 92%.

Key Words : *Tomato, crop disease, feature selection, machine learning.*

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. Several agrochemicals are applied to the plantation in an effort to minimize and control pathogens. However, the agrochemicals are usually harmful to the human health, can increase production costs, and may contaminate water and soil. Aiming at minimizing agrochemicals use, ensuring product quality and minimizing inherent agricultural production problems, computer applications have been developed and revealed high efficacy. The use of computers in agriculture has been subject of several scienti?c works, many of them focusing on the identi?cation of diseases of various crops.

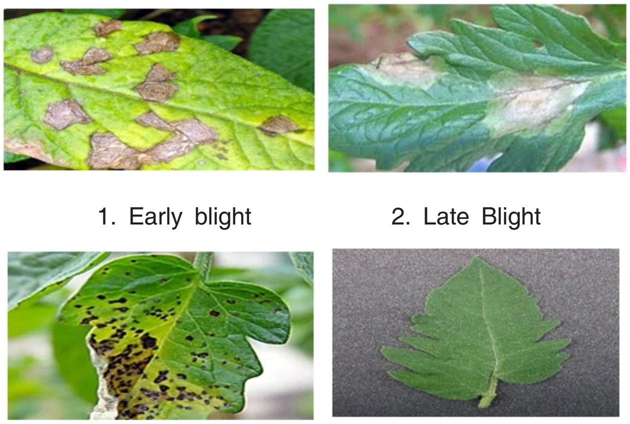
Tomato (*Solanum lycopersicum* L.), one of the most popular vegetable crops in the world, shares a coveted position in India as fresh vegetable and also being used as a variety of processed products such as juice, ketchup, sauce, canned fruits, puree, paste, etc. There has been a gradual increase in area under tomato cultivation in India while the productivity has been fluctuating ranging from 14.70t/ha in 1991-92 to 18.20t/ha in 2005-06. The major limiting factors towards production of optimum yield are considerable biotic stresses caused by fungi, bacteria, viruses, viroids, nematodes and insect-pests in existing varieties and hybrids. Outdoor production of tomato is seriously impaireddue to increasing infections with evolving early blight (*Alternaria solani*), late blight (*Phytophthora infestans*) populations and leaf curl virus (ToLCV) diseases particularly in the India.

The present work is aimed to develop a feature selection mechanism for identifying four types of leaf diseases in tomato. 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 tomato leaf disease identification.

Kai (1992) proposed a method to discriminate between pairs of diseases in wheat and grapevines. The images are segmented by a K-means algorithm, and then 50 color, shape and texture features are extracted. For the purpose of classification, the authors tested four different kinds of neural networks: Multilayer Perceptron, Radial Basis Function, Generalized Regression, and Probabilistic. The authors reported good results for all kinds of neural networks. Sanyal *et al.* (1997) tackled the problem of detecting and classifying six types of mineral deficiencies in rice crops. First, the algorithm extracts a number of texture and color features. Each kind of feature (texture and color) is submitted to its own specific MLP neural network. Both networks have one hidden layer, but the number of neurons in the hidden layer is different (40 for texture and 70 for color). The results returned by both networks are then combined, yielding the final classification. Xu *et al.* (2000) proposed a method to detect nitrogen and potassium deficiencies in tomato plants. The algorithm begins extracting a number of features from the color image. The color features are all based on the b* component of the L*a*b* color space. The texture features are extracted using three different methods: difference operators, Fourier transform and Wavelet packet decomposition. The selection and combination of the features was carried out by means of a genetic algorithm. Finally, the optimized combination of features is used as the input of a fuzzy K-nearest neighbour classifier, which is responsible for the final identification.Yan Cheng Zhang, *et al.* (2007) 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. Hetzroni *et al.* (2008) was carried out to monitor plant health and their system tried to identify iron, zinc and nitrogen deficiencies by monitoring lettuce leaves. Those parameters are finally fed to neural networks and statistical classifiers, which are used to determine the plant condition.

MATERIALS AND METHODS

Dataset of Pomegranate Leaf Image : Plant Village (www.plantvillage.org), a publicly available image database, contains 54,306 images of diseased and healthy plant leaves of 14 crop species collected under controlled conditions and the ground truths are also provided. We analyze200 images of tomato leaves, which have a spread of the following class labels assigned to them: Early blight affected tomato leaf, Late blight affected tomato leaf, Bacterial leaf spot affected tomato leaf, Healthy or Non-diseased tomato leaf.



1. Early blight 2. Late Blight

3. Bacterial leaf spot 4. Healthy leaf

Fig-1 : Different Class of Tomato Leaf Diseases.

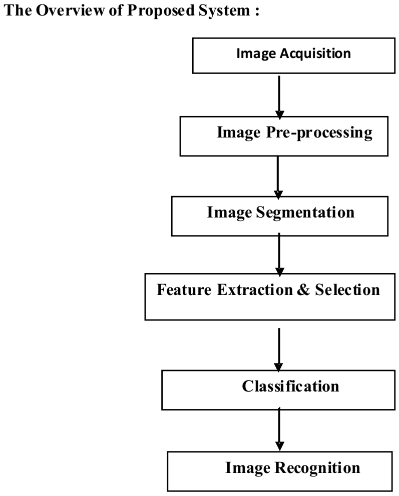


Fig.-2 : Overview of Proposed System**.**

Image Acquisition : The image acquisition is required to collect the actual source image. An image must be converted to numerical form before processing. This conversion process is called digitization.

Image Pre-processing : The principle objective of the image enhancement is to process an image for a specific task so that the processed image is better viewed than the original image. Image enhancement methods basically fall into two domains, spatial and frequency domain. Spatial domain : as the name suggests in this approach different methods are used, which will affect the manipulation of pixel values of an image.

Frequency domain : in this method first, a Fourier transform of the image is computed and then different operations are performed on them and finally results are obtained by getting the inverse Fourier transform of the image.

Image Segmentation : In image processing, segmentation falls into the category of extracting different image attributes of an original image. Segmentation subdivides an image into constituent regions or objects. The level to which that subdivision carried out is a problem specific. The simplest method among all segmentation methods is threshold-based method, whose volume uses either a manually or automated generated threshold values for segmentation.

Feature Extraction and Selection : Feature Extraction and Selection almost always follow the output of a segmentation stage. The first decision must be made whether the data should be represented as a boundary or complete region. Boundary representation is appropriate when the focus is on external shape characteristics whereas regional representation is focusing on internal properties, such as texture and skeletal shape.

Classification : Classification is a usual process used to recognize image. Classification is needed to distinguish a plant species with other species based on the data obtained from feature selection. Artificial neural network (ANN), Support vector machine (SVM) and fuzzy logic are the most commonly techniques used in classification.

Image Recognition : The descriptors from the image data stored in database are compared with the descriptors from the query image. The closer gap within those descriptors is then chosen to appoint the query image to be in which class.

Feature Extraction : 30 different features related to colour, 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, cluster performance and dissimilarity. The shape features like Solidity, Eccentricity, Perimeter and the colour features such as mean R, mean G, mean B are extracted from the leaf images. Feature selection is then applied to identify the best features from this dataset for accurate classification.

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.

Principal Component Analysis : 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.

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. Shannon entropy is the common measure for the information. Concretely, it measures the expected reduction in entropy.

$$IG = H(Y) - H(Y/X)$$

$$H(Y) = - \quad P(Y)LOG(P(Y))$$

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

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.

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.

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.

Confusion Matrix :

$\text{Precision} = \frac{TP}{TP + FP}$, where TP is True positive, FP is False positive

$\text{Recall} = \frac{TP}{TP + FN}$, where FN is False negative

$$F = \frac{2 \quad \text{precision} \quad \text{recall}}{\text{precision} + \text{recall}}$$

RESULTS AND DISCUSSION

The performance evaluation of the classifiers utilizing all 30 features before performing feature selection.

After each feature selection technique is

Table-1 : Predictive Performance of Classifiers with Feature Selection.

Method	Evaluation Measure	SVM	NB	DT
Principal Component Analysis	Accuracy	92%	83.5%	80.5%
	F-Measure	0.8655	0.8174	0.8005
Information Gain	Accuracy	90%	80%	77%
	F-Measure	0.9115	0.8034	0.7655
Relief Attribute Evaluation	Accuracy	81%	78%	75%
	F-Measure	0.8211	0.7805	0.7535

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 measured the accuracy of the classifiers as shown in Table-1.

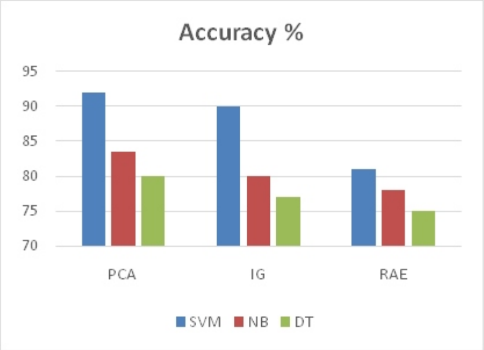


Fig-3 : Comparison of Accuracy.

Comparison of Accuracy : SVM and Naive Bayes showed better accuracy with Information gain. The comparative results indicate that the combination of SVM with Information gain results in a better performance when compared to other models. SVM provides prediction of accuracy with 92%.

F-Measure Comparison : SVM performs better in all feature selection techniques and SVMalso performs better for F-measure.

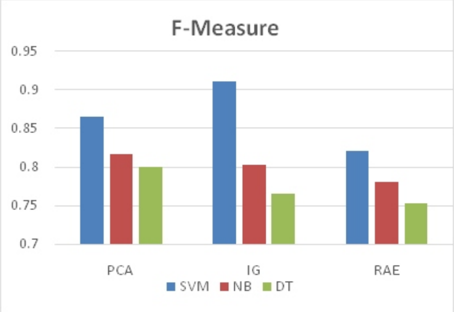


Fig-4 : F-Measure Comparison.

CONCLUSIONS

Plant disease has become a major threat to global food security. Plant diseases contribute 10-16% losses in the global harvest of crops each year. The proposed work involves feature extraction and selection and machine learning techniques to investigate four types of disease classes in tomato namely Early blight affected tomato leaf, Late blight affected tomato leaf, Bacterial leaf spot affected tomato leaf, Healthy or Non-diseased tomato leaf. 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 tomato leaf diseases using Support Vector Machine pooled with information gain gives better accuracy of 92% when compared to other algorithms.

REFERENCES

1. Hairuddin, S.P., J. Sil (2008).“rice disease Identification using pattern recognition techniques, *in Proc. 11th Int. Conf. on computer and information technology*, Khulna, Bangladesh, pp. 420-423.
2. Hetzroni, Hossein Nejati, ZohrehAzimifar, Mohsen Zamani (2008).*Using Fast Fourier Transform for weed detection in corn fields,* IEEE.
3. Kai, K., Ackley, D., Littman, M. (1992).Interactions between learning and evolution,In C. G. Langton, C. Taylor, J.D. Farmer, and S. Rasmussen, eds., Artificial Life II. Addison-Wesly.
4. Meunkaewjinda. A., Kumsawat, P., Attakitmongcol, K., Sirikaew, A. (2008). Grape leaf disease Detection from color imaginary using Hybrid intelligent system”, Proceedings of ECTI-CON. 2008
5. Sanyal, A., Barýþ, T.U., Sankur, B.(1997). The Performance of Thresholding Algorithms for Optical Character Recognition,*Int. Conf. on Document Analysis and Recognition: ICDAR’97,* Ulm., Germany, pp.697-700.
6. Xu, C.C.Y., Prasher, S.O., Landry, J.A., Ramaswamy, H.S., Ditommaso, A. (2000).Application of artificial neurol networks in image precongation and classification of crop and weeds,*Canadian Agricultural Engineeering*, vol. 42,no. 3, pp. 147-152.
7. Yan, C.Z., Mao, H.P., Ming Xili, B.H., (2007). 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.