Automated Classification of Bacterial and Fungus Infection in Crops

S. Raja[1], Dr. C. Ashok Kumar[2]

[1] *Research Scholar, Department of Computer and Information Science, Annamalai University, Tamilnadu, India.* rajadeo@gmail.com

[2] *Research Supervisor, Department of Computer and Information Science, Annamalai University, Tamilnadu, India*

Abstract: The classification of diseases in the crops and plants are very interesting research area in the field of agricultural image processing. The diseases and the parasite infections present in the crops will multiply uncontrollably if left unnoticed. Hence the detection of such infection at the early stage is imperative for the benefit of the productivity of the crop yield. Automating this process is a cumbersome task as it needs intense mechanism and process to identify the type of infection (normal to abnormal). This paper introduces a new method using gray level co-occurrence matrix GLCM to extract the defects from the images of the crop. The proposed algorithm categorizes the images into two classes (normal and abnormal) and the extracted features are then modeled using support vector machines, decision tree and k-nearest neighbor algorithms. The experimental result showcased that the decision tree performed with 98%accuracy in classifying the abnormal crops and outscored the other two algorithms by a good margin.

Keywords: GLCM, KNN, SVM, Normal and abnormal

1. INTRODUCTION

It is a known fact that most of the Indian population depends on the agricultural income and most of the population earn their bread and butter from the agri based employment especially in the rural India. As per the proverb “Prevention is better than cure”, it is quite important for the farmers to adapt to some technological advance method to discover the infections at the early stage to evade huge losses in the yield of the crop. India is now the second biggest agricultural producer in the world and this agricultural sector plays a pivotal role in the socio-economic development of India. The productivity in this sector is low since only one third of the agricultural land is being cultivated. So naturally the demand for the food is perpetually increases to the supply decreases constantly. The government of India is taking many steps to increase this productivity by educating the farmers with some advanced methods, technologies, and processes but the farmers are adopting the old custom manual method during irrigation thus incurring heavy losses due to infections and natural calamities.

`The data in agriculture industry is very precious as it will produce a pattern which will be useful for the farmers to identify the infections and diseases present in the leaves very quickly. Classification is an important form of data analysis that predicts and separate the raw data into classes and used to find the future data trends. This process learns to predict from the training sets which is further used to predict the discrete classes on new data sets.

K-nearest neighbor classifiers (KNN) classify a data occurrence by considering just the k most comparative data occasions in the input data set [16].. A preparation set is known and ordering tests of obscure classification is utilized in the process of clsssification. The essential supposition in the k-NN algorithm is that comparative examples ought to have comparable classification to predict the upcoming data trend for the new samples. As in the k-implies approach, the likenesses between tests are estimated utilizing reasonable distance capabilities.

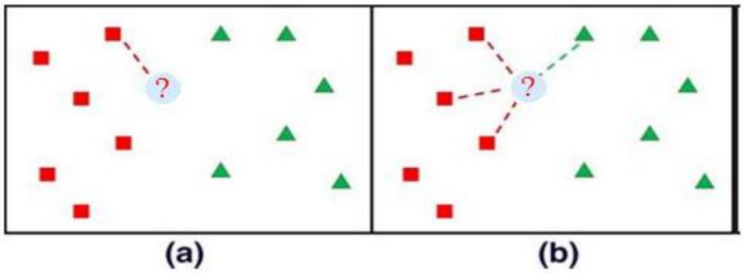


Figure 1: K-nearest neighbor where ? is the point where classification starts

Support vector machines are binary classifiers and it is shown in the figure 2.

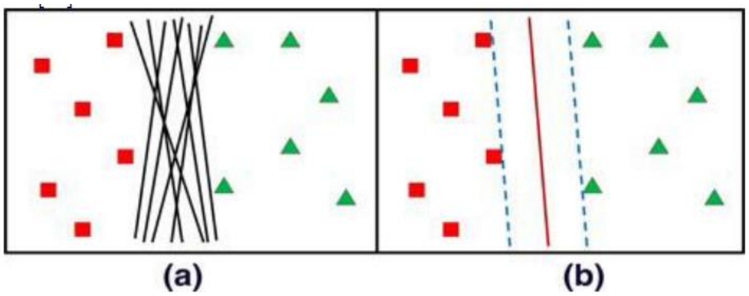


Figure 2: SVM classifier

The decision tree is a flowchart type of classification algorithm where there are two important components, namely,

1. Leaf – Nodes that assign the classes for observation
2. Internal nodes – specify test of each attributes and represent the outcome of the classification result.

2. PROPOSED APPROACH

The overall architectural diagram of the proposed method is shown in the figure 3. The input imageis preprocessed initially using the median filter to smoothen the raw image to extract minute features used for classification. The raw input image is first divided into two sets, namely normal and abnormal. The important features that are extracted are then modeled using SVM, KNN and DT for classification [8].

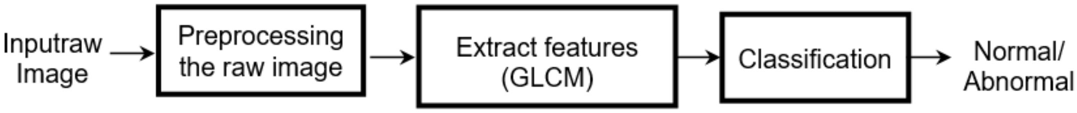


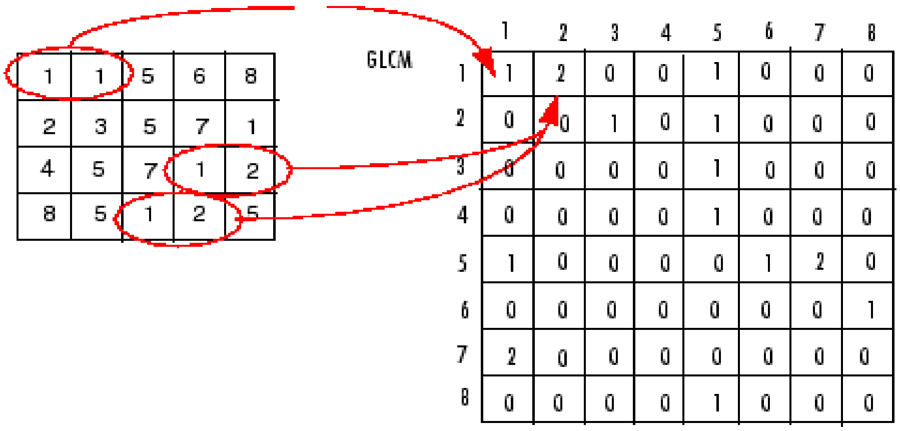
Figure 3: Architectural diagram of the proposed approach

3. EXTRACT FEATURES

The feature extraction is an important process where very minute and key parts present in the leaf, crops are extricated for further study to find the exact disease and infection that has been occurred in the plants. The upcoming section illustrates the feature extraction clearly.

4. GLCM

A mathematical technique for inspecting surface those arrangements with the spatial association of pixels is the gray level co-event lattice. The methodology and execution behind the Gray Level Co-event Matrix (GLCM) technique are introduced in the cited reference [3]. GLCM is calculated by computing how often a pixel with gray-scale intensity values *M* occurs adjacent neighboring to a pixel with the valuecorresponding to *N*. Here every element*(M,N)* in GLCM clearly denoteshow many times the pixel with the value *M* occurred adjacent neighboring pixel value*N*.



1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0
5	1	0	0	0	0	1	2	0
6	0	0	0	0	0	0	0	1
7	2	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0

Figure 4: GLCM Matrix for distance 1 and 0° direction

The adjacent pixels can be computed from the following directions like 0°(horizontal), 45°(vertical), 90°(left) and 135°(right) degrees in a two-dimensional pixel input image as shown in the figure. Only the element (1,2) is occurring twice as shown in the figure. The co-occurrence matrix P for an image is shown in the following section,

where (Δx, Δy) are offsets for the pixel we work upon and the neighboring pixel.

The next attribute is contrast, and this contrast is used to measure the local variation in the pixel intensity of the given input raw image.

$$P(i,\, j) = \sum_{x=1}^{G}\sum_{y=1}^{G} \begin{cases} s1, & if\ I(x,y) = i\ and\ I(x+\Delta_x, y+\Delta_y) = j; \\ 0, & Otherwise \end{cases}$$

$$Contrast = \sum_{i,j=0}^{G-1} (i-j)^2 p(i,j)$$

The next computation is entropy as it is the measure of the random intensity distribution. The formula for the entropy is shown here,

$$Entropy = \sum_{i,j=0}^{G-1} p(i,j)\ (\text{-}Inp(i,j))$$

The next calculation is the dissimilarity which is the separation of high and low contrast region present in the input image.

$$Dissimilarity = \sum_{i,j=0}^{G-1} |\, i-j \,|\, P(i,j)$$

The next few attributes are sum of square variance, sum average and sum variance.

$$Variance = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} (i-\mu)^2\ p(i,j)$$

$$Sum\ Average = \sum_{i=0}^{2G-2} i\ P_{x+y}(i)$$

$$Sum\ Variance = \sum_{i=0}^{2G-2} (i - aver)^2\ P_{x+y}(i)$$

5. SUPPORT VECTOR MACHINE

The fundamental gamble minimization principle is what the Support Vector Machine (SVM) [6] is dependent on (SRM). Support vector machines can be used for nonlinear relapse and design categorization. It creates a direct model based on non-straight class bounds and support vectors to evaluate the choosing capability. SVM develops direct machines for an ideal hyperplane that accurately separates the data and into the largest distance between the hyperplane and the closest preparation focuses, assuming that the data are straightly distinct. Support vectors are the preparation focuses that are closest to the ideal isolating hyperplane. Through nonlinear planning picked deduced, the SVM maps the information designs into a more layered. Consequently, SVM is a straight classifier in the boundary space, yet it turns into a nonlinear classifier because of the nonlinear planning of the space of the information designs into the high layered include space.

Various types of SVM kernels are shown in the following table 1,

Table 1: SVM kernels

Types of kernels	Inner Product Kernel	Details
Polynomial	$(x^T x_i + 1)^p$	Where x is input patterns, x_i is support vectors, σ^2 is variance, $1 \le i \le N_s$, N_s is number of support vectors, β_0, β_1 are constant values. p is degree of the polynomial
Gaussian	$exp\left[-\frac{\lVert x^T - x_i \rVert^2}{2\sigma^2}\right]$	
Sigmoidal	$tanh(\beta_0 x^T x_i + \beta_1)$	

6. DECISION TREE

One of the early learning algorithms, decision trees create meaningful rules by classifying the input raw data into a classification tree [5]. The element space is divided recursively with respect to a training or practice set to create the classification tree. A decision tree is a visual representation of the information and the issue, with the output shown in illustrative form. A decision tree helps to break down a complex problem into smaller, more manageable tasks. A decision tree is a common and natural method for classifying examples through a series of questions, where each question is dependent upon the previous one's outcome.

The commonly used technique for design classification is decision trees. A decision tree is a common and organic technique to deal with classifying examples through a series of inquiries, where the next inquiry depends on the outcome of the previous one. A visual representation of an issue is a decision tree. A decision tree helps break down a challenging problem into smaller, more manageable tasks. This enables the decision makers to reach more modest findings while still achieving the best overall decision. A systematic, structured approach to dealing with decision-making is decision tree investigation.

7. EXPERIMENTAL RESULTS

The experiments are performed using MATLAB 2013a under Windows 10and a machine with an Intel I5 processor clocked at 2.43 GHz and 4 GB of RAM. In order to create a model for each class, the collected GLCM features are input into supervised classifiers like SVM, K-NN, and Decision Tree. These models are then used to assess how well the proposed features perform.

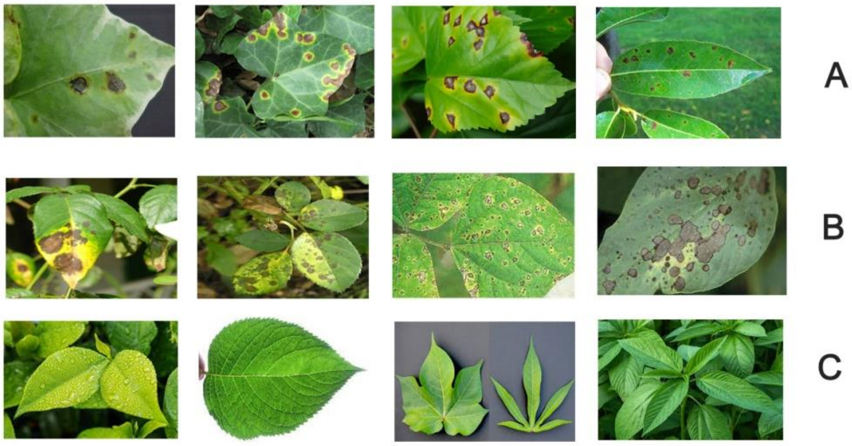


Figure 5: A - bacterial infection, B – Fungus infection, C – Normal leaf

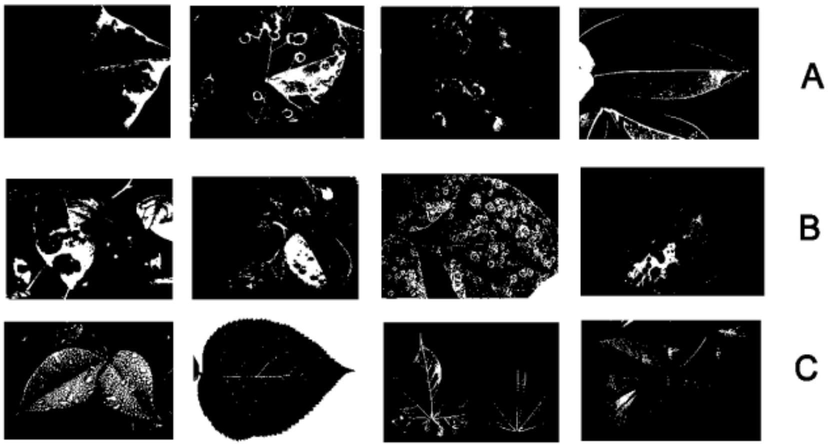
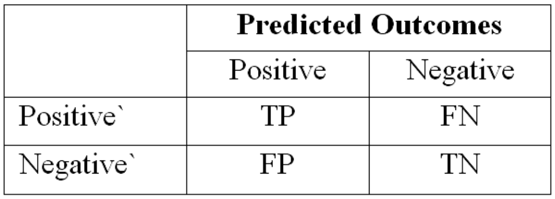


Figure 6: Binarized input image dataset

8. EXPERIMENTAL EVALUATION

The most commonly used metrics in the evaluation of accuracy are precision, recall and F-measure [8]. These three metrics tend to give the best gauge for the performance of the classification. The confusion matrix for the evaluation is shown in the following figure 7.



	Predicted Outcomes	
	Positive	Negative
Positive`	TP	FN
Negative`	FP	TN

Figure 7: Confusion matrix for classification

A classification system's actual and anticipated classifications are detailed in the confusion matrix, where TP and TN represent the number of accurate positive and accurate negative predictions for a given class, respectively. The number of false negatives and false positives for a given class is represented by the letters FN and FP. Precision, Recall, and F-measure, which are frequently used in classification, cannot identify changes in TN while all other matrix values remain unchanged.

$$\text{Precision (P)} = \frac{TP}{TP+FP}$$

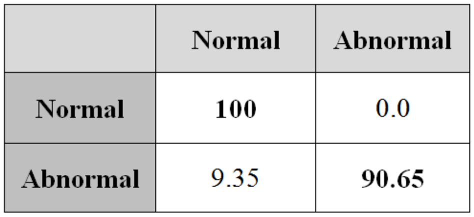
$$\text{Recall (R)} = \frac{TP}{TP+FN}$$

$$\text{F-Measure (F)} = 2\times\frac{P\times R}{P+R}$$

$$\text{Accuracy (A)} = \frac{TP+TN}{TP+FP+TN+FN}$$

9. SVM - RESULT

Table 2 displays the confusion matrices of the SVM classifier on the input raw dataset shown in the figure8, and the table's diagonal displays the accuracy of the infection related to bacteria and fungus. SVM has a 94.91 percent recognition rate on average bacterial infection and 95.23 percent recognition rate on fungus infection. In SVM, the normal class is almost accurately categorized, but in 9.35 percent of cases, the abnormal class is mistaken with the normal class. Therefore, it requires more care.

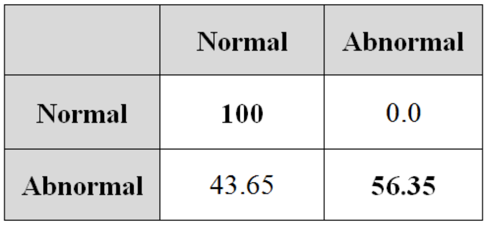


	Normal	Abnormal
Normal	**100**	0.0
Abnormal	9.35	**90.65**

Figure 8: Confusion matrix for SVM classifier

10. KNN - RESULT

In figure 9, which is the diagonal of the table, the confusion matrices of the k-NN classifier on the raw input dataset are displayed. The k-NN system's average recognition rate is 79.03 percent. In k-NN, the normal class is well and accurately categorized, but the abnormal class is mistaken for the normal class in 43.65 percent of cases.

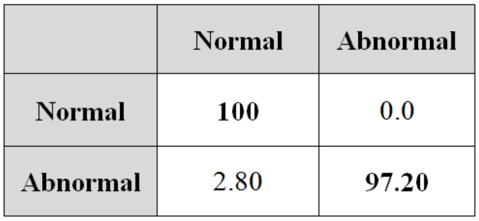


	Normal	Abnormal
Normal	**100**	0.0
Abnormal	43.65	**56.35**

Figure 9: Confusion matrix of K-NN classifier

11. DT – RESULT

In figure 10, which is the diagonal of the table, the confusion matrices of the Decision Tree classifier on the raw input dataset are displayed. DT has a 98.35 percent average recognition rate. In DT, the normal class is correctly identified, but 2.80 percent of the abnormal class is mistaken for the normal class.



	Normal	Abnormal
Normal	**100**	0.0
Abnormal	2.80	**97.20**

Figure 10: Confusion matrix of decision tree classifier

From the results it is clear that the DT algorithm has a higher accuracy than the SVM and K-NN algorithms.

Table 3: Metrics compared for the algorithms

Classifiers	Precision	Recall	F-measure
SVM	93.00	94.65	94.52
K-NN	78.73	77.28	72.89
Decision Tree	**97.61**	**98.89**	**98.18**

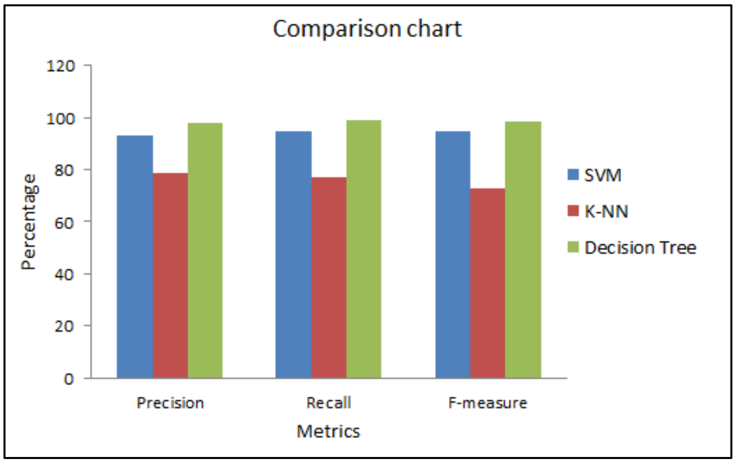


Figure 11: Comparison chart

12. CONCLUSION

This study uses an SVM, k-NN, and Decision Tree to efficiently classify crop/leaf pictures into normal and infected ones. In this research, a statistical feature extraction technique known as Gray Level Co-occurrence Matrix (GLCM) is presented. GLCM statistical features represent the significant textural properties of bacterial and fungus infectionin the leaves are taken and show highly promising results in the classification of the raw input images. The experimental results reveal that Decision Tree has a classification accuracy of 98.55%, proving that the proposed feature method works well and produces positive bacterial and fungus classification results. The experiments revealed that the system could not accurately distinguish between the abnormal class, which is something that will be of interest in the future work.

13. REFERENCES

1. Keller, James M., Michael R. Gray, and James A. Givens. "A fuzzy k-nearest neighbor algorithm." IEEE transactions on systems, man, and cybernetics 4 (1985): 580-585.
2. Vladimir NaumovichVapnik and VlamimirVapnik, Statistical learning theory, vol. 1, Wiley New York, 1998.
3. F. Albregtsen, “Statistical texture measures computed from GLCM”, Image processing Laboratory, Dept of Informatics, University of Oslo, 2008.
4. Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen, Classification and regression trees, CRC press, 1984.
5. J. Ross Quinlan, “Induction of decision trees,” Machine learning, vol. 1, no. 1, pp. 81–106, 1986.
6. NelloCristianini and John Shawe-Taylor, An introduction to support vector machines and other kernel-based

learning methods, Cambridge university press, 2000.

7. F. Albregtsen, “Statistical texture measures computed from GLCM”, Image processing Laboratory, Dept of Informatics, University of Oslo, 2008.
8. Marina Sokolova and Guy Lapalme, “A systematic analysis of performance measures for classification tasks,” Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.

DOI: https://doi.org/10.15379/ijmst.v10i2.3195