RICE PLANT DISEASE DETECTION USING TWIN SUPPORT VECTOR MACHINE (TSVM)

Bikash Chawal[1], Sanjeev Prasad Panday[2]

[1]Department of Computer Engineering, Khwopa Engineering College, Purbanchal University, Nepal
[2]Department of Electronics and Computer Engineering, Pulchowk Campus,
Institute of Engineering, Tribhuvan University, Kathmandu, Nepal

Abstract

Crop disease epidemics can cause severe losses and affect agricultural products and food security especially in south Asian countries and Nepal where rice is enjoyed as a staple throughout the year. To achieve automatic diagnosis of crop disease the proposed system aims to develop a prototype system for detection of the paddy disease. Image recognition of the disease would be conducted based on Image Processing techniques to enhance the quality of the image and Twin Support Vector Machine (TSVM) technique to classify the paddy disease. The methodology involves image acquisition, pre-processing, analysis and classification of the paddy disease. All the paddy sample images will be passed through the RGB calculation before it proceeds to the binary conversion. If the sample is in the range of normal paddy RGB, then it is automatically classify as normal. Then, all the segmented paddy disease sample will be converted into the binary data in data base before proceed through the TSVM for training and testing. The proposed system is targeted to achieve better recognition results.

Keywords: *TSVM, Image Processing, Binary Conversion, Image Recognition, Classification*

1. Introduction

One of the most utilized food plants and widely grown in Asia as an important crop is the Paddy plant. Around half of the world population relies on it as a major food product. Security of food through the prevention and detection of paddy plant disease is the major concern. The prevalence of paddy plant disease should be well known, and disease prediction should be caused out. Image Recognition of Paddy disease and automatic classification of disease severity are achieved by using image processing technologies. In Nepal, rice is the most important cereal crop and in 1966 total rice production amounted to a little more than 1 million tons; by 1989 more than 3 million tons were produced. Fluctuation in rice production was very common because of changes in rainfall, disease; overall, however, rice production had increased following well as increases in cultivated land. By 1988 the introduction of new cultivation techniques as approximately 3.9 million hectares of land were under paddy cultivation. Many people in Nepal devote their lives to cultivating rice to survive (Savada, 1991). Rice is the major food amongst all the ethnic groups in Nepal. In the Terai region, most rice varieties are cultivated during the rainy season. The principal rice growing season is from June to July when water is sufficient for only some part of the fields; the subsidiary season, known as "Ropai”, is from April to September, when there is usually enough water to sustain the cultivation of all rice fields. Farmers use irrigation channels throughout the cultivation seasons.

Being important crop that has deep impact in country’s economy, various plans have been proposed to upgrade rice cultivation, growth and reduce the impact of diseases. Rice is the seed of the grass species Oryza sativa (Asian rice) or Oryzaglaberrima (African rice) (UN-FAOSTAT, 2017). Rice leaves are the most vulnerable part that is easily affected by various diseases. Rice blast, caused by the fungus Magnaporthegrisea, (Talbot, 2005) is the most significant disease affecting rice

*Corresponding author: Bikash Chawal
Department of Computer Engineering,
Khwopa Engineering College,Libali-8 Bhaktapur, Nepal
Email:bikash.chawal@gmail.com
(Received: July 10, 2019 Accepted: November 10, 2019)

cultivation.Bacterial blight is caused by Xanthomonasoryzaepv. oryzae.Other major rice diseases include: sheath blight, rice ragged stunt and tungro (IRRI database,2013).

The aim of this paper was to develop the system that automatically detect and classify the paddy disease by using Image Processing technique with prediction accuracy with the support of k means clustering algorithm and cluster classifier TSVM (Twin Support Vector Machine) which ensure food security, paddy plant disease is to be controlled timely and effectively. Recognition and diagnosis of paddy disease relies on visual identification by agricultural technicians and personnel which requires high professional knowledge with rich experience. Besides, disease diagnosis through pathogen detection requires satisfactory laboratory and molecular biological techniques which is more time consuming and costly in terms of money and machinery equipment.

2.Methodology

The methodology for diagnosing paddy disease for the research purpose has been presented in the figure below. The given block diagram shows the training of the TSVM using the features extracted and the final classification by taking the weight values stored in the database.

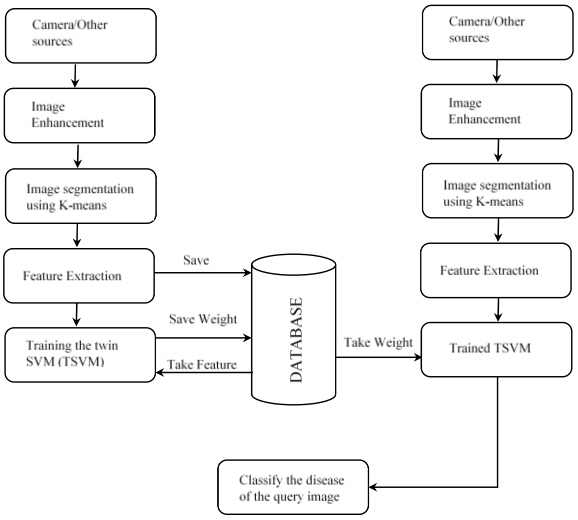


Fig.1. System Block Diagram

Here in the block diagram in fig 1, image segmentation using k-means is shown for feature extraction. However feature extraction has also been done from the images without k-means.

2.1 Image Acquisition

Digital images are acquired on environment conditions using digital cameras or imaging sensors. Those acquired image are enhanced and used to extract useful features that are necessary for detection of disease in plants.

2.2 Image Enhancement

It is the process to enhance the image in such way that increases the chance for success of other processes. It includes contrast stretching, noise removal, filtering etc.

2.3 Image Segmentation

Image segmentation is the process of partitioning a digitalimage into multiple segments (sets of pixels, also known as super-pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze.

2.4 Feature Extraction

In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.

2.5 K-means Clustering

K-means Clustering is used to decompose an image into meaningful partitions. Using k-means, defected portion and healthy portions of leaf can be extracted in separate cluster which help in further analysis of type of disease and image segmentation. The k-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed, and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop, we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$\boldsymbol{J(V)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{c_i}} \left|\left|\mathbf{xi}-\mathbf{vj}\right|\right|^{\mathbf{2}}$$

where

'||xi - vj||' is the Euclidean distance between data point xi and cluster center vj.

'ci' is the number of data points in ith cluster.

'c' is the number of cluster centers.

Deciding Number of Clusters: The number of clusters should match the data. An incorrect choice of the number of clusters will invalidate the whole process. An empirical way to find the best number of clusters is to try K-means clustering with different number of clusters and measure the resulting sum of squares.

K-means Algorithm: The Lloyd's algorithm, mostly known as k-means algorithm, is used to solve the k-means clustering problem and works as follows.

First, decide the number of clusters k. Then:

1. Initialize the center of the clusters

vj = set no. of cluster, j=1,...,k

2. Attribute the closest cluster to each data point

ci = { i:d(xi, vj) ≤ d(xi, vl), l≠i, i=1,...,n}

3. Set the position of each cluster to the mean of all data points belonging to that cluster

vj = (1|ci|) $\sum_{i\in\mathrm{ci}} x_i$,∀j

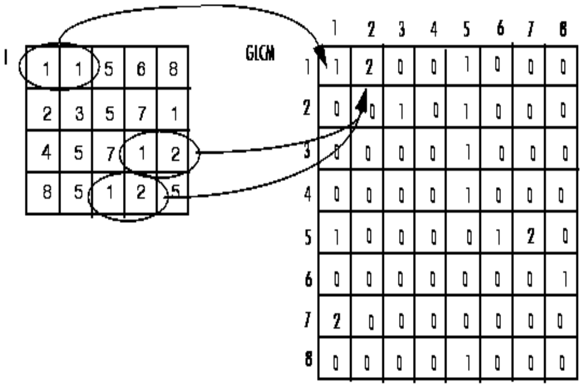
Notation:

|c|= number of elements in c

2.6 Gray Level Co-occurrence Matrix

A co-occurrence matrix is a matrix that is defined over an image to be the distribution of co-occurring pixel values (grayscale, values or colors) at a given offset. A statistical method of examining texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. The GLCM (fig 2) functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix. Statistical measures like Contrast, Correlation, Energy, Homogeneity and dissimilarities.

GLCM is used to extract various features from segmented image obtained from K-means clustering and then are fed into SVM for training, classification & prediction (testing). Various features that can be extracted using GLCM are: Contrast, Energy, Entropy,Homogeneity, RMS, Variance,Smoothness, Kurtosis, Skewness,Correlation and so on.



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

Fig.2. Gray Level Co-occurrence Matrix

2.7 Support Vector Machine (SVM)

In machine learning, support vector machines (SVMs, shown in fig 3) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

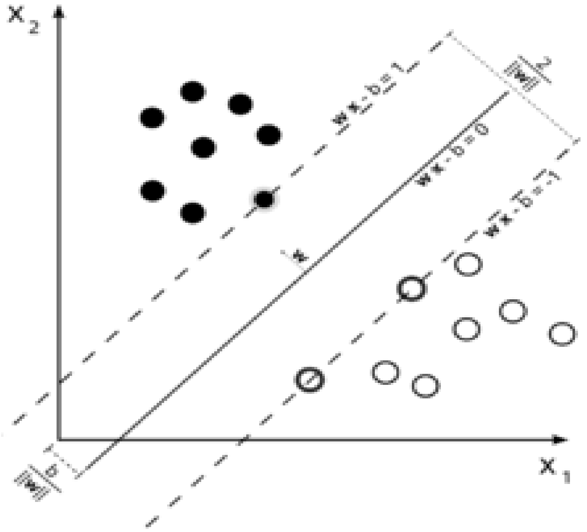


Fig.3. Maximum-margin hyper plane and margins for an SVM trained with samples from two classes.

Samples on the margin are called the support vectors.

Mathematically expressing linear SVM where the goals are to separate the data with hyperplane and extend it to non-linear boundaries .For mathematical calculations we have,

i. If Yi= +1; wxi + b ≥ 1

ii. If Yi= -1; wxi + b ≤ 1

iii. For all i; yi (wxi + b) ≥ 1

where x is a vector point, b is bass and w is weight and also a vector.

2.8 Twin SVM:

Twin support vector machine (TSVM), a useful extension of the traditional SVM, has become the current researching hot spot in machine learning during the recent years. For the binary classification problem, the basic idea of TSVM is to seek two nonparallel proximal hyperplanes such that each hyperplane is closer to one of the two classes and is at least one distance from the other. TSVM has lower computational complexity and better generalization ability, therefore in the last few years it has been studied extensively and developed rapidly. Consider the following originally proposed models for the binary classification problem with the training set as:

$$T = \{(x_1, +1), \dots ., , (x_{p+1}, -1), (x_{p+q}, -1)\}(1)$$

where

$x_i \epsilon R^n$, i= 1,…, p+q let l= p+q and

$$\text{A} = \ (x_1, \dots . x_p)^T \epsilon R^{p \times n},$$

$$\text{B} = \ (x_{p+1}, \dots . x_{p+q})^T \epsilon R^{q \times n}(2)$$

Unlike the standard SVM solving one QPP (Quadratic Programming Problem), TSVM constructs two smaller QPPs as given below:

$$\underset{w_+, b_+, \xi_-}{min} \frac{1}{2}(Aw_+ + e_+b_+)^T(Aw_+ + e_+b_+) + c_1 {e_-}^T \xi_-,$$

$$s.t - (Bw_+ + e\_b_+) + \xi_- \geq e_-, \xi_- \geq 0(3)$$

and

$$\underset{w_-, b_-, \xi_+}{min} \frac{1}{2}(Bw_- + e\_b_-)^T(Bw_- + e\_b_-) + c_2 {e_+}^T \xi_+,$$

$$s.t - (Aw_- + e_+b_-) + \xi_+ \geq e_+, \xi_+ \geq 0 \qquad (4)$$

to seek the pair of nonparallel hyperplanes

$$(w_+ \, .x) + b_+ = 0 \; and$$

$$(w_- . x) + b_- = 0 \qquad (5)$$

such that the positive hyperplane $(w_+ \, .x) + b_+ = 0$ is proximal to the positive class (measured by the quadratic loss $\frac{1}{2}(Aw_+ + e_+b_+)^T(Aw_+ + e_+b_+)$) and far from the negative class (measured by the hinge loss $\xi_- = \max\{0, e_- + (Bw_+ + e\_b_+)\})$ and vice versa for the negative hyperplane$(w_- . x) + b_- = 0$. $c_i, i = 1{,}2$ are the penalty parameters and e_+ and $e_$are the vectors of ones of appropriate dimensions. In order to get the solutions of the above QPPs, their dual problems

$$\underset{\alpha}{max} \left({e_-}^T \alpha - \frac{1}{2} \alpha^T G (H^T H)^{-1} G^T \alpha\right),$$

$$s.t \; 0 \leq \alpha \leq c_1 e_- \qquad (6)$$

and

$$\underset{\gamma}{max} \left({e_+}^T \gamma - \frac{1}{2} \gamma^T H (G^T G)^{-1} H^T \gamma\right),$$

$$s.t \; 0 \leq \gamma \leq c_2 e_+ \qquad (7)$$

are solved respectively, where

$$H = [Ae_+] \epsilon R^{p \times (n+1)},$$

$$G = [Be_-] \epsilon R^{q \times (n+1)}, \qquad (8)$$

Therefore the solutions of (3) and (4) can be obtained by

$$({w_+}^T, b_+)^T = -(H^T H)^{-1} G^T \alpha, (9)$$

$$(w_-{}^T, b_-)^T = -(G^TG)^{-1}H^T\gamma \qquad (10)$$

Thus an unknown point $x\epsilon R^n$ is predicted to the Class by

$$\text{Class}=arg_{k=-,+}^{\;\;min}|w_k.x+b_k| \qquad (11)$$

3. Results

The collected sample images are resized to size 256 X 256 which is the mostly used size by different researchers for their research works atdifferent noise levels. Then the resized images are contrast enhanced before they are segmented using k-means clustering and the segmented ROI is used to extract the features which TSVM uses to classify the disease in the query image provided as input.

3.1 Contrast Enhancement

The sample images after resizing to the standard 256 X 256 resolution size are denoised by using Median Filter Algorithm. Median filtering shown in fig 4 is very widely used in digital image processing because, under certain conditions, it preserves edges while removing noise.

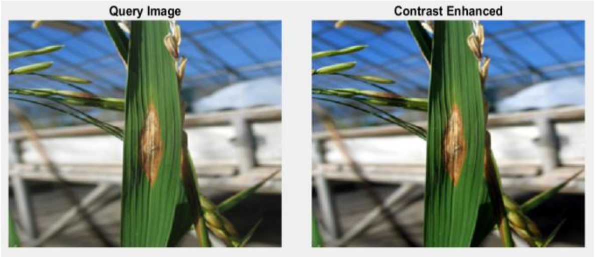


Fig.4. Contrast enhanced image using Median Filter Algorithm

3.2 Segmentation

Next, the contrast stretched (enhanced) images are segmented by using K means algorithm. In the process, the image submitted is segment into three clusters and the cluster containing the ROI (Region of Interest) is selected. ROI is a portion of an image that you want to filter or perform some other operation on. ROI is defined by creating a binary mask, which is a binary image that is the same size as the image that is to be processed. In the mask image, the pixels that define the ROI are set to 1 and all other pixels set to 0.

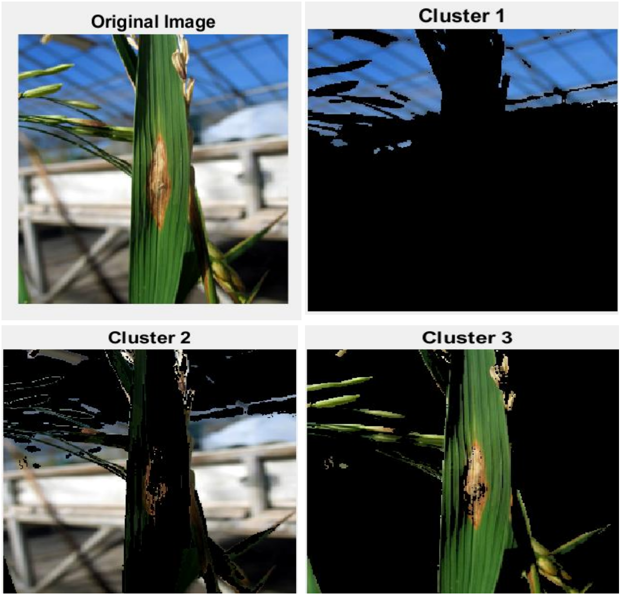


Fig.5. Segmentation using K means clustering

3.3 Feature Extraction

After selecting the appropriate ROI from the above listed images various features are extracted using GLCM matrix such as: Contrast, Energy, Entropy Homogeneity, RMS, Variance, Smoothness, Kurtosis, Skewness, Correlation and so on which will be used for the purpose of training the TSVM. These features will be stored in the database which in turn will be used while training the TSVM and the corresponding weight of the features will be stored in the database during the training phase. These features will later be used during classification phase.

Table 1: Features extracxted from GLCM of the image

S.N	Features	Value
1.	Mean	68.2172
2.	S.D.	83.4849
3.	Entropy	4.142624
4.	RMS	10.9867
5.	Variance	5880.14
6.	Smoothness	1
7.	Kurtosis	2.30266
8.	Skewness	0.86788
9.	IDM	255
10.	Contrast	0.451808
11.	Correlation	0.961456
12.	Energy	0.278708
13.	Homogeneity	0.935678

3.4 Training Data Set for Blast Disease and Bacterial Blight

100 images of blast affected rice leaf and 36 images of blight affected plants and 10 healthy images have been used to train and test the TSVM and the above mentioned features are extracted for each image which combine together to form the training data matrix and accuracy data matrices. The matrices are used for testing and validation of other test/query images.

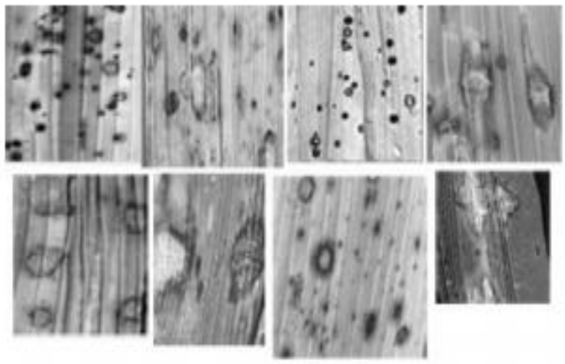
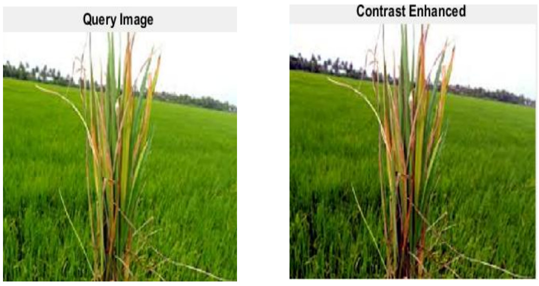


Fig.6. Images of infected leaves for test session (source:www.knowledgebank.irri.org/blast leaf - collar)

4. Testing and Validation

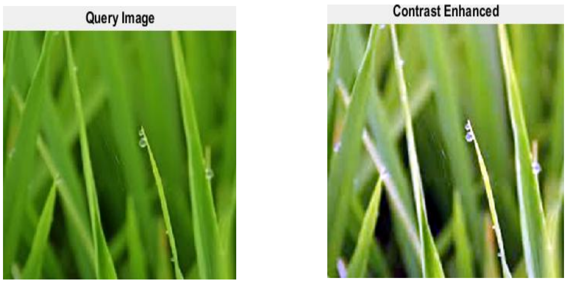
After the generation of training data matrix and accuracy data matrix, 36 images of rice leaf affected with blast disease are used for the testing and validation purposes. The images are loaded first and then the contrast enhancement is done at first. Next the segmentation of the contrast enhancement by k means clustering follows that helps us in selecting a particular ROI that likely has the disease affected part in the image. Feature extraction from the image is then performed and based on these features TSVM classifies the disease. The accuracy of disease classification in this system prototype is evaluated using 500 iterations.

However the system does classify adversely when the k mean clustering not done on the contrast enhanced image and we straight forward resume with the disease classification. The blight affected and healthy image given below were classified as blast. The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.



(a) (b)

Fig.7. a) i/p image with blight b) contrast enhanced classified as blast



(a) (b)

Fig 8. a) i/p image healthy b) contrast enhanced classified as blast

The disease classification without using k means clustering in Fig. 7 may have been affected be due to the ratio of blast portion to the image and in fig 8 due to the dew on the healthy plant might have caused the system prototype to classify the disease as bacterial blight. The healthy query image is predicted correctly by the system as healthy.

The prediction and classification of rice diseases as blast and bacterial blight have so far been correct when segmentation using k-means clustering is used in the system prototype. However, the system prototype significantly loses its credibility for correctly classifying the diseases when it is implemented without segmentation using k-means clustering. The system classifies correctly one healthy query image as healthy, while the other healthy query image as bacterial blight.

Clustering refers to a process of partitioning set of data or objects into a set of meaningful subclasses called clusters (Gonzalez, 2011). It is an unsupervised learning technique and this technique works well when there are finite set of clusters. The main goal of using clustering is to find similarities within a given dataset that uses finite set of data.

The problem in implementing the system prototype without segmentation using k-means clustering seems to be the input to the system. The implementation of the system without segmentation using k-means clustering uses the whole contrast enhanced image to extract the different features from the GLCM of the enhanced image which have very fluctuated values as we can see from the features table. On the other hand, using k-means clustering only desired ROI is fed as input to extract the features i.e. the partitioning into sub classes and only the meaningful ROI is given as input to the system. Hence the accuracy of classification is high. But the system without clustering takes the whole image and since the search area is larger, it fails to accurately classify the diseases for the same dataset as it wasused for the system with segmentation using k-means clustering. The table given below shows the different values of the GLCM features of 6 test input images (2 from each class: blast, blight and healthy) that were obtained during the testing phase.

Table 2: GLCM features for various test images

features	***Test image***					
	blast1.jpg	***blast1.jpg***	***blight1.jpg***	***blight2.jpg***	***healthy1.jpg***	***healthy2.jpg***
Mean	7.04	18.5	40.65	33.71	78.69	51.05
S.D.	30.93	52.5	67.88	67.04	57.79	43.77
Entropy	0.84	1.67	3.458	2.91	7.455	6.93
RMS	2.64	4.70	8.950	7.68	15.08	13.2
Variance	798.7	2302	4382	3639	664	540.5
Smoothness	0.99	1	1	1	1	1
Kurtosis	27.31	10.48	3.53	4.42	1.91	2.85
Skewness	4.88	2.90	1.41	1.74	0.2	0.69
IDM	255	255	255	255	255	255
Contrast	0.43	0.55	0.68	2.01	0.18	0.20
Correlation	0.62	0.86	0.91	0.73	0.9	0.86
Energy	0.83	0.73	0.43	0.45	0.26	0.24
Homogeneity	0.957	0.95	0.92	0.85	0.92	0.90

The GLCM features Table 2 above has 13 different features that were extracted from the test images and these values are now tested against the previous features that were used to train the TSVM during the training phase.

The features of the different class of images as shown in the table above present a strong agreement for different feature values of the images in the same class. The smoothness of all images is found to be near to or equal to 1, whereas the IDM value is found to be 255.

In the validation phase 15 query images were taken and tested 5 images from each class (blast, blight and healthy). The accuracy for blast was found to be 96.7742, 98.3871, 95.1613, 96.7742, and 97.6342. The mean accuracy was found to be 96.9462. The accuracy for blight was found to be 95.1633, 98.3871, 96.7742, 96.5782, and 97.6374. The mean accuracy was found to be 96.94081. The accuracy for blast was found to be 98.3871, 96.7742, 97.8832, 98.6754, and 96.6754. The mean accuracy is found to be 97.6967. The system accuracy was hence calculated as the mean of above three means and was calculated approximately as 97.1837.

Table3: Accuracy of classification for different test images by the system

Query image	**Accuracy**	**µ**	**System Accuracy**
blast1.jpg	96.7742	96.9462	97.1837
blast2.jpg	98.3871		
blast3.jpg	95.1613		
blast4.jpg	96.7742		
blast5.jpg	97.6342		
blight1.jpg	95.1633	96.9081	
blight2.jpg	98.3871		
blight3.jpg	96.7742		
blight4.jpg	96.5782		
blight5.jpg	97.6374		
healthy1.jpg	98.3871	97.6967	
healthy2.jpg	96.7742		
healthy3.jpg	97.8832		
healthy4.jpg	98.6754		
healthy5.jpg	96.7648		

Comparison with different existing systems

Amit Kumar Singh et al have also proposed a rice disease classification system with only the blastdisease in "Classification of Rice Disease Using Digital Image Processing and SVM Classifier" (Singh,2015). The system developed was able to predict and classify blast disease with and accuracy of 82%.

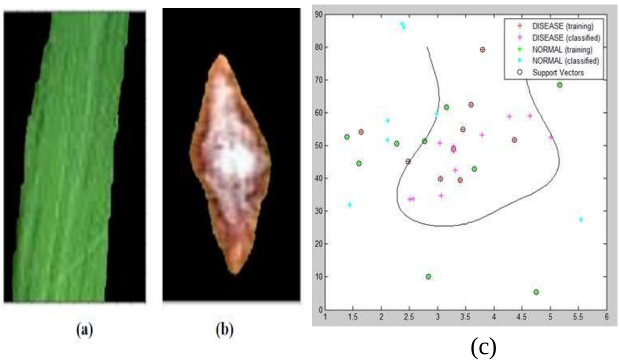


Fig.9. segmented image (a) normal image (b) infected image and (c) is the output of the SVM classification

R.Rajmohan, M.Pajany et al in (R.Rajmohan,2018) have developed a Deep CNN & SVM classifier and compared the output against k-means & fuzzy logic system and KNN & SVM classifier. The results they found is given in the table below:

Table 4: Comparison of K-means& fuzzy logic vs. KNN & SVM vs. Deep CNN & SVM classifier

Disease	Accuracy %		
	K-means & Fuzzy Logic	KNN & SVM classifier	Deep CNN & SVM classifier
Blast	80	84	92
Brown spot	80	84	88
Bacterial Leaf Blight(BLB)	76	84	92
Sheath Blight	76	80	88
False Smut	72	80	84
Udbatta	72	76	80
Root Knot nematode	80	88	92
White tip nematode	76	80	84

Now, it can be seen that our system prototype has classification accuracy more than the above mentioned systems. Our system prototype has presented an accuracy of over 95% in most of the cases during testing as shown in the section above.

5. Conclusion & Recommendations:

In this paper, a system prototype was used to classify rice leaf's Blast disease and Bacterial blight using K-means clustering for segmentation and Twin Support Vector Machine (TSVM). It was tested on rice leaves to detect if they are affected by blast disease or bacterial blight. With 146 total number of rice leaf's images (100 blast affected, 36 blight affected and 10 healthy ones); training and testing of the system was performed. Once, the system was trained, the detection of disease or not by the trained TSVM was high with accuracy of more than 95%. Accuracy can surely be increased if system can be trained with high number of dataset of rice leaf affected by blast disease and blight disease and accuracy more than the current scenario can be obtained. This work can be extended to detect other diseases by training the TSVM with various other disease datasets.

References

[1] Dean, R. A.; Talbot et al. (2005). "The genome sequence of the rice blast fungus Magnaporthegrisea". Nature. 434 (7036): 980–6.

[2] Gonzalez Rafael C., Woods Richard E. (2011) Digital Image Processing, Pearson Education in South Asia.

[3] IRRI Rice Diseases factsheets Archived October 14, 2013, at the WaybackMachine,knowledgebank.irri.org.

[4] Rajmohan R., Pajany M., R.Rajesh, Raghu RamanD., PrabuU. (2018)"Smart Paddy Crop Disease Identification and Management Using Deep Convolution Neural Network and SVM Classifier" International Journal of Pure and Applied Mathematics special issue.

[5] Savada, Andrea Matles (1991). "Nepal: A Country Study:Agriculture". Washington GPO for the Library of Congress

[6] Singh A.K.,Rubiya .A, B.Senthil Raja(2015), "Classification Of Rice Disease Using Digital

ImageProcessing And SVM Classifier" International Journal of Electrical and Electronics Engineers

[7] UN Food and Agriculture Organization, Corporate Statistical Database (FAOSTAT). 2017)"Crops/Regions/World list/Production Quantity (pick lists), Rice (paddy), 2014".

[8] YingjieTian, Zhiquan Qi(2014), “Review on: Twin Support Vector Machines”, published online 2015, Springer-Verlag Berlin Heidelber