

# Effective Feature Extraction and Classification Method for Potato Foliar and Tuber Disease Detection using Machine Learning

Megha Rani Raigonda<sup>1</sup>, Sujatha P. Terdal<sup>2</sup>

<sup>1</sup>Research Scholar, Department of CSE, PDA College of Engineering, Kalaburagi, Karnataka, India

<sup>2</sup>Professor & Head, Department of CSE, PDA College of Engineering, Kalaburagi, Karnataka, India

**Abstract:** *The Indian economy heavily depends on agriculture. Consequently, it is crucial to detect diseases in the agricultural sector. There is a necessity to identify the disease at the initial stage because farmers struggle to produce crops properly because of several plant diseases. The potato is a vegetatively propagated crop. It is a host to many bacterial fungal and viral diseases. A potato crop can be infected by more than 30 plant viruses, a viroid, and phytoplasmas. Viral diseases are of major concern nowadays in potato crops. Once discovered in the field, virus-infected plants result in declassification or even rejection of the seed lots, which results in a monetary loss. The viruses Potato Virus A (PVA), Potato Virus X (PVX), Potato Virus S (PVS), Potato Virus Y (PVY), Potato Virus M (PVM), Potato leaf roll virus (PLRV), and Tomato leaf curl New Delhi virus are known to infect potatoes in India (ToLCNDV), Potato spindle tuber viroid (PSTVD) and Groundnut bud necrosis virus (GBNV). PVM, PVY, PVA, PVX, and PVS occur commonly. Nowadays many computer vision technologies like machine learning, deep Learning are employed in building a prediction model for the effective, rapid, and accurate detection of potato plant disease. In the proposed work the viral disease considered is Potato Leaf Roll Virus (PLRV), Mosaic Virus, Leaf curl, and tuber diseases Potato Tuber Viroid Disease (PSTVD), Potato Virus Y (PVY) Tuber cracking. The foliar image is initially resized to 256\* 256, applied contrast enhancement and then filters are applied for denoising removing high frequency, and smoothing the image then the segmentation using Canny Edge Detection is applied to the blurred image to accurately detect the edges of the leaf and then the suitable features are extracted. The disease is classified using classification methods like Support Vector Machine (SVM) and Random Forest. The Random Forest classifier outperforms all other classifiers and produces a classification accuracy of 98.12%.*

**Keywords:** Gray level Cooccurrence matrix (GLCM), Support Vector Machine (SVM), Random Forest (RF), Global Features (GF), Viral Disease

## 1. Introduction

India is second in the world for the production of crops like rice, wheat, lentils, and spices. Due to losses that occur during cultivation, crop production in India only yields between 30 and 60 percent [1]. The crop's yield and quality determine a farmer's financial situation. By looking at a leaf's shape, area, color, texture, etc., researchers can identify a particular plant type. Plants play a major role in regulating the environment and the climate. Some plants are also raised for the generation of biofuels since they are less harmful than other toxic gases. A wide range of viral infections can threaten a crop of potatoes, causing a variety of symptoms. One of the most common and significant viruses affecting potatoes worldwide is PVY (genus Potyvirus, family Potyviridae), which is also in the top ten most harmful plant viruses [2] [3]. Early plant disease diagnosis is highly advised for enhancing the farmer's financial situation. The food quality is lowered by leaf diseases. [4] Monitoring crops from remote places for detecting diseases play a key role in successful cultivation. In the traditional method, naked eye observation is done with a large team of experts for disease detection, which is very expensive [5].

In this paper, the proposed methodology uses different computer vision and artificial Intelligence methodology is used for automatic disease detection [6]. The objective of this research is the detection and classification of potato foliar and tuber diseases. This objective is achieved in four stages. Acquiring image from the captured image

database, Pre-processing the image by resizing, applying filters, denoising the image, converting RGB to grayscale for fast access and less storage, Segmenting the image into clusters using the k-means clustering method, Extracting the features using Gray Level Cooccurrence matrix (GLCM), shape, and colored features from the segmented clusters and with the help of extracted features the classifier classifies the disease accurately.

Image processing starts with image acquisition, which is the process of collecting images from the database. These images are in the form of .png, .jpeg, .gif, and so on. After acquiring the image, it can be segmented into different parts that are clustered by using the k-means clustering method. In this RGB color space is transformed into Hue, Saturation, and value (HSV) color space which shows the color distribution.

The rest of the paper is arranged as follows in section 2 literature review is presented, in section 3 illustrates proposed methodology is illustrated, in section 4 results and discussion are discussed and finally, section 5 concludes the work.

## 2. Literature Review

Modern research is heavily focused on the use of computers and information technology in the agriculture sector. Various researchers have focused on different types of diseases affected by bacteria, fungi, and viral to the whole plant [7]. Khirade et al. [8] utilized SVM for

classification and hue-saturation thresholding for picture segmentation in their technique. Wang et al. [9] A dataset with two classes for grape and two classes for wheat leaves, and 50 features were retrieved from it. For disease classification, back propagation neural networks were combined with K-means segmentation and PCA dimensionality reduction. Orillo et al. [10] employed a back propagation neural network with statistical features from a color picture dataset with the diseases bacterial leaf blight, brown spot, and blast. Suman et al. [11] worked with bacterial leaf blight, brown spot, narrow brown spot, and rice blast diseases. For illness segmentation and SVM for classification, they used an 8-connected component analysis. Their model had a 70% accuracy rate after they retrieved numerous color and shape features. Various color and shape features were also extracted as features in the work of Chawathe [12]. Sumathi, et al. [13] introduced a feature fusion method that fused the features that were produced using the Gabor filter in the frequency domain with the extraction of features using edges. To assess the accuracy of the retrieved features, 10-fold cross-validation training was used, followed by tests using CART and RBF classifiers. Deep convolutional neural networks were recently employed by Sharada Prasanna Mohanty et al. to identify 26 illnesses and 14 crop types [14]. A strategy that combines relatively easy techniques that made use of shape and texture features was introduced in the research work supplied by Beghin, et al., [15]. The contour signature from each leaf is extracted using the shape-based technique, after which the differences between the leaves are computed. Analysis of the macro-texture of the leaf is done using the edge gradient orientations. With the aid of an incremental classification algorithm, which offers an accuracy of 81.1%. The researcher in [16] uses convolution Four convolutional layers, each with 32, 16 or 8 filters, make up the architecture of a neural network. The proposed model's acquired training accuracy is 99.47%. The results are categorized once the ratio of leaf area and disease spot quotients has been calculated. With the use of a hybrid intelligent system, grape leaf disease is recognized from color imaging [17]. Self-organizing maps

and back-propagation neural networks were employed. Utilizing image processing and segmentation, leaf diseases can be identified from simple leaf images. [18] By combining Otsu's approach with k-means clustering, a defective leaf region is identified, allowing for the identification of the best course of action. Decision trees, the neural networks, Naive Bayes theorem, Random Forest, and K-Means algorithms were developed for the classification of leaf diseases in [19] utilizing variables such as wilting, dryness, size, and shape.

In this proposed system global features are used to classify the plant diseases and segmented the foliar and tuber for regions of interest using the canny edge algorithm and green masking algorithm. Comparative results of different machine learning algorithms are also presented in this study.

### 3. Proposed Method

The proposed work involves different types of viral diseases affecting foliar and tuber of potato plants. The dataset includes 1, 709 images of diseased and healthy tubers and leaves. The leaf and tuber images are first put through pre-processing, which includes scaling them down to a standard size and turning them into grayscale images then segmentation like canny edge detection and green masking is used for extracting the region of interest. In training, the region of interest of leaf and tuber images is trained by applying Gray Level Cooccurrence Matrix methodology next the extracted features are stored in a csv file. The features extracted in this study are texture using GLCM, shape features using Hu Moments, and color histogram for extracting color features. The classifier used for this work is Support Vector Machine (SVM) and Random Forest (RF). Based on the above features the classifier works accurately and classifies various diseases on both foliar and tuber of the potato crop.

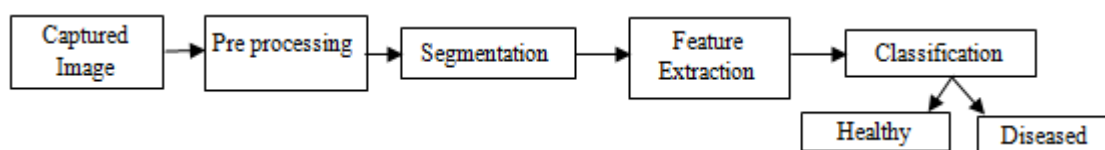


Figure 1: Block diagram of proposed methodology

#### a) Feature extraction

Feature extraction is a crucial step to recognizing the pattern in leaf disease detection. Accurate and minimal features led to fast and efficient classification of the disease. The features are color, shape, and texture.

#### Global Feature Descriptor (GFD)

Using global feature descriptors, features from images are retrieved. GFD processes the complete leaf or tuber image to process and extracts features rather than just the specific region of interest from an image.

#### Hu Moments:

Hu Moments measure how a shape of an object appears in an image. The shape of the object is typically denoted by it. It computes the image's moments after converting an RGB image to a grayscale version. Once that is done, shape feature vectors are returned.

Shape feature Hu moments are considered where image moments are calculated by the below formulae.

$$M_{ij} = \sum_x \sum_y x^i y^j I(x, y)$$

where  $i$  and  $j$  are integers (e. g., 0, 1, 2 ...). These moments are often referred to as raw moments to distinguish them from central moments.

For the Color feature, the histogram is generated for the HSV image where the color distribution is known accurately.

```
image = cv2.cvtColor(image, cv2.COLOR_BGR2HSV)
hist = cv2.calcHist([image], [0, 1, 2], None, [bins, bins, bins], [0, 256, 0, 256, 0, 256])
cv2.normalize(hist, hist)
```

**Gray Level Co-occurrence matrix Haralick Texture:**

For the purpose of extracting texture features, the Haralick Texture feature descriptor is employed. When using the Haralick feature descriptor to extract texture features from an image, it is first necessary to transform the colour image to grayscale. The Gray Level Co-occurrence Matrix is the primary idea used in computing the Haralick texture feature (GLCM).

**Energy**

The GLCM's sum of squared elements is provided by energy. Its values range from 0 to 1.

$$Energy = \sum_{i,j} P(i,j)^2$$

**Contrast**

A pixel's intensity contrast with its neighbour over the entire image is measured as contrast. Contrast is 0 for a picture that is "constant" (has no variation).

$$Cont. = \sum_i \sum_j |i - j|^2 p(i,j)$$

**Homogeneity**

The degree to which GLCM elements are distributed near to the diagonal is measured by local homogeneity. Homogeneity is equal to one for a diagonal GLCM.

$$Homog. = \sum_i \sum_j \frac{1}{1 + |i - j|^2} p(i,j)$$

**Correlation**

Correlation measures how closely connected one pixel is to another throughout the entire image. A fully positively or negatively correlated image has a value of 1 or -1, and a constant image has a value of infinity.

$$Correl. = \sum_i \sum_j \frac{(i - \mu_i)(j - \mu_j) p(i,j)}{\sigma_i \sigma_j}$$






**Dissimilarity**

Distance between adjacent pairs of objects (pixels) in the region of interest is measured by dissimilarity.



$$Dissimilarity = \sum_i \sum_j |i - j| p(i,j)$$




For texture features, Haralick texture is used which are normalized gray level Cooccurrence Matrix (GLCM). Where RGB is converted to grayscale and then given to the Haralick function defined in the mahotas package. Here generally correlation, homogeneity, contrast, energy, and dissimilarity are considered to extract the texture features of a leaf. In Table 1 and Table 2 the GLCM properties of both leaf and tuber are shown.

**Table 1: GLCM for texture features of potato foliar**

S. No.	Foliar Image	Contrast	Correlation	Energy	Homogeneity	Entropy
1		1.546	0.473	0.997	0.998	5.42
2		0.797	0.187	0.999	0.999	4.52
3		0.484	0.412	0.999	0.999	4.60
4		0.943	0.248	0.998	0.998	4.78
5		1.022	0.653	0.999	0.999	4.51

**Table 2: GLCM for texture features of potato tuber**

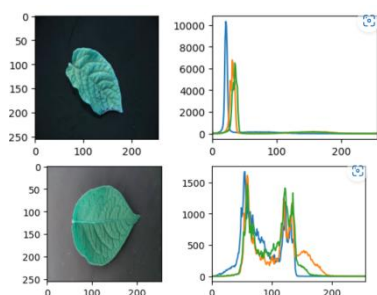
S. No.	Foliar Image	Contrast	Correlation	Energy	Homogeneity	Entropy
1		145.695	0.882	0.921	0.934	9.451
2		492.460	0.883	0.758	0.809	12.166

3		524.196	0.718	0.920	0.927	9.379
4		220.278	0.921	0.823	0.883	14.399
5		283.302	0.906	0.825	0.886	14.456

**Color Histogram:**

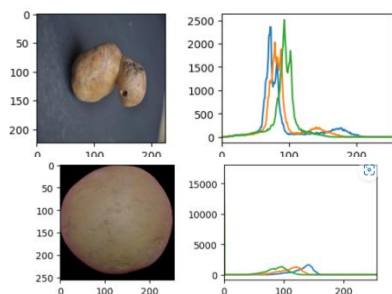
It displays the amount of pixels in each colour range and the colour distribution in the input image. The colour intensity of a picture is calculated using the colour histogram descriptor.

Fig.1 show the graphical view of the color histogram of the healthy and color histogram of a diseased potato leaf is shown in Fig.2.



**Figure 1:** Color histogram of a diseased mosaic leaf  
**Figure 2:** Color histogram of a healthy potato leaf

The graphical view of color histogram of healthy and diseased tuber is shown in fig.3 and fig.4.



**Figure 3:** Color histogram of a PSTVD leaf  
**Figure 4:** Color histogram of a healthy tuber

**F. Classification using different classifiers**

The extracted feature vectors are then used to train different classifiers and the results were analyzed.

**Support Vector Machine:** SVMs are very efficient in high-dimensional spaces and generally are used in classification problems. This is a novel development in the field of machine learning, and it is used for a variety of texture classification difficulties as well as other pattern recognition issues. SVMs use a small selection of training points in the decision function, which makes them well-liked and memory-efficient. Given labeled training data, SVM outputs an optimal separating hyperplane. A hyperplane is used to classify fresh data points. Some SVM classifier parameters need to be adjusted in order to

increase the accuracy of the algorithm. One of the parameters is the kernel which defines whether separation should be linear or non-linear. The importance of misclassifications is controlled by the regularisation parameter (lambda). SVM provides a quadratic optimization problem that seeks to maximise the margin between both classes and reduce the number of misclassifications. The value of the linear kernel and regularisation parameter relies on the training samples. If it guarantees the least amount of misclassification of training instances, a higher value of regularisation chooses a tiny margin of hyperplane. In this classifier SVC (random\_state=seed) is used where the seed value is any random integer, the train\_test\_split will return the same results for each execution. Multiclass classification is also possible and is essentially constructed by different two class SVMs to handle the issue, either by employing one-versus-all or one-versus-one. The highest output function or the most votes, respectively, determines the winning class next [20].

**Random Forest:** A decision tree-based supervised learning technique is called random forest. It is used widely in Classification and Regression problems. It builds a forest of decision trees. Many trees fit into a random forest classifier. Data with labels makes up the leaf node in the tree. The decision tree commonly leads to overfitting. Unlike decision trees, random forests handle both numerical and categorical data and overcome the drawback of overfitting their training data set. Each tree in the forest receives the extracted feature vector as an input vector, which produces a decision rule. So, a class is chosen by the trees. The forest selects the class based on the majority of tree votes.

RandomForestClassifier (n\_estimators =num\_trees, random\_state=seed))

The classifier sets the random seed value as 9. Where the results do not change for random samples.

**4. Experimental Results**

The proposed system is evaluated 1, 600 images of both healthy and diseased leaves of potato. We have used 80%of the images for training and 20% of images for testing. For tuber our model uses 500 images where 400 images are for training and 100 images for testing. The classification method used in this work is Support vector Machine (SVM) and Random Forest. For 10-fold cross validation, the classifiers mean and standard deviation is calculated and the results are shown below in fig.5.

```

# 10-fold cross validation
for name, model in models:
    kfolds = KFold(n_splits=10, random_state=seed, shuffle=True)
    cv_results = cross_val_score(model, trainDataGlobal, trainLabelsGlobal, cv=kfolds, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

# boxplot algorithm comparison
fig = pyplot.figure()
fig.suptitle('Machine Learning algorithm comparison')
ax = fig.add_subplot(111)
pyplot.boxplot(results)
ax.set_xticklabels(names)
pyplot.show()

RF: 0.957031 (0.013189)
SVM: 0.916406 (0.020886)
    
```

Figure 5: Comparison of classification algorithms

**Random Forest**

Mean: 0.957 and Standard Deviation: 0.013

**SVM**

Mean: 0.916 and Standard Deviation: 0.020

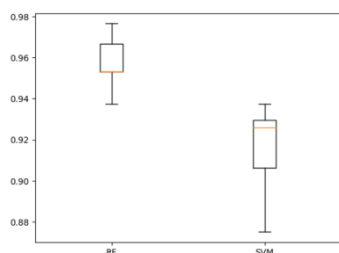


Figure 6: Graphical representation for leaf

**Machine Learning Algorithm Comparison**

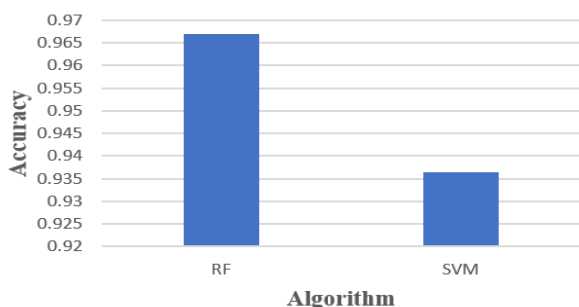


Figure 7: Graphical representation for tuber

The classification algorithms used in this work are SVM and Random Forest. The dataset is divided into 80% training and 20% testing data. Therefore, on the test data i.e., 320 images the Random Forest classifier is giving highest accuracy of 98.1% in accurately classifying the disease. The confusion matrix is shown in the below.



Figure 8: Confusion matrix of the Random Forest classifier for foliar

In the above figure it is clear that 154 images are correctly classified as diseased, 160 as healthy and wrongly predicted is 6. The precision and recall are calculated from the confusion matrix. Based on precision and recall, F1-score is calculated for both the healthy and diseased classes are shown in table 2 & table 3. Accuracy is not only the metrics to evaluate the performance of the classifier. The tuber confusion matrix is shown in Fig.9, classified on test data 100 is 55 diseased and healthy data is 43 and wrongly predicted is 2. Classification accuracy alone is typically not enough information to make the decision. The F1-score value is a harmonic mean between precision and recall. The precision value should be greater than the recall then the F1-score value will also be larger value.

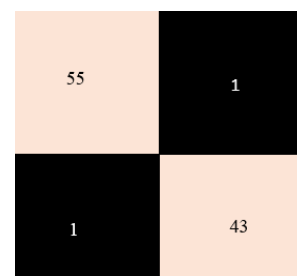


Figure 9: Confusion matrix of the Random Forest classifier for tuber

The precision, recall, F1-score, and support are also the metrics used to evaluate the classifier performance the values are shown in Table 3 and Table 4.

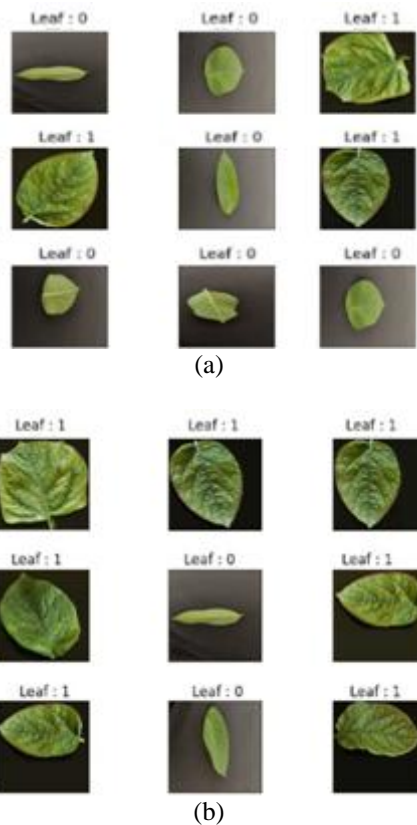
Table 3: Accuracy metrics for calculating F1 score on leaf test data

Class	Precision	Recall	F1-score	Support
0 (Diseased)	0.99	0.97	0.98	158
1 (Healthy)	0.98	0.99	0.98	162
Accuracy			0.98	320

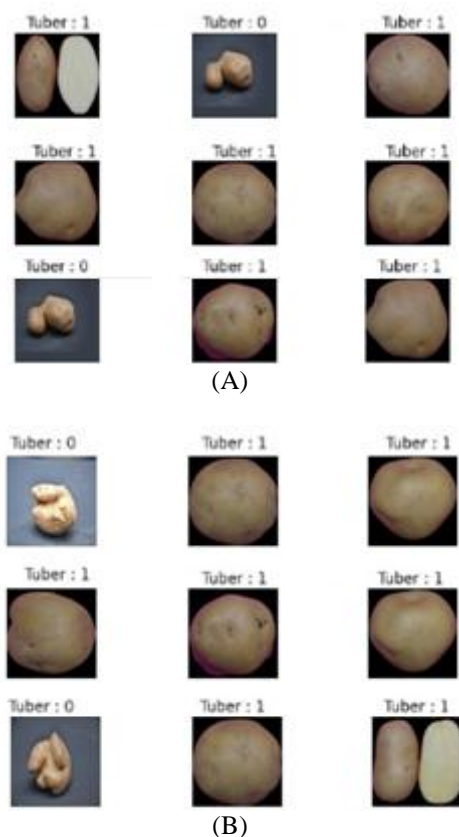
Table 4: Accuracy metrics for calculating F1 score on tuber test data

Class	Precision	Recall	F1-score	Support
0 (Diseased)	0.98	0.98	0.98	56
1 (Healthy)	0.97	0.97	0.97	44
Accuracy			0.98	100

The Random Forest classifier is given leaf and tuber dataset as input which classifies the foliar and tuber either under class 0 or class 1. Class 1 resembles unhealthy and class 0 resembles healthy in classifying potato foliar as shown in fig.9 (A&B). In case of tuber the classifier classifies class 1 as healthy and class 0 as unhealthy as shown in fig.10 (A&B).



**Figure 9:** Classified leaf images after applying Random Forest classifier 1-Unhealthy 0-Healthy



**Figure 10:** Classified tuber images after applying Random Forest classifier 1-Healthy 0-Unhealthy

## 5. Conclusion

In this proposed prediction model, image processing and machine learning techniques are used to detect the diseases affecting the plant leaves. The proposed method takes an RGB image as the input and applies pre-processing methods such as resizing the image to standard size 256\*256, contrast enhancement, grayscale conversion, and gaussian filters are applied. After applying the filter, that image is segmented using Canny Edge Detection. The features considered for disease detection are color, shape, and texture features. The effective features are given to the classifier like SVM and Random Forest for accurately classifying the disease. The Random Forest classifier outperforms the other two classifiers with a greater accuracy of 98.1% in the identification and classification of potato leaf disease. In future, dataset must be increased and different types of diseases in potato crop can be employed and deep learning can be employed for accurately detecting the disease.

## References

- [1] Sukhvir Kaur, Shreelekha Pandey, Shivani Goel, Semi-automatic leaf disease detection and classification system for soybean culture, ISSN 1751-9659, IET Image Processing.
- [2] Scholthof, K. B. G., Adkins, S., Czosnek, H., Palukaitis, P., Jacquot, E., Hohn, T., et al. (2011). Top 10 plant viruses in molecular plant pathology. *Mol. PlantPathol.*12, 938–954. Doi: 10.1111/j.1364-3703.2011.00752. x.
- [3] Valkonen, J. P. T. (2007). "Viruses: Economical Losses and BiotechnologicalPotential, " in *Potato Biology and Biotechnology*, eds D. Vreugdenhil, J. Bradshaw, C. Gebhardt, F. Govers, D. K. L. ackerron, M. A. Taylor, (San DiegoCA: Elsevier Science), 619–641.
- [4] Syafiqah Ishak, Mohd Hafiz Fazalul Rahiman, Siti Nurul Aqmariah Mohd Kanafiah, Hashim Saad, Leaf disease classification using artificial neural network, Syafiqah Ishak et al. / *Jurnal Teknologi (Sciences & Engineering)* 77: 17 (2015) 109114.
- [5] Arti N. Rathod, Bhavesh A. Tanawala, Vatsal H. Shah, Leaf Disease Detection Using Image Processing and Neural Network, *IJAERD*, Volume 1, Issue 6, June 2014, e-ISSN: 2348 4470.
- [6] Jimita Baghel, Prashant Jain, K-Means Segmentation Method for Automatic Leaf Disease Detection, *IJERA*, ISSN: 2248-9622, Vol.6, Issue 3, (Part-5) March 2016, pp.83-86.
- [7] H. B. Prajapati, J. P. Shah, and V. K. Dabhi, "A survey on detection and classification of rice plant diseases, " 2016 IEEE Int. Conf. Current Trends in Advanced Computing (ICCTAC), 2016, pp.1-8.
- [8] S. D. Khirade and A. B. Patil, "Plant disease detection using image processing, " *Int. Conf. on Computing Co mmunication Control and Automation (ICCUBEA)*, IEEE, 2015, pp.768-771.
- [9] H. Wang, G. Li, Z. Ma, and X. Li, "Image recognition of plant diseases based on Backpropagation Networks, " 5th Int. Congress on

- Image and Signal Processing (CISP), IEEE, 2012, pp.894-900.
- [10] J. W. Orillo, J. D. Cruz, L. Agapito, P. J. Satimbre, and I. Valenzuela, "Identification of diseases in rice plant (*Oryza sativa*) using Backpropagation Artificial Neural Network, " Int. Conf. on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), IEEE, 2014, pp.1-6.
- [11] T. Suman and T. Dhruvakumar, "Classification of paddy leaf diseases using shape and color features, " Int. Journal of Electrical and Electronics Engineers, vol.7, no.1, pp.239-250, 2015.
- [12] S. S. Chawathe, "Rice disease detection by image analysis, " 10th Annual Computing and Communication Workshop and Conf. (CCWC), Las Vegas, NV, USA, 2020, pp.0524-0530.
- [13] C. S. Sumathi and A. V. Senthil Kumar, "Edge and Texture Fusion for Plant Leaf Classification", International Journal of Computer Science and Telecommunications, Vol 3, Issue 6, June 2012, pp.6-9.
- [14] Mohanty Sharada P., Hughes David P., Salathé Marcel, "Using Deep Learning for Image-Based Plant Disease Detection, " Frontiers in Plant Science, Vol 7 (2016) DOI: 10.3389/fpls.2016.01419.
- [15] T. Beghin, J. S. Cope, P. Remagnino, & S. Barman, "Shape and texture based plant leaf classification", Advanced Concepts for Intelligent Vision Systems (ACVIS), Vol 6475, 2010, pp.45-353.
- [16] Mohit Agarwal et. al., "Potato Crop Disease Classification Using Convolutional Neural Network", Smart Systems and IoT: Innovations in Computing Proceeding of SSIC 2019.
- [17] Meunkaewjinda, P. Kumsawat, K. Attakitmongcol & A. Srikaew [2008] "Grape leaf disease detection from color imagery system using hybrid intelligent system", proceedings of ECTICON, 2008, IEEE, PP-513-516.
- [18] S. Maity et al., "Fault Area Detection in Leaf Diseases Using K-Means Clustering, " 2018 2<sup>nd</sup> International Conference on Trends in Electronics and Informatics (ICOEI), 2018, pp.1538-1542, doi: 10.1109/ICOEI.2018.8553913.
- [19] G. PremRishiKranth, HemaLalitha, LaharikaBasava, AnjaliMathurh: Plant disease prediction using machine learning algorithms. International Journal of Computer Applications 18 (2) (2018).
- [20] Er. Varinderjit Kaur, Dr. Ashish Oberoi, " WHEAT DISEASE DETECTION USING SVM CLASSIFIER", Journal of Emerging Technologies and Innovative Research, Vol.5 Iss.8, (2018), PP.779-788.