

Plant Leaf Disease Detection Using Computer Vision Techniques and Machine Learning

Kalpesh Joshi^{1,*}, Rohan Awale², Sara Ahmad³, Sanmit Patil⁴, and Vipul Pisal⁵

¹⁻⁵Department of Electronics Engineering, Vishwakarma Institute of Technology, Pune, India

Abstract. Agriculture production is extremely important in today's economy because disease development in plants is relatively common, early detection of disease in plants is critical in the agriculture field. The automatic finding of such early-stage disease detection is helpful as it decreases a great effort of supervising in large farmhouses of yields. Using digital image processing and machine learning algorithms, this paper presents a method for detecting plant disease. The disease detection is done on the yields' various leaves. The presented system for plant disease detection is simple and computationally efficient which requires less time for prediction than other deep learning-based approaches. The accuracies for the various plant and leaf diseases are calculated and presented in this paper.

1 INTRODUCTION

Developed countries are now putting more money into agriculture. Early detection of plant diseases is critical to avoid sufferers during the harvest. Different lesions on the leaf can be seen in the early stages of the plant disease. Physically detecting plant diseases is extremely difficult [1] It necessitates a significant amount of effort, knowledge of plant diseases, and an inordinate amount of time. The use of image processing aids in the accurate detection of plant diseases. The strategies for detecting plant diseases using their leaf pictures are presented in this paper. Image processing is a method of performing a few operations on a photograph to obtain a better photograph or extract a few useful records from it. A subset of Artificial Intelligence, machine learning (ML), works automatically or gives instructions to complete a task. Machine learning's main goal is to comprehend training data and fit that data into models that should be useful to people. Machine learning aids in the automatic detection of plant diseases [2]. It has aided good decision-making and forecasting due to the large amount of data generated. The color of the leaves, the amount of damage to the leaves, and the area of the unhealthy plant leaf are used to classify them. The system shown here is useful for monitoring large crop fields. The proposed solution for plant disease detection is less computationally expensive and takes less time to predict than other deep learning approaches because it uses statistical machine learning and image processing algorithms.

The remainder of the paper is divided into three sections. Section 2 shows similar work that has been done. Section 3 presents the proposed system. Section 4 elaborates on the experimental findings, and Section 5 brings the paper to a close.

2 LITERATURE REVIEW

Kothawale, et al. presented a system for detecting and classifying grape leaf disease using a support vector machine (SVM) classifier [3]. The leaf area is obtained using histogram thresholding, Grey Level Co-occurrence Matrix (GLCM) is employed for the feature extraction. They classified each image as healthy or affected and achieved an accuracy of 89.90%.

Pavithra, et al. implemented a method for rice leaf disease detection [4]. The shape and color-based features were extracted using Scale Invariant Feature Transform (SIFT). They have performed experimentation using SVM and Artificial Neural Network (ANN) classifiers and achieved accuracies of 95.2% and 92.2% respectively.

Islam et. al. presented a Multiclass Support Vector Machine-based method for detection of Potato Diseases[5]. Their experimentation performed on the public database 'Plant Village' yielded an accuracy of 95%.

* Corresponding author: rohan.awale18@vit.edu

Shima Ramesh et. al. implemented a technique using Random Forest for plant disease detection [6]. They created the dataset and extracted the features using the Histogram of an Oriented Gradient (HOG). The final classification was performed using different classifiers yielding maximum accuracy of about 70%.

Jagan Mohan et.al. presented a system for the detection and recognition of Paddy Plant Leaf diseases [7]. The disease was identified with an accuracy of 83.33 percent using Haar-like features and the Ada-Boost classifier. The SIFT feature and classifiers k-Nearest Neighbour (k-NN) and SVM are used to identify the paddy plant disease type. The accuracy rate for disease recognition was 91.10 percent and 93.33 percent, respectively.

John William et. al. identified three common diseases on rice plants using Back Propagation Artificial Neural Network [8]. The basic statistical and color features were extracted for detection of the disease using Neural Networks. They claimed an accuracy achieved of 100%.

Kusumo et. al. proposed a method for the disease detection of corn-plant [9]. In this technique, the features based on color, SIFT, speeded-up robust features (SURF), and object detectors such as HOG are utilized. They performed the performance analysis using various classifiers and achieved maximum accuracy of 87%.

H. Sabrol and S. Kumarn presented a decision tree classifier-based method for detecting tomato plant disease[10]. They demonstrated the detection of six basic diseases using statistical features, with a 78 percent overall accuracy.

3 IMPLEMENTATION

3.1 Dataset

The disease detection experiment was carried out with the help of a publicly available dataset called Plant Village, which was curated by Sharada P. Mohanty et al. [2]. As shown in Table 1, the dataset contains 87000 RGB images of healthy and unhealthy plant leaves divided into 38 classes.

Table 1. Specifications for Datasets

Plant	Disease Name & its Types	No. of images
Apple	Healthy	2008
	Scab	2016
	Black rot	1987
	Cedar apple rust	1760
Corn	Healthy	1859
	Cercospora leaf spot gray leaf spot	1642
	Common rust	1907
	Northern Blight	1908
Grapes	Healthy	1692
	Black rot	1888

	Esca (Black Measles)	1920
	Leaf blight (Isariopsis Leaf Spot)	1722
	Healthy	1824
Potato	Early blight	1939
	Late blight	1939
	Healthy	1926
Tomato	Bacterial spot	1702
	Early blight	1920
	Late blight	1851
	Leaf Mold	1882
	Septoria leaf spot	1745
	Spider mites	1741
	Two-spotted spider mite	
	Target Spot	1827
	Tomato Yellow Leaf Curl Virus	1961
	Tomato mosaic virus	1790

3.2 Methodology

The presented system for leaf disease detection is implemented using major steps as shown in Fig. 1. The image acquisition step captures the image of the plant, then these images are preprocessed for making these images suitable for further process. The image segmentation step performs the localization of the disease pattern-like areas for further steps. The feature extraction step extracts the features related to the disease for classification and identification purposes.

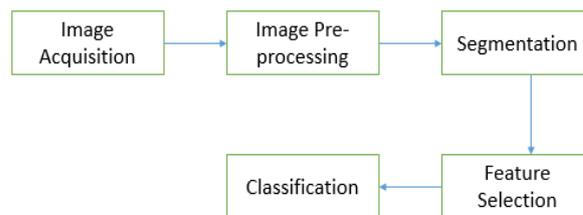


Fig. 1. Proposed System Flow

3.2.1 Pre-processing

In any computer vision-based system, data preprocessing is essential. The preprocessing steps for each image are depicted in Figure 2. Before extracting features, background noise should be removed to ensure precise results. The RGB image is converted to greyscale first, and then the image is smoothed with a Gaussian filter. After that, Otsu's thresholding algorithm is used to binarize the image. The morphological transform is then used on the binarized image to fill in the small holes in the foreground. Following foreground detection, a bitwise AND operation on the binarized image and the original color image yields the RGB image of the segmented leaf. The shape, texture, and color characteristics of the image are extracted.

Contours are used to calculate the area and perimeter of the leaf. The lines that run along the edges of objects that are all the same colour or intensity are known as contours. In an RGB image, the mean and standard deviation of each channel are calculated. Convert an image to HSV colour space and divide the total number of pixels in one channel by the number of pixels with hue (H) channel pixel intensities between 30 and 70 to find the amount of green colour. By subtracting the green colour part from 1, the non-green part of the image is calculated. The texture features were extracted from the grey level co-occurrence matrix after the colour features were extracted from the image (GLCM). The image's unique pixel relationship is represented by the GLCM. Extracting texture features from GCLM is a traditional method in computer vision. Contrast, Dissimilarity, and other characteristics were extracted from GCLM. Homogeneity, Energy, and Correlation are all words that come to mind when thinking about homogeneity.

The feature selection task is performed after all of the features from all of the images in the dataset have been extracted.

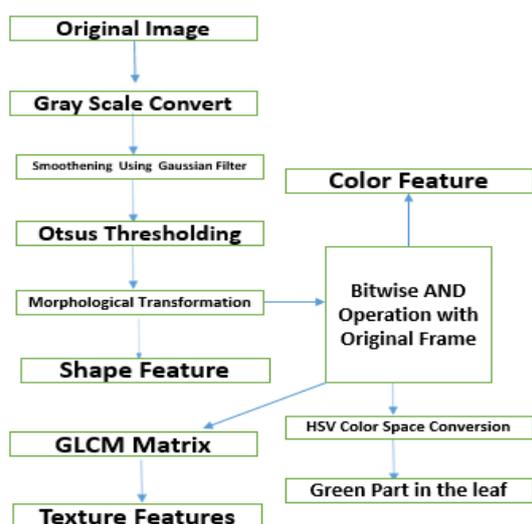


Fig. 2. Image Processing

3.2.2 Feature Extraction

The features characterize appropriate & refined qualities/evidence related to objects, distinguishing one

object from other objects. The features are supportive of identifying objects and labeling the category tag to an object. The feature abstraction stage is most substantial in making the classification/recognition model and searching for the extraction of relevant attributes characterizing each class. The features used in the presented system are explained as follows.

i) Grey Level Co-occurrence Matrix (GLCM):

It is used in a series of "second-order" texture calculations to measure the combinations of picture element brightness values (grey levels) that occur in a picture. The matrices are used to figure out how pixels are related spatially. The GLCM technique is extremely sensitive to any changes in the images, such as rotation, scale, and so on. By observing repeating patterns, spatial distribution, colour arrangement, and intensity, we can visually distinguish the texture of an image.

ii) Shape Features:

For the calculation of the shape feature, we have used the contour. Contours are described as a curve joining all the continuous points (along the boundary), having the same color or intensity. The contours are a useful tool for shape analysis and object detection and recognition.

Statistical and Color Feature:

The standard statistical feature such as mean, variance, and standard deviation are utilized here. The green feature is extracted as the amount of green color in the leaf. The RGB value of true Green color is (0,255,0). In the actual images there is always variation in the green color values due to various lighting conditions, low & even due to noise added by the camera while capturing & subsequently processing the image. To calculate the amount of green color in an image, convert the image to HSV color space and divide the total number of pixels in the green channel by the number of pixels with hue channel pixel intensity between 30 and 70. Subtracting the green color portion from One yields the non-green portion of the image.

3.2.3 Classification

The correlation matrix method is being used to select features. The correlation matrix can be used to predict how features are related to one another. In all machine learning problems, feature selection is critical. The features are selected based on the variables' correlation with the target variable. Figure 3 depicts the correlation of each variable with each other for the apple dataset. The F1 and F2 features have a high correlation (1), indicating that they are interdependent. As a result, one of them has been removed from the list (F2). Green channel mean, blue channel mean, green channel standard deviation, f4, f5, f6, f7, and f8 are less correlated features that will not contribute significantly to the model for apple disease prediction.

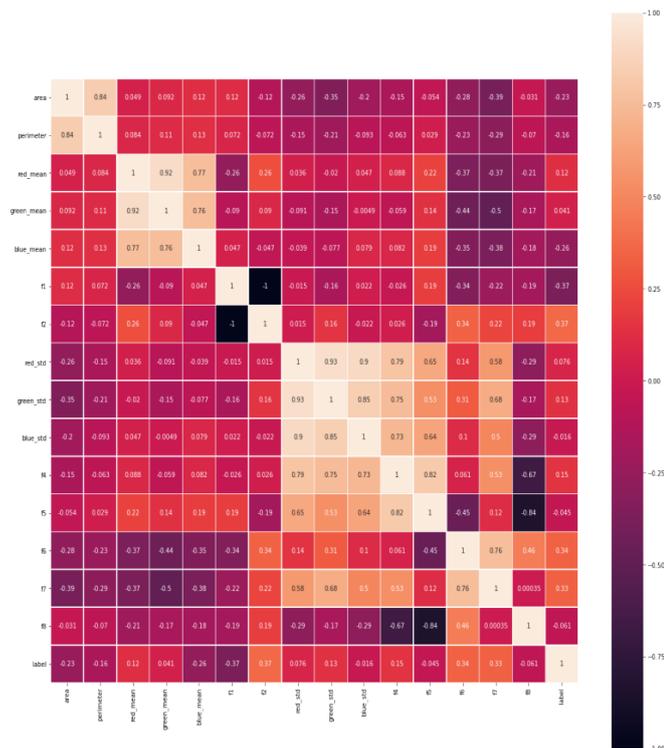


Fig. 3. Correlation map for apple dataset

3.2.4 Classical Model Development

For classification, the Random Forest algorithm is used. For regression, classification, and other tasks, a random decision forest is a mixed methods training technique. Several decision trees make up a random forest. In the case of a classification problem, Random Forest aggregates the results of all decision trees during training and outputs the class, while in the case of a regression model, it outputs the mean prediction. One of the main issues in decision tree algorithms is overfitting, which is overcome by using a random decision tree. Random forest is a classifier that uses many decision tree models to create an ensemble classifier.

4 RESULT & DISCUSSION

The proposed algorithm is tested on a database of 300 images of potato leaves taken from the publicly available dataset 'Plant Village,' which includes 100 healthy and 200 diseased leaves. The database was divided into two sets during the experiment: the training set, which contained 180 images (70 percent), and the testing set, which contained 120 images (30 percent). The accuracy, sensitivity, recall, F1-score, and precision of the presented system were calculated to evaluate its

performance. The classification testing accuracy is 95 percent when the train-test split is 60 percent -40 percent. Furthermore, 5-fold cross-validation was used to make the model more robust, and 93.7 percent accuracy was achieved. Table 2 illustrates the performance indicators.

Table 2. Performance Measure Classification

Plant dataset	Accuracy	F1 score	Precision
Tomato	87%	0.93	0.91
Potato	98%	0.99	0.99
Grapes	95%	0.99	0.99
Corn	94%	0.98	0.97
Apple	91%	0.91	0.92

The graphical representation of the confusion matrix containing Tomato test data for 0: healthy, 1: Bacterial spot, 2: Early blight, 3: Late blight, 4: Leaf Mold, 5: Septoria leaf spot, 6: Spotted spider mite, 7: Target Spot, 8: Yellow Leaf Curl Virus, 9: the mosaic virus is shown in Fig 4. It depicts the visual representation of the presented algorithm's performance on the testing dataset. As shown in Fig. 5, the area under the ROC curve for our classification is 87 percent, indicating that our classification is extremely accurate.

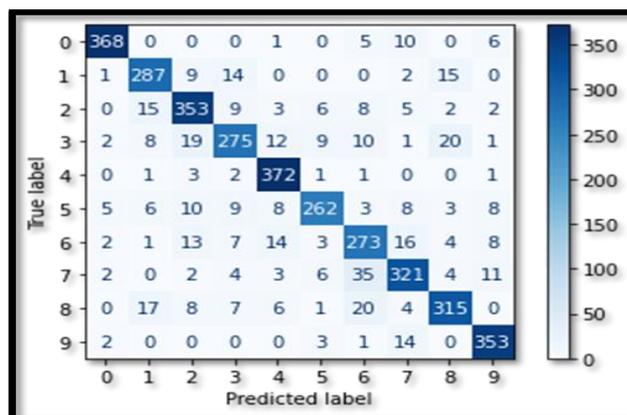


Fig. 4. Statistical data of confusion matrix on tomato dataset

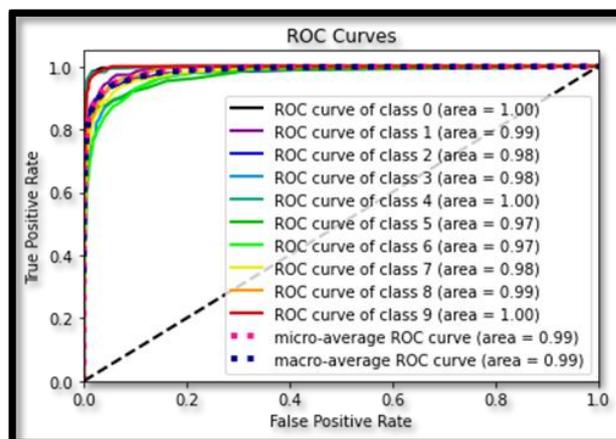


Fig. 5. ROC for tomato dataset

The graphical representation of the confusion matrix containing Potato test data for 0: healthy, 1: Early blight, 2: Late blight is shown in Fig.6. It represents a visual representation of our algorithm's performance on the testing dataset. As shown in Fig 7, the area under the ROC curve for our classification is 98 percent, indicating that our classification is extremely accurate.

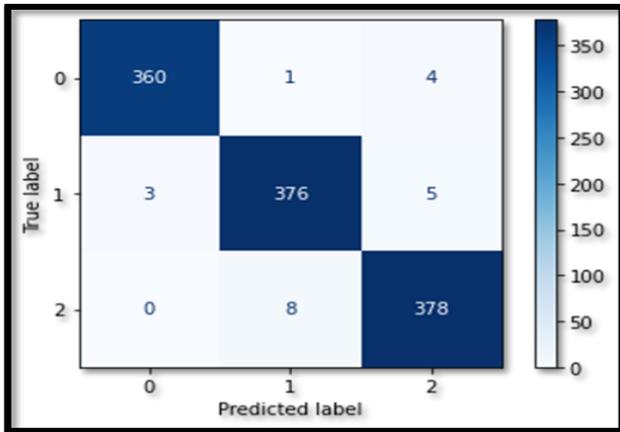


Fig. 6. Statistical data of confusion matrix on potato dataset

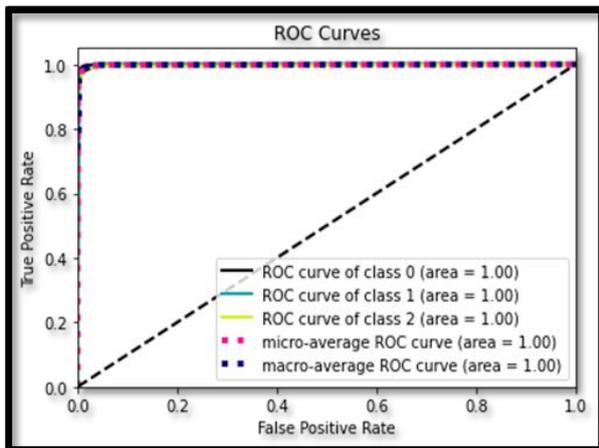


Fig. 7. ROC curve on potato dataset

The graphical representation of the Grape test data confusion matrix. for 0: healthy, 1: Black_rot, 2: Esca(Black Measles), 3: Leaf blight (Isariopsis Leaf Spot) as shown in Fig. 8. The ROC curve achieved 95% as indicated in Fig. 9.

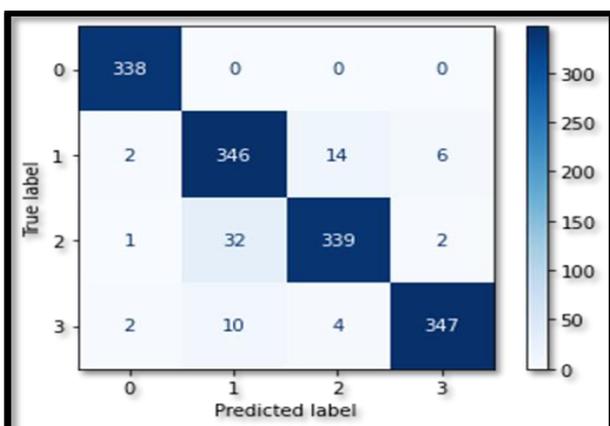


Fig. 8. Statistical data of confusion matrix on grapes dataset

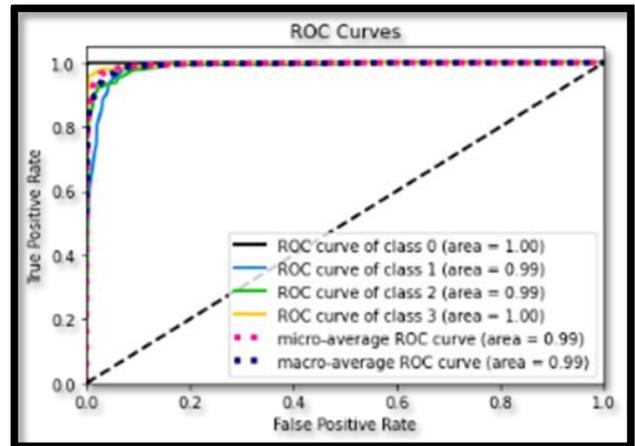


Fig. 9. ROC curve on the

The confusion matrix containing the Corn test data for which 0: healthy, 1: Cercospora leaf spot Gray leaf spot, 2: Common rust, 3: Northern Leaf Blight is presented in Fig. 10. Fig. 11 shows that the realm under the ROC curve for our classification is 94%.

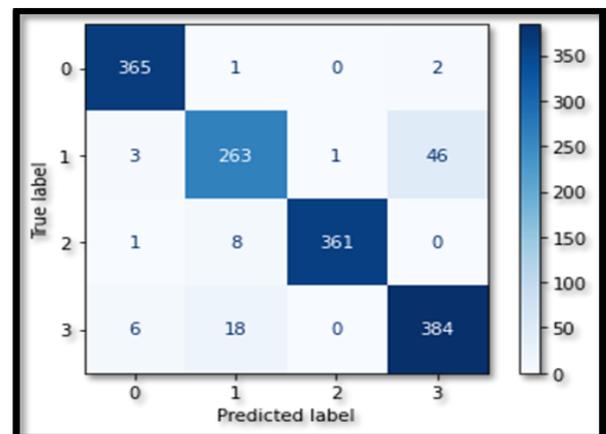


Fig. 10. Confusion matrix for Corn dataset

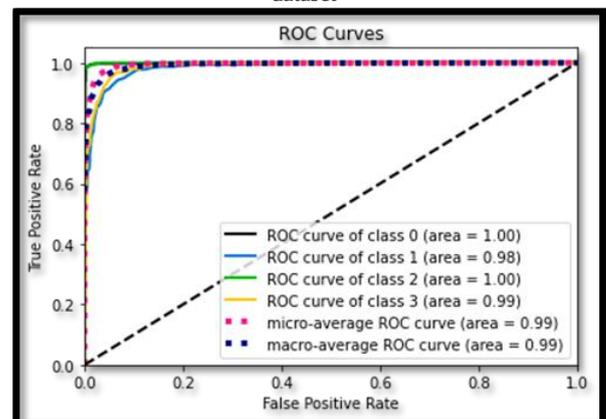


Fig. 11. ROC Curve for Corn dataset

The confusion matrix containing Apple test data for 0: healthy, 1: Apple scab, 2: Black rot, 3: Cedar apple rust is shown in Fig. 12. The realm under the ROC curve for our classification is 91% as seen from Fig. 13.

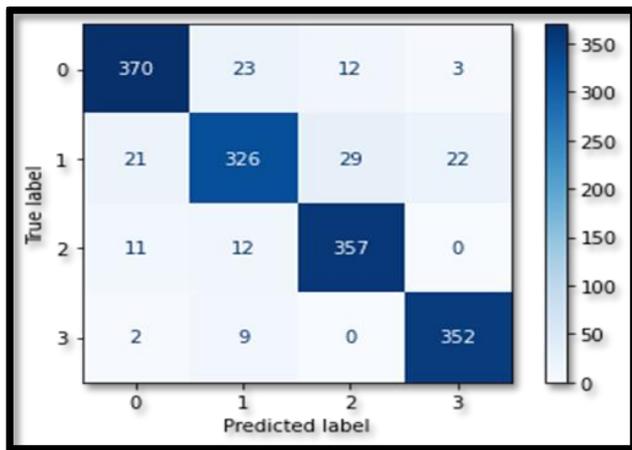


Fig. 12. Confusion matrix for Apple dataset

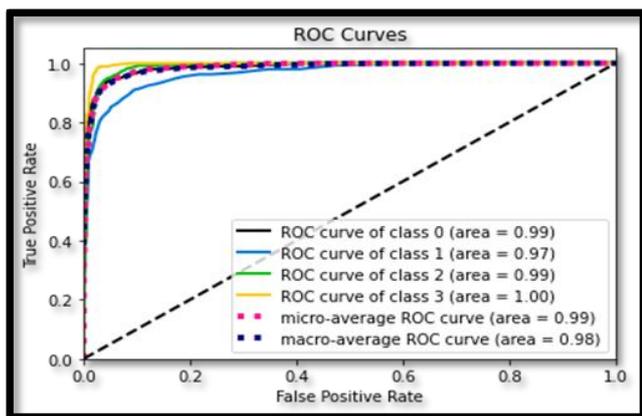


Fig. 13. ROC Curve for apple dataset

5 CONCLUSION

This paper describes a novel plant leaf disease classification technique that can be used for both automatic detection and classification of plant leaf diseases. As a result, related diseases for these plants were investigated. The best results were obtained with very little computational effort, demonstrating the efficacy of the proposed algorithm in the recognition and classification of leaf diseases. Another benefit of using this method is that plant diseases can be detected at an early or early stage. A 93 percent accuracy rate has been achieved. The research has real-world implications for crop disease classification, as it can be used to classify disease symptoms in fruits, vegetables, commercial crops, and other crops. The accuracy of the same can be further improved by implementing other classifiers.

REFERENCES

- [1] E. David et al. (2020).: Global Wheat Head Detection (GWHD) dataset: a large and diverse dataset of high resolution RGB labeled images to develop and benchmark wheat head detection methods. <https://arxiv.org/abs/2005.02162>.
- [2] SP . Mohanty, DP. Hughes and M. Salathe Using Deep Learning for Image –Based Plant Disease Detection .Front .Plant Sci. 7:1419 .DOI :10.3389/fpls.2016.01419(2016)
- [3] Kothawale, Shweta S., S. R. Barbade, and P. P. Mirajkar. "Grape leaf disease detection using SVM classifier." International Journal of Innovative Research in Computer and Communication Engineering 6.4 (2018).
- [4] S.Pavithra, A.Priyadharshini, V.Praveena and T.Monika Electronics and Communication Engineering, VSB Engineering College, Karur. "PADDY LEAF DISEASE DETECTION USING SVM CLASSIFIER." International Journal of Communication and Computer Technologies 3.1 (2015), 16-20.
- [5] Islam, Monzurul, et al. "Detection of potato diseases using image segmentation and multiclass support vector machine." 2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE). IEEE, (2017).
- [6] Ramesh, Shima, et al. "Plant disease detection using machine learning." 2018 International conference on design innovations for 3Cs compute communicate control (ICDI3C). IEEE, (2018).
- [7] Mohan, K. Jagan, M. Balasubramanian, and S. Palanivel. "Detection and recognition of diseases from paddy plant leaf images." International Journal of Computer Applications 144.12 (2016).
- [8] Orillo, J. William, et al. "Identification of diseases in rice plant (Oryza sativa) using backpropagation Artificial Neural Network." 2014 International Conference on Humanoid, Nanotechnology, Information Technology, Communication, and Control, Environment and Management (HNICEM). IEEE, (2014).
- [9] Kusumo, B. Suryo, et al. "Machine learning-based for automatic detection of corn-plant diseases using image processing." 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA). IEEE, (2018).
- [10] Sabrol, H., and S. Kumar. "Intensity-based feature extraction for tomato plant disease recognition by classification using the decision tree." International Journal of Computer Science and Information Security 14.9 (2016): 622.
- [11] SE. Raza, G. Princ, JP. Clarkson, NM. Rajpoot Automatic detection of diseased tomato plants using thermal and stereo visible light images. PLoS ONE 10(4):1–20 (2015)
- [12] MM. Hasan et al.: Detection and analysis of wheat spikes using Convolutional Neural Networks. Plant Methods. 14 DOI: 10.1186/s13007-018-0366-8. (2018).
- [13] Wu Wei et al.: Detection and enumeration of wheat grains based on a deep learning method under various scenarios and scales. Journal of Integrative Agriculture 19:1998–2008. DOI: 10.1016/s2095-3119(19)62803-0, ISSN 2095-3119(2020).
- [14] NGowda, S.N., Yuan, C (2019).: ColorNet: Investigating the Importance of Color Spaces for Image Classification. Computer Vision – ACCV 2018 Lecture Notes in Computer Science. 581– 596 (2019).