2020 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS 2020), 20 June 2020, Shah Alam, Malaysia.

Plant Leaf Diseases Identification using Convolutional Neural Network with Treatment Handling System

Koay K. Leong
Department of Electrical and Electronics Engineering
Faculty of Engineering and Technology
Tunku Abdul Rahman University College,
Jalan Genting Klang,
53300 Setapak, Kuala Lumpur
simon_kkl@hotmail.com

Lim L. Tze
Department of Electrical and Electronics Engineering
Faculty of Engineering and Technology
Tunku Abdul Rahman University College,
Jalan Genting Klang,
53300 Setapak, Kuala Lumpur
limlt@tarc.edu.my

***Abstract*— Agriculture is a very crucial industry to Malaysia where it would bring a huge impact on the country's wealth. All plant species, regardless of cultivated or wild, are prone to diseases and it is often inevitable. In the older days, identification of plant diseases was done by the experts of the field through observing the plants manually which can be rather tedious, time consuming and often inaccurate. With the advent of image processing technologies, automatic detection of plant disease can be done by capturing and processing the image of the plant leaves. In this paper, the K-means clustering and color thresholding are applied for image segmentation of the plant leaf. Both algorithms are compared in the study to evaluate which serves as a better segmentation algorithm on a target dataset to segment the region of interest (ROI). Moreover, two feature extraction approaches of ResNet-50 Convolutional Neural Network (CNN) and Gray-Level Co-Occurrence Matrix (GLCM) are also studied to evaluate their performance as a feature extractor. With applying the CNN to the support vector machine (SVM) classifier, it is investigated that an average classification accuracy of 96.63% can be achieved to perform the classification of the leaf diseases. A graphical interface for the system is also developed to provide the treatment handling method for the detected diseases.**

Keywords*— *Leaf diseases detection, image processing, color thresholding, K-means clustering, CNN, GLCM, SVM

I. Introduction

Agriculture plays a crucial role in the everyday life of human beings. Not only does it act as a food supply to most of the living creatures on earth, it also provide numerous raw materials such as latex, crude oil, cotton and log that has enabled humans to build everyday items. In Malaysia, although it is not as significant as compared to its golden age a few decades ago, the agriculture sector still contributes a huge amount in the country's wealth. In fact, 12% of the country's overall GDP are all contributed by the agriculture industry alone which includes palm oil, rubber, rice, cocoa [1].

Every farmer has one common fear and that is to have their crops infected by diseases. All plant species regardless of wild or cultivated are subjected to diseases and it is inevitable. Since plant diseases are usually unavoidable, disease detection plays a crucial role in the agriculture industry. When a plant is infected by a certain disease, it is required to be treated as soon as possible to prevent the degradation of crop quality and quantity which would lead to a loss in profit. However, as simple as it may sound, the detection of plant diseases is not an easy task to be carried out. To be able to identify these diseases, it requires special knowledge which normally the farmers doesn't incorporate. Therefore, the farmers would need to consult experts to assist in identifying the plant diseases which can be rather costly, tedious and the result are often inaccurate.

In recent years, image processing has been contributing a significant part in the agriculture industry for a variety of usages from fruit grading to weed detection [2]. Therefore, image processing is indeed not an unfamiliar term in the agriculture sector. Plant disease detection using image processing makes no exception as it has been gaining popularity among researchers in recent years and has seen as a potential replacement of the traditional method.

The detection and identification of plant leaf disease using image processing have been studied by researchers over the past years. As can be seen from these papers [3][4][5], the authors have suggested the use of K-means algorithm for the segmentation of the image. The algorithm was used to partition the dataset into k numbers of distinct non-overlapping subgroups (clusters) where each data point belongs to only one group so as to differentiate the diseased area on the leaf, the healthy part and the background. However, in [6], the author have suggested the use of color-based thresholding for segmentation. Images were converted to other color space like HSV, LAB and setting threshold at the individual channels to mask away unwanted region.

As for the extraction of features attributes, in this paper [7], the author have suggested the use of GLCM to extract the texture features of the image. The GLCM would determine the texture values by considering its neighbouring pixels of the region of interest. Specifically, this paper [8] uses SGDM for colour co-occurrence texture analysis to extract statistical texture features from the H image.

In these papers [3][5], the authors have proposed the use of support vector machine (SVM) as the algorithm to classify the data obtained from the features extracted. All the features that has been extracted from the previous stage to map into high dimensional to construct hyperplanes that separate the classes. In this paper [3], it suggests that using the Reproducing Kernel Hilbert Space (RKHS) can reduce the computational complexity where the author was able to obtain a classification accuracy of 90% from this while [5] were only able to obtain a classification accuracy of 88.89%.

According to this paper [4], the author has suggested to use the Random Forest classifier for the training and the classification of the mentioned diseases. As the name implies, the Random Forest classifier is like a forest with many individual decision trees which operate as an ensemble. From each of the trees, a class prediction will be produced and the one with the most votes are selected as the model's prediction. The system was able to predict up to four types of diseases.

The highlight of this paper [7] is that the author suggested the use of back-propagation and Particle Swarm Optimization (PSO) methods for training the neural network for the classification of plant leaf disease. The back-propagation was used for training the feed forward neural network, followed by optimizing it using particle swarm optimization. The backpropagation algorithm consists of two phases which is a feed-forward phase and a backward error correction phase. Overall the system was able to achieve an accuracy of 96.72%

In this paper, the plant leaf detection system which can process the leaf image and detect the infected area on the leaf for diagnosis of the disease using image processing is designed and developed. For implementation of the system in image segmentation and classification, several algorithms are studied and investigated to determine which algorithm is best suit to be applied to obtain better accuracy in detecting the disease on the plant leaf. Investigation is performed on using the K-means algorithm clustering and color thresholding for image segmentation. While, the GLCM and CNN (ResNet-50) are studied for leaf image classification using SVM (Support Vector Machine) classifier. The outline of the paper is as such the design methodology of the system is discussed in Section II, the result and discussion obtained from the studies will be demonstrated in Section III and lastly, the conclusion will be presented in Section IV with recommendation of future work.

II. Research Methodology

There are a few steps to apply in identifying the plant disease using image processing technology. As shown in Fig. 1, the steps include image acquisition, image pre-processing, segmentation, feature extraction and classification.

A. Image Acquisition and Preparation

In image acquisition, a tomato leaf image is acquired and used as the target image for further processing in the system. Three associated diseases are studied on tomato leaf image in this project such as Septoria leaf spot, Alternaria (Early Blight) and Bacteria disease. This would make up a total of four classes to be classified including 'healthy' tomato leaf as one of them. About 80% of the images of each class will be used for training the classifier while the remaining 20% is used for testing. Fig. 2 shows a sample of tomato leaf image from each class in the dataset.

B. Pre-processing

The next step after acquiring tomato leaf image is to pre-process the raw image to improve the quality and enhance the features of the image so that more information can be extracted from it. Since the dataset that is used in this project is very decent and contains very little noise, only a few pre-processing steps are done. There exist two image pre-processing steps which is resizing and contrast enhancement. All the images obtained from PlantVillage dataset are to be resized to 224 x 224 to ensure linearity in the datasets. After resizing it, the contrast of the image is then enhanced to make the objects in the images more distinguishable to be processed.

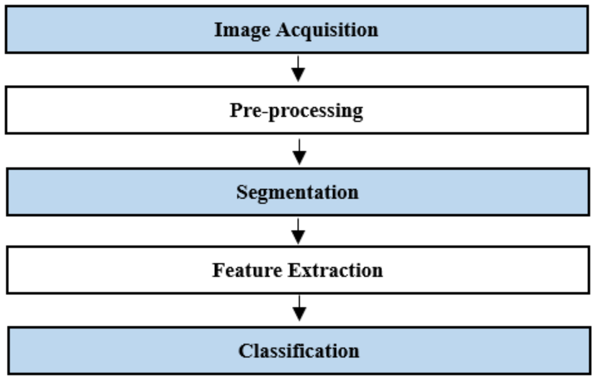

Fig. 1. Image processing phases.

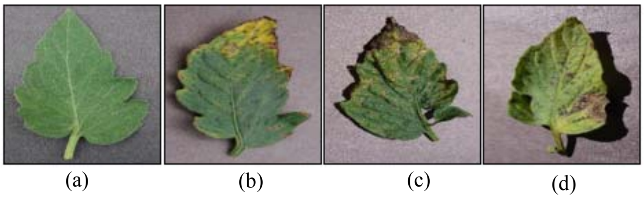

Fig. 2. Sample image from the dataset (a) Healthy, (b) Bacteria Disease, (c) Septoria Leaf Spot and (d) Alternaria (Early Blight).

C. Segmentation

The main goal of image segmentation is to obtain the region of interest (ROI) of the tomato leaf from image. In this case, the background of the dataset, which is in grey contains no information while the foreground is the plant leaf, which it would be analyzed by the system. Therefore, a mask must be created to mask out the background of the image which result the image to be left with only pixels containing the leaf with black background. After masking the background of the image, the healthy part of the leaf image will be further masked for calculating the diseased affected area in percentage. The affected area is calculated by the following formula:

$$Affected\ Area\ (\%) = \frac{Diseased\ area}{Total\ area\ of\ the\ leaf} \times 100\ \% \quad (1)$$

There are two segmentation approaches that will be used, namely, K-means clustering and color thresholding method. Each of these methods is investigated to evaluate which method is most suitable to be deployed with the given dataset.

K-means is an unsupervised machine learning algorithm that would make inferences from the dataset only based on its input vectors without referring to any labelled outcomes [9]. The 'K' in K-means clustering refers to the number of centroids to be used where a centroid refers to the imaginary or real location which would represent the center of the cluster. A cluster refers to a collection of data points aggregated together due to the certain similarities it has. Basically, what it does is to categorize pixels with similarities together. It takes advantage of the algorithm to separate the background, healthy and diseased area of the leaf into three different clusters. Apart from the RGB value, the spatial information and the texture information from a Gabor filter will be supplemented into the K-means algorithm as inputs.

Another segmentation method used to segment the image is color thresholding. Color thresholding is a method that is used to remove parts of the image that falls under a specific value of a color range. This method is suitable to be applied as the region that is required to be separated in the dataset happens to all fall under distinct colors. Initially, the RGB images are to be converted to HSV color space as it serves digital systems better for color perception. After converting the image to HSV color space, the hue (H) and Saturation (S) value of the image would be limited to a certain range value as both properties can be used for masking out dull color such as the image background in grey.

D. Feature Extraction

Once the image segmentation has been completed, the features of the segmented image would be extracted to train the classifier. Two different algorithms used for feature extraction are GLCM and CNN. The algorithm of GLCM is used for the feature extraction of the texture information from the image while feeding it into a classifier.

GLCM is defined over an image to be the distribution of co-occurring values at a given offset [10]. In other words, it represents the distance as well as the angular spatial relationship over an image sub-region of specific size. GLCM, as its name implies is created from a gray-scale image. Therefore, to perform GLCM, the image must be converted to gray scale in advance. In this study, eight features information that are associated with the texture of the image are extracted using the GLCM. The features are Mean, Variance, Correlation, Contrast, Dissimilarity, Homogeneity, Angular Second Moment and Entropy.

Another approach to perform the feature extraction is by using CNN. Training a CNN with large collection of diverse images is not an easy task. However, to make it easier, a pre-trained CNN can be used to leverage the power of CNN. It saves a huge amount of time and effort when a pre-trained CNN is used as feature extractor. Hence, the model ResNet-50 is used as a pre-trained CNN model to perform the feature extraction of plant leaf images.

E. Classification

Next, the features that are extracted from the previous stage would fed to the SVM classifier as inputs and it maps these features into high-dimensional feature spaces using its kernel function. SVM is a type of kernel-based supervised machine learning algorithm that would provide analysis of data for classification tasks [9]. In this work, multiclass SVM is used as the basic SVM which only works with 2 classes. The algorithm would attempt to find natural clustering of the labelled training data to groups and then map new unlabelled data into these newly formed groups.

III. Experimental Results and Discussion

Fig. 3 shows an enhanced contrast of the image in the pre-processing stage. It can be clearly observed that the colors look more vibrant and clearer as compared to the original image. Every element of the image looks very distinguishable and the texture is more emphasized. As can be seen from the plotted histogram, the contrast of enhanced image has a broader histogram which reflects a high contrast making it easier to be processed.

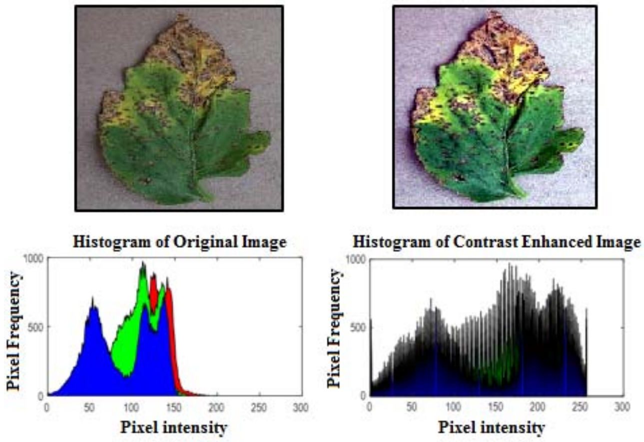

Fig. 3. Results of Contrast Enhancement.

To remove the background of the images in dataset, two methods of segmentation which are K-means clustering method and color thresholding are used respectively for evaluation of their performance. Fig. 4 shows the result of using K-means clustering method to segment the image into two clusters. The two clusters contain the image of the leaf and the background using k = 2. The first image in Fig. 4(a) shows the result of applying K-means algorithm with RGB as input to segment the image while the second image (Fig. 4(b)) is fed with additional spatial information and texture information from Gabor filter concatenated together. As can be compared from the results of K-means segmentation, the performance of image segmentation by using only the RGB information provided, is less encouraging. However, the results are slightly improved by using additional texture and spatial information as more features are fed into the K-means algorithm. Different number of clusters are experimented as well with k = 3, 4 and 5. Fig. 5 shows the result of using different number of clusters. As seen from the figure, with the increase of number of clusters, the result got worsen as it starts to blends part of the leaf with the background instead of differentiating the details of the leaf.

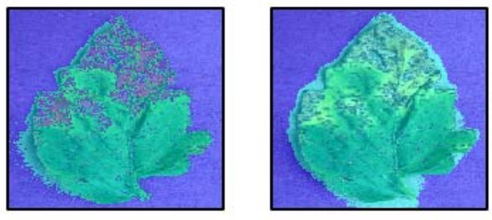

Fig. 4. Segmentation on leaf image using (a) K-means clustering and (b) K-means with textual and spatial information when k = 2.

Fig. 5. Segmentation result using k-mean with different number of clusters (a) k = 3, (b) k = 4 and (c) k = 5.

Fig. 6 shows the image segmentation using color thresholding. As depicted in Fig. 6(a), the background pixels' values are eliminated with remaining the foreground pixels of the leaf. However, it is observed that certain pixels in the region of interest has masked away (as indicated in black spots within the leaf area in Fig. 6(b)) due to certain area on the leaf has a similar hue (H) and saturation (S) content with the background. To reduce the black spot within the leaf area, an improved binary mask is formed (Fig. 6(c)) by having the unwanted black spots within the leaf removed and the outer part of the leaf shape is connected.

Hence, it can be compared and deduced that the leaf images being segmented using color thresholding method are more precise than using K-means based on the results compared in Fig. 7. Using K-means has limitation as some part of the background are not completely removed from the image due to the shadow at the edges. Thus, the color thresholding method is chosen to implement in the system. The segmentation process using color thresholding is repeated in masking the healthy part of the leaf in order to acquire the infected area with diseases from the image for calculation of affected area on leaf.

Next, the feature extraction using GLCM and CNN approaches are investigated and implemented to feed into the SVM classifier. Table I and Table II shows the tabulated confusion matrix result that are produced by SVM classifier for the respective approaches of GLCM and CNN. 200 images from the testing dataset are used for each class to carry out the testing. The diagonal elements in the confusion matrices represent the number of images that are correctly matched for the predicted label and true label, while the off-diagonal elements are those that are mislabelled by the classifier. The higher the diagonal values of the confusion matrix, the better as it indicates more correct predictions.

As can be observed from the result on both the confusion matrices, an inference can be made that by using the two different proposed methods to perform feature extraction, both produce a fairly good outcome on a SVM classifier. The confusion matrices are computed to determine to overall accuracy for each system and it is tabulated in Table III for the four classes of the leaf condition for healthy, bacterial, early blight and Septoria. Both of the feature extraction approaches are able to achieve high accuracy of more than 80% in overall. However, it can be seen that CNN has outperformed the GLCM in feature extraction as it attained an accuracy of 96.63% in average. This is due to the reason that the pre-trained CNN model which is the ResNet-50 is being trained with over a million of images to construct better representation of the image while the GLCM only extracted eight texture features of each image where the number of representation information are low as compared to CNN. As a result from this investigation, the ResNet-50 CNN model is determined to apply as the feature extractor to feed into the SVM classifier in the system, with color thresholding for image segmentation of the leaf.

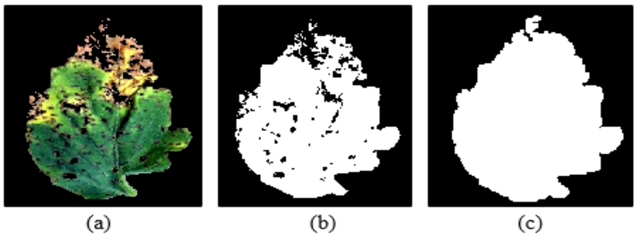

Fig. 6. Segmentation using color thresholding with (a) initial result, (b) binary mask from the initial result and (c) improved binary mask.

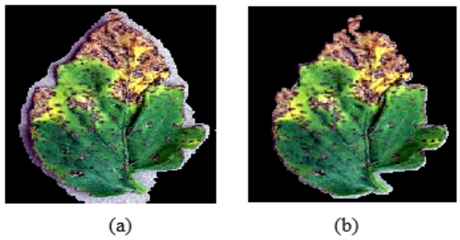

Fig. 7. Comparison between (a) K-means and (b) color thresholding on segmentation of leaf image.

TABLE I. CONFUSION MATRIX OBTAINED USING GLCM

		TRUE CLASS			
PREDICTED CLASS		**Healthy**	**Bacterial**	**Early Blight**	**Septoria**
	Healthy	168	8	16	8
	Bacterial	4	170	15	11
	Early Blight	20	16	159	5
	Septoria	3	18	15	164

TABLE II. CONFUSION MATRIX OBTAINED USING RESNET-50

		TRUE CLASS			
PREDICTED CLASS		**Healthy**	**Bacterial**	**Early Blight**	**Septoria**
	Healthy	179	13	0	8
	Bacterial	2	194	1	3
	Early Blight	0	0	200	0
	Septoria	0	0	0	200

TABLE III. ACCURACY COMPARISON USING GLCM AND CNN

	GLCM, Accuracy (%)	**CNN, Accuracy (%)**
Healthy	**84.00**	**89.50**
Bacterial	**85.00**	**97.00**
Early Blight	**79.50**	**100.00**
Septoria	**82.00**	**100.00**
	TOTAL: 82.63	**TOTAL: 96.63**

For user to apply the detection system, a Graphical User Interface (GUI) with the ability to display the disease information and the possible treatment to prevent the diagnosed disease on the infected plant leaf is developed. As shown in Fig. 8, the user is allowed to interactively load the leaf image into the system to identify and diagnose whether the target image is healthy or being infected. Throughout the process, the system would prompt user to specify the path to the appropriate dataset used for training through the click of 'TRAIN' button. Once the training has been completed, the user could load the desired leaf image from any source into the system for processing. The subsequent processing step of the image segmentation would be performed on the leaf image through the click of 'CLASSIFY' button after the leaf image is loaded. The processed image will then be used to extract its features and map into the trained SVM classifier for classification. The processed image of the leaf condition after image segmentation and classification are displayed in the four columns reserved in the interface. As a result, the condition of leaf whether it is healthy or being infected would be decided from the classification. The information of the diagnosed diseases from the plant leaf will be displayed through the designated column of classification result with the percentage of affected area shown if the leaf is infected with disease. The treatment procedure for an infected leaf will be recommended based on the type of disease detected from the system to the user in the description box as illustrated from Fig. 8 to Fig. 11 for the four classes of leaf images tested.

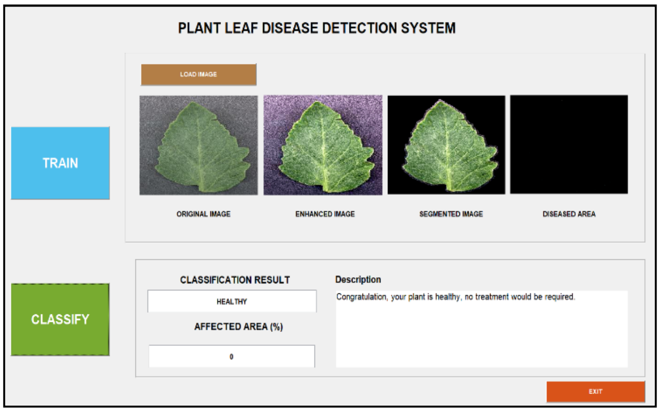

Fig. 8 Detection of healthy leaf image.

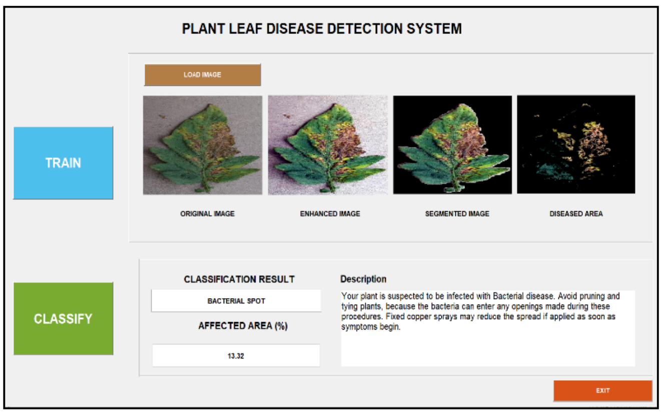

Fig. 9. Detection of leaf image with bacterial disease.

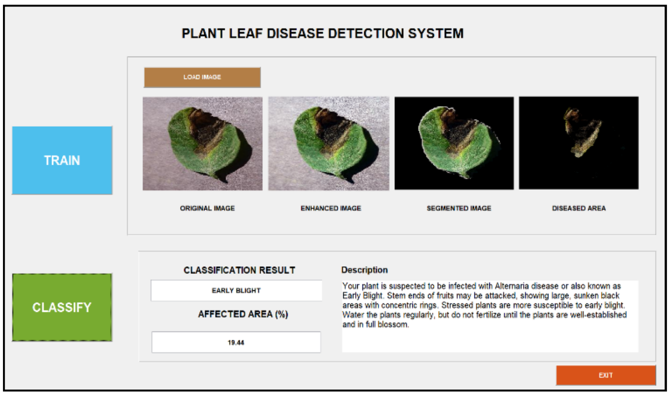

Fig. 10. Detection of leaf image with Early Blight.

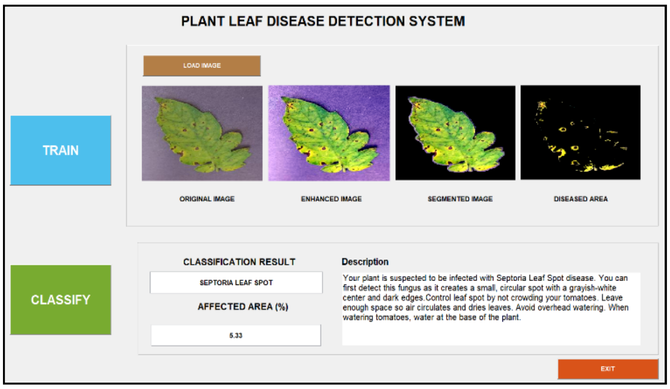

Fig. 11. Detection of leaf image with Septoria Leaf Spot.

IV. Conclusion and Future Work

In conclusion, the proposed system using color thresholding for image segmentation and ResNet-50 CNN model for feature extraction using MATLAB R2019a has implemented successfully. The accuracy of detecting the healthy, bacterial, early blight and Septoria of leaf image reach 89.5%, 97%, 100% and 100% respectively using the developed system. Moreover, the proposed system has the flexibility to work with most leaf diseases, depending on the dataset that it is being trained with. The proposed system design with the user interface in this project is able to perform comparatively well in detecting and classifying plant leaf disease and with further fine tuning, it has the potential to be integrated into the agricultural system. More works can be done in future to make the system work more effectively. For example, a more advanced image segmentation method like Deep CNN can be researched to obtain a more precise segmentation of the image and enable the system to work with dataset with a much more complex background.

References

[1] "Malaysia Agriculture, Information about Agriculture in Malaysia." [Online]. Available: https://www.nationsencyclopedia.com/economies/Asia-and-the-Pacific/Malaysia-AGRICULTURE.html. [Accessed: 13-Jan-2020].

[2] A. Vibhute and S. K. Bodhe, "Applications of Image Processing in Agriculture: A Survey," *Int. J. Comput. Appl.*, 2012.

[3] R. Meena Prakash, G. P. Saraswathy, G. Ramalakshmi, K. H. Mangaleswari, and T. Kaviya, "Detection of leaf diseases and classification using digital image processing," *Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017*,

vol. 2018-Janua, pp. 1–4, 2018.

[4] A. Devaraj, K. Rathan, S. Jaahnavi, and K. Indira, "Identification of plant disease using image processing technique," *Proc. 2019 IEEE Int. Conf. Commun. Signal Process. ICCSP 2019*, pp. 749–753, 2019.

[5] P. B. Padol and A. A. Yadav, "SVM classifier based grape leaf disease detection," *Conf. Adv. Signal Process. CASP 2016*, pp. 175–179, 2016.

[6] C. G. Dhaware and K. H. Wanjale, "A modern approach for plant leaf disease classification which depends on leaf image processing," *2017 Int. Conf. Comput. Commun. Informatics, ICCCI 2017*, pp. 5–8, 2017.

[7] M. Chanda and M. Biswas, "Plant disease identification and classification using Back-Propagation Neural Network with Particle Swarm Optimization," *2019 3rd Int. Conf. Trends Electron. Informatics*, no. Icoei, pp. 1029–1036, 2019.

[8] M. S. Arya, K. Anjali, and D. Unni, "Detection of unhealthy plant leaves using image processing and genetic algorithm with Arduino," *EPSCICON 2018 - 4th Int. Conf. Power, Signals, Control Comput.*, pp. 1–5, 2018.

[9] "Understanding K-means Clustering in Machine Learning." [Online]. Available: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. [Accessed: 27-Jan-2020].

[10] J. Hayes, "Image Classification Image Classification-Gray Level Co Gray Level Co-Occurrence Matrix Occurrence Matrix (GLCM) (GLCM)."