TEXTURE FEATURES EXTRACTION TECHNOLOGY USING GREY LEVEL CO-OCCURRENCE MATRIX FOR THE KNN CLASSIFICATION OF CITRUS DISEASE

Wilis Kaswidjanti[1], Hidayatulah Himawan[2] and Galih Wangi Putri[3]
[1]Department of Informatic, UPN Veteran Yogyakarta, Indonesia
[2]Department of Informatic, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia
[3]Department of Informatic, UPN Veteran Yogyakarta, Indonesia
E-Mail: if.iwan@upnyk.ac.id

ABSTRACT

The citrus disease is a problem affecting the decrease of agricultural commodity yields. One way to determine disease in citrus is through the leaves. Leaves, as a place for photosynthesis, with disease will cause stunted plant growth. Therefore, the fruit can experience a quality decrease. This study aims to classify citrus diseases based on leaf images by applying extraction technology of GLCM (Gray Level Co-occurrence Matrix) using KNN (K-Nearest Neighbor). Citrus disease classification has four main stages, namely preprocessing, segmentation, feature extraction, and classification. The preprocessing stage converts the LAB color space. Segmentation stage uses Otsu Thresholding. Texture features extraction uses GLCM. Classification uses KNN. KNN classification uses several distances, namely Chi-Square, City Block (Manhattan), Correlation, Cosine, Euclidean, and Hassanat. Comparisons are made based on the normalization of the dataset and the KNN distance used. The dataset without normalization gets the best results with Hassanat distance KNN (k = 29) with an accuracy of 91.86% and the dataset with normalization gets the best results at Euclidean distance (k = 7) with an accuracy of 98.84%. This research was expected to find out the accuracy of the method mentioned above in the classification of citrus diseases.

Keywords: gray level Co-occurrence matrix, K-nearest neighbour, classification, extraction technology.

Manuscript Received 13 October 2022; Revised 23 May 2023; Published 15 June 2023

1. INTRODUCTION

Recently, computer vision has been studied from many angles. This extends combining digital image processing, pattern recognition, machine learning and computer graphics, from raw data recording into techniques and ideas. computer vision technology is used in various aspects of life [1]. One field that applies it is agriculture. In agriculture, computer vision technology is used for various things such as the classification of plant types, sorting fruit, classification of plant diseases or pests, and determining the yield quality of these plants [2].

Plant disease is one of the factors that can be found in the field. The disease attacks several parts of the plant such as fruit, leaves and stems. Leaves as a place for photosynthesis will experience stunted growth if they are attacked by disease. As a result, the fruit can experience a decrease in quality [3]. The leaf section contains the most accurate information regarding plant identification [4]. Early detection of the types of diseases that attack plants can be used to find appropriate treatments. In general, the classification of plant diseases has four main phases, namely preprocessing, segmentation, feature extraction, and classification [5].

At the preprocessing stage, there are several processes that can be carried out aimed at improving image quality. Some of them include image enhancement, color space conversion, image filtering, image resizing, and image rescaling. Gavhale *et al*. used SF-CES for image enhancement, color space conversion (YCbCr, LAB), and grayscale [6]. Ali *et al*. used the equalization histogram for image enhancement and conversion of the RGB color space to LAB [7]. Sharif *et al*. used a Top-hat filter and a gaussian function for image enhancement [8]. Sunny & Gandhi used the resizing method, Gaussian, Median, Linear, Low Pass, High Pass to remove noise, and CLAHE for (image enhancement) [9].

The segmentation stage is the process of separating the image of the disease lesion area from healthy leaves and other backgrounds. K-Means Clustering was used in the following study to separate the areas of healthy and diseased leaves into clusters. The clusters represent each disease, so that one leaf can be identified as having more than one disease [3] [5] [6] [9] [10]. Arivazhagan et al. used green pixel masking and threshold [11]. Ratnasari *et al.* used Otsu thresholding based on the a* and b* components of the LAB color space [12]. The use of Otsu Thresholding is also used in research [13]. Rakib Hassan *et al.* used RGB-HSV Thresholding based on the RGB color space [14].

In the feature extraction stage, the features can be in the form of color, texture, and shape. Each object has unique characteristics that help in the process of object recognition for classification [5]. The GLCM for texture feature was used in the following studies [6] [8] [9] [10]. The LBP (Local Binary Paterrn) method for texture feature is also used by [5] [7]. RGB histogram and HSV histogram is for color feature [5] [7]. Extraction of color features using RGB, HSV, HIS, LAB, LUV was used in the following studies [5][8]. The Color Moments method based on the LAB color space which includes calculating the mean, standard deviation, skewness, and kurtosis was used in the following study [12].

The classification stage used the extracted feature value. The SVM (Support Vector Machine) method is a method that is widely used based on reference journals because it is easily recognized and gives correct results related to other classification approaches [5] [6] [7] [9] [10] [11]. M-SVM (Multiclass Support Vector Machine) is an SVM method that can classify more than two classes, whereas SVM can only classify two classes. The KNN method was used in the following research [3]. KNN is a method that is easy to implement [5].

This study performs citrus disease classification based on leaf images by implementing the GLCM method, Color Moments, Min-max Scaling normalization, and KNN with ChiSquare, City Block (Manhattan), Correlation, Cosine, Euclidean, and Hassanat distances. In the preprocessing, the image was converted from RGB to LAB color space. LAB has the most complete color space because it can describe all colors like a human perspective [15]. In addition, the LAB color space produces good segmentation on Otsu Thresholding with various background conditions [13]. Furthermore, Otsu Thresholding segmentation is carried out to separate the image affected by the disease from the healthy leaves. The Otsu Thresholding method works well for separating the foreground and background [5]. The GLCM method has been used in several studies for the extraction of texture features. Color Moments are used for the extraction of color features based on LAB images like the research conducted by Ratnasari *et al.* [12]. The classification used KNN method. This method is a classification method that is easy to implement [5]. In this study, the classification with KNN compares the distances of Chi-Square, City Block (Manhattan), Correlation, Cosine, Euclidean, and Hassanat. Chi-Square distance is the best accuracy distance in the study of [16] and Hassanat is the best accuracy distance in the study of [17]. Min-max scaling normalization method is used on the dataset. The normalization method is used to equalize the value scale on the feature value so that no feature value is the most prominent [18]. It is also used to know the effect of normalization on the dataset on the level of classification accuracy.

2. METHOD

This study performs a classification of citrus diseases based on leaf images by implementing the GLCM (Gray Level Co-Occurrence Matrix), Color Moments, Min-max Scaling normalization, and KNN (K-Nearest Neighbor) with Chi-Square, City Block (Manhattan) distance. Correlation, Cosine, Euclidean, and Hassanat. Sources of data sets for lime leaf disease images were taken from the digipathos repository [19], Mendeley data [20], and Citrus Disease [21]. In preprocessing, the images are converted from RGB to LAB color space. LAB has the most complete color space, because it can describe all colors like a human perspective [15]. In addition, the LAB color space produces good segmentation on Otsu Thresholding with various background conditions [13]. Furthermore, Otsu Thresholding segmentation is carried out to separate the image affected by the disease from the healthy leaves. The Otsu Thresholding method works well for separating the foreground and background [5]. The GLCM method has been used in several studies for the extraction of texture features. Color Moments are used for the extraction of color features based on LAB images such as research conducted by Ratnasari *et al.* [12]. For the classification method using the KNN, this method is a classification method that is easy to implement [5]. In this study, the classification with KNN will compare the distances of Chi-Square, City Block (Manhattan), Correlation, Cosine, Euclidean, and Hassanat. Chi-Square distance is the best accuracy distance in the study of [16] and Hassanat is the best accuracy distance in the study of [17]. Min-max scaling normalization method is used on the dataset. The normalization method is used to equalize the value scale on the feature value so that no feature value is the most prominent [18].

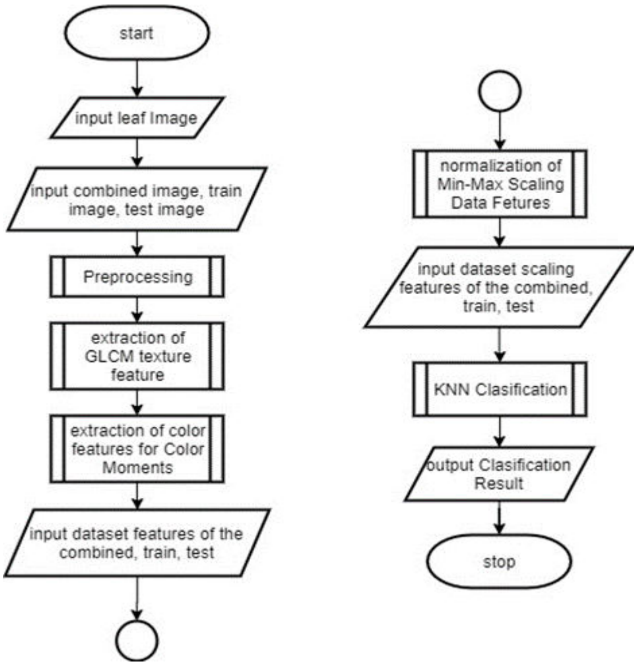
start
input leaf Image
input combined image, train image, test image
Preprocessing
extraction of GLCM texture feature
extraction of color features for Color Moments
input dataset features of the combined, train, test
normalization of Min-Max Scaling Data Fetures
input dataset scaling features of the combined, train, test
KNN Clasification
output Clasification Result
stop

Figure-1. The main process flowchart.

2.2 Preprocessing

The data set used in this study is Mendeley's data by Rauf *et al.* [20]. From this data set, three leaf image conditions were taken, namely canker, greening, and healthy. The images have RGB or BGR color space format with the same row (M) and column (N) size (M = N). This study uses dimensions (256 x 256).

Table-1. Category of citrus leaf condition and number of images.

Category	Number of Images
canker	163
greening	204
healthy	58

a) Color Space Conversion

a. BGR image to LAB image conversion

The following is the calculation for the conversion from the RGB color model to the LAB color space. CIE L*a*b* is also referred to as CIELAB color model as the second uniform color space of the CIE XYZ space with a white reference point. CIE L*a*b* is the most complete color space, because it can depict all colors like a human perspective [15]. The following is the process of converting from RGB to CIE XYZ, then from CIE XYZ to CIELAB [22]:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0{,}4125 & 0{,}3576 & 0{,}1804 \\ 0{,}2127 & 0{,}7152 & 0{,}0722 \\ 0{,}0193 & 0{,}1192 & 0{,}9502 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

$$L = \begin{cases} 116\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - 16 & jika, \quad \frac{Y}{Y_n} > 0.008856 \\ 903.3 & lainnya \end{cases} \tag{2}$$

$$a = 500\left[\frac{X^{\frac{1}{3}}}{X_n} - \frac{Y^{\frac{1}{3}}}{Y_n}\right] \tag{3}$$

$$b = 200\left[\frac{Y^{\frac{1}{3}}}{Y_n} - \frac{Z^{\frac{1}{3}}}{Z_n}\right] \tag{4}$$

b. Channel separation (split) in the LAB images into 'L', 'A', 'B' channels.

Each channel in the LAB images is separated into channel L, channel A, and channel B. Channel A is used for the segmentation process of Otsu Thresholding.

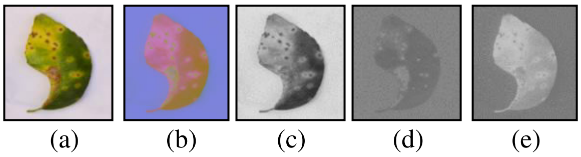

Figure-2. (a) BGR leaf image, (b) LAB leaf image, (c) 'L' channel, (d) 'A' channel, (e) 'B' channel.

b) Segmentation

Otsu thresholding, one of the important steps in preprocessing, is used to separate objects from the background [23]. Otsu Thresholding is done on channel 'A' from LAB. The results of the Otsu Thresholding process are used in the masking process. An image of a BGR lesion is produced and converted to grayscale.

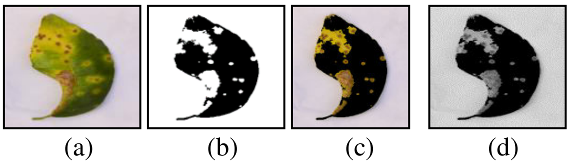

Figure-3. (a) BGR leaf image, (b) The result of Otsu thresholding image, (c) Image masking result of BGR image and Otsu thresholding image, (d) grayscale leaf lesion image.

2.2 Extraction of Texture and Color Features

GLCM method is one of the methods used in extraction technology of texture features, which estimates the image properties related to second-order statistics. GLCM calculates the level of gray that occurs in partial relationship with certain other features [24]. Texture features are calculated spatially based on spatial relationships or gray patterns between one image and another. The gray pattern in GLCM has a dependency matrix between one another. The neighbor relationship of this matrix uses a degree of distribution which has four degrees, namely 0°, 45°, 90°, and 135°. In this study, 0° degree is used. The GLCM calculation process uses a grayscale image. GLCM features consist of ASM (Angular Second Moments), Contrast, IDM (Inverse Different Moment) or Homogeneity, Correlation, and Energy [25].

Color Moments is a representation of the color feature of an image. The representation of color moments is displayed in the form of a color distribution [26]. The color moment feature consists of the mean, standard deviation, skewness, and kurtosis [27]. Calculating color moments in the LAB image will get 12 features. Those are three mean features (L, A, B mean channels), three features of standard deviation, three features of skewness, and three features of kurtosis. M, N are dimensions of image height and width is the pixel value in row i of column j of each channel.

2.3 Classification

K-Nearest Neighbor is a method that classifies new objects based on the training data [28]. K-Nearest Neighbor is an instance-based method. The proximity between new objects and old objects is calculated. The new objects are classified based on the majority vote of the number of k of neighbors. The value of k taken is usually a small number of k [29]. The basic steps of the KNN classification consist of two phases, namely the training phase and the classification phase [17]. There are several distances in the different KNN categories. Here are some of them [17]: Minkowski Distance (Manhattan distance, Euclidean distance), Inner product distance family, Cosine distance, L2 distance measure family, Squared Chi-Squared distance, Other measure, Correlation distance, and Hassanat.

The process of training and testing requires training data and test data. The training data and the test data are (80%) and (20%) of the total image respectively.

3. RESULT AND DISCUSSIONS

Accuracy comparisons are made based on the KNN distance and the application of the normalization method to the data set. Based on the KNN distance, there are six distances being compared namely, Chi-square, Correlation, Cosine, Euclidean, Hassanat, and Manhattan. Based on the application of the normalization method, there are two data set namely, (1) data set without normalization and (2) data set with Min-max Scaling normalization.

Figure-4 and Table-2 show the results of the classification using the data set without the normalization process. From the table, it can be concluded that the highest accuracy obtained is 91.86047 by the Hassanat distance (k = 29).

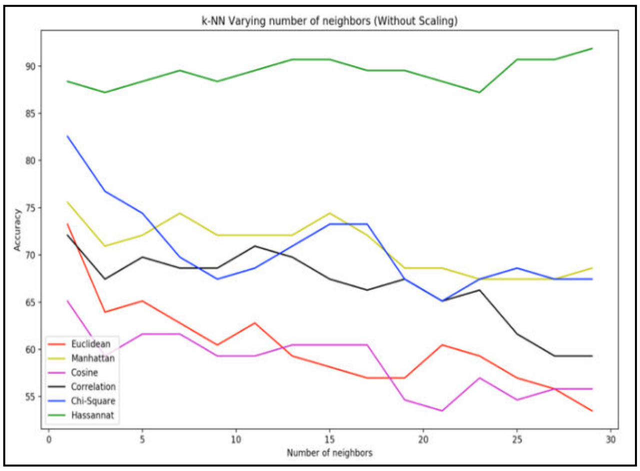

Figure-4. Diagram of accuracy without normalization.

Table-2. Comparison of the accuracy of each distance without normalization.

k	Distance					
	Euclidean	**Manhattan**	**Cosine**	**Correlation**	**Chi-square**	**Hassanat**
1	**73,25581**	**75,5814**	**65,11628**	**72,09302**	**82,55814**	88,37209
3	63,95349	70,93023	59,30233	67,44186	76,74419	87,2093
5	65,11628	72,09302	61,62791	69,76744	74,4186	88,37209
7	62,7907	74,4186	61,62791	68,60465	69,76744	89,53488
9	60,46512	72,09302	59,30233	68,60465	67,44186	88,37209
11	62,7907	72,09302	59,30233	70,93023	68,60465	89,53488
13	59,30233	72,09302	60,46512	69,76744	70,93023	90,69767
15	58,13953	74,4186	60,46512	67,44186	73,25581	90,69767
17	56,97674	72,09302	60,46512	66,27907	73,25581	89,53488
19	56,97674	68,60465	54,65116	67,44186	67,44186	89,53488
21	60,46512	68,60465	53,48837	65,11628	65,11628	88,37209
23	59,30233	67,44186	56,97674	66,27907	67,44186	87,2093
25	56,97674	67,44186	54,65116	61,62791	68,60465	90,69767
27	55,81395	67,44186	55,81395	59,30233	67,44186	90,69767
29	53,48837	68,60465	55,81395	59,30233	67,44186	91,86047

While the classification using data set with Min-Max Scaling normalization gets the highest accuracy of 98.83721 by the Euclidean distance (k = 7). From these two comparisons it can be concluded that the highest accuracy of 98.83721 is resulted from the Euclidean distance with a value of k = 7 using the Min-max Scaling normalized data set. The use of the Min-Max normalization method has an effect in determining the accuracy value. For the comparison of the KNN distance used, it can be seen that the accuracy of the Euclidean distance is better than the Chi-Square and Hassanat distances which were the best distances in previous studies. The research results obtained are also influenced by the quality of the data set, such as the number of data sets, the uniformity of the number of data set distribution in each class (label).

This study conducted a test using confusion matrix. Confusion matrix is a technique used to determine the performance of a classification algorithm. The confusion matrix is a matrix of the number of true and false numbers of the predicted results. N is the number of classes (labels). The number of cells in the matrix is N x N. The x-axis is the predictive label and the y-axis is the actual label.

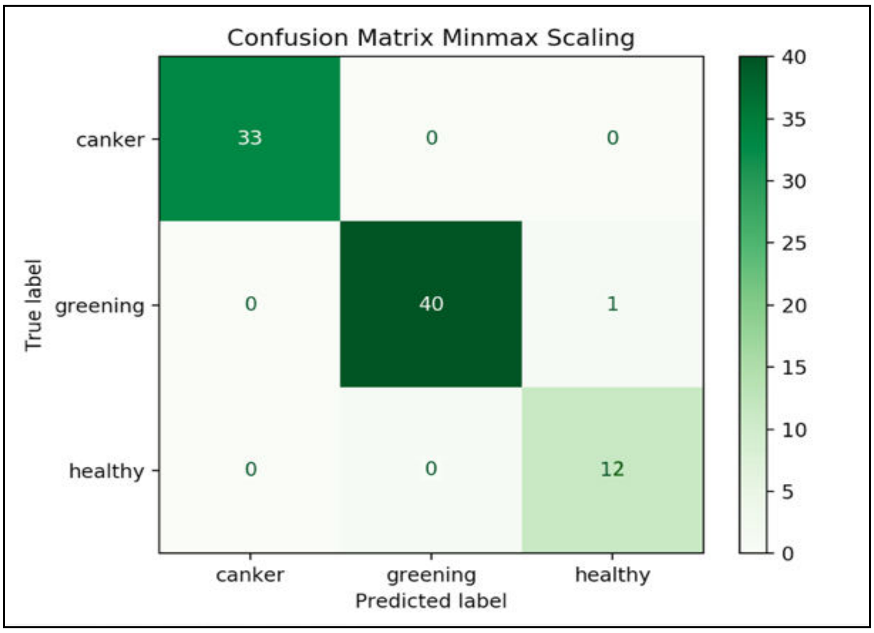

Figure-5. Min-max scaling confusion matrix of Euclidean distance (k=7).

Calculations using the Confusion Matrix resulted in 85 correct detected images and 1 incorrect detected image. The results of this calculation are used to measure the performance of the model using classification metrics. The measurements taken are precision, recall, specificity, and F1-score.

Recall, also called as sensitivity, is part of the test data from a positive event that is predicted correctly. Precision is the piece of test data from a positive event that is predicted to be actually positive. Specificity, also known as TNR (True Negative Rate), is a measure of the proportion of correctly identified negative actual conditions. F1-Score is the mean harmonic of recall and precision. The higher the value, the better the model a classification.

Table-3. Measurement of classification performance based on type of disease.

Class	Precision	Recall	Specificity	F1-Score	Class
Canker	1.00	1.00	1.00	1.00	Canker
greening	1.00	0.975	1.00	0.987	greening
healthy	0.923	1.00	0.986	0.96	healthy

4. CONCLUTIONS

Based on the results of the implementation and testing of the proposed method, several conclusions can be drawn: 1. The Otsu thresholding segmentation method can work well using the LAB color space on citrus leaves which is a dicot type plant by producing a fairly good feature extraction. Feature extraction as a dataset is used for the classification process. Therefore, the preprocessing stage which functions to improve image quality will affect the accuracy results; 2. The normalization method affects the level of accuracy. The dataset before normalization has an accuracy of 91.86% (Hasssanat distance, k = 29). After the normalization, the accuracy is 98.83% (Euclidean distance, k = 7); 3. In the dataset without normalization, the accuracy obtained is 73.26% from Euclidean distance (k = 1), 75.58% from Manhattan (k = 1), 65.11% from Cosine (k = 1), 72.09% from Correlation (k = 1), 82.56% from Chi-Square (k = 1), and 91.86% from Hassanat (k = 29); 4. In the dataset with normalization, the accuracy of the Euclidean distance (k = 7) is 98.84%, Manhattan (k = 21) is 96.51%, Cosine (k = 7) is 96.51%, Correlation (k = 7) is 97.67%, Chi-Square (k = 7) is 97.67%, Hassanat (k = 19) is 96.51%

REFERENCES

[1] V. Wiley and T. Lucas. 2018. Computer Vision and Image Processing : A Paper Review. 2(1): 28-36.

[2] Jagadeesh D. Pujari; Rajesh Yakkundimath; Abdulmunaf Syedhusain Byadgi. 2014. Automatic Fungal Disease Detection based on Wavelet Feature Extraction and PCA Analysis in Commercial Crops. (November 2013): 24-31.

[3] F. G. C. D. A. T. Febrinanto. 2019. The Implementation of K-Means Algorithm as Image

Segmenting Method in Identifying the Citrus Leaves Disease The Implementation of K-Means Algorithm as Image Segmenting Method in Identifying the Citrus Leaves Disease.

[4] A. Verroust-blondet. 2013. A Shape-based Approach for Leaf Classi fi cation using Multiscale Triangular Representation. pp. 127-134.

[5] Z. Iqbal, M. Attique, M. Sharif, J. Hussain, M. Habib and K. Javed. 2018. An automated detection and classi fi cation of citrus plant diseases using image processing techniques: A review. Comput. Electron. Agric. 153(July): 12-32.

[6] M. K. R. Gavhale. 2014. Unhealthy Region of Citrus Leaf Detection Using Image Processing Techniques. pp. 2-7.

[7] H. Ali, M. I. Lali, M. Z. Nawaz, M. Sharif and B. A. Saleem. 2017. Symptom based automated detection of citrus diseases using color histogram and textural descriptors. Comput. Electron. Agric. 138: 92-104.

[8] M. Sharif, M. Attique, Z. Iqbal, M. Faisal, M. I. Ullah and M. Younus. 2018. Detection and classi fi cation of citrus diseases in agriculture based on optimized weighted segmentation and feature selection. 150(April): 220-234.

[9] S. Sunny and M. P. I. Gandhi. 2018. An Efficient Citrus Canker Detection Method based on Contrast Limited Adaptive Histogram Equalization Enhancement. 13(1): 809-815.

[10] R. M. Prakash. 2017. Detection of Leaf Diseases and Classification using Digital Image Processing.

[11] S. Arivazhagan, R. N. Shebiah, S. Ananthi, and S. V. Varthini. 2013. Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features. 15(1): 211-217.

[12] C. Ratnasari, E.K, Ginardi, H, Fatichah. 2017. Classification of Staining Disease in Cane Leaf Image Based on Texture and Color Properties Using Segmentation-Based Gray Level Cooccurrence Matrix and Lab Color Moments. Sci. J. Inf. Syst. Technol. 3(1): 1-10.

[13] P. Chaudhary, A. K. Chaudhari and S. Godara. 2012. Color Transform Based Approach for Disease Spot Detection on Plant Leaf. 3(6): 4-9.

[14] R. R. E. Rakib Hassan Tajul Islam, Md. 2017. Color Image Segmentation using Automated K-Means clustering with RGB and HSV Color Spaces. Glob. J. Comput. Sci. Technol. Vol 17, No 2-F Glob. J. Comput. Sci. Technol.

[15] P. J. Baldevbhai and R. S. Anand. 2012. Color Image Segmentation for Medical Images using L * a * b * Color Space. 1(2): 24-45.

[16] L. Hu, M. Huang, S. Ke and C. Tsai. 2016. The distance function e ff ect on k-nearest neighbor classi fi cation for medical datasets.

[17] H. Arafat, A. Alfeilat, A. B. A. Hassanat, O. Lasassmeh and A. S. Tarawneh. 2019. Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. 00(00).

[18] A. Pandey. 2017. Comparative Analysis of KNN Algorithm using Various Normalization Techniques. (November): 36-42.

[19] J. Garcia Arnal Barbedo *et al.* 2018. Annotated Plant Pathology Databases for Image-Based Detection and Recognition of Diseases. IEEE Lat. Am. Trans. 16(6): 1749-1757.

[20] H. T. Rauf, B. A. Saleem, M. I. U. Lali, M. A. Khan, M. Sharif, and S. A. C. Bukhari. 2019. A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning. Data Br. 26: 104340.

[21] Citrus Disease Image Gallery and Fact Sheet. 2013. [Online]. Available: http://idtools.org/id/citrus/diseases/.

[22] A. Kaur. 2012. Comparison between YCbCr Color Space and CIELab Color Space for Skin Color Segmentation. 3(4): 30-33.

[23] A. B. Patil and J. A. 2016. OTSU Thresholding Method for Flower Image Segmentation. pp. 1-6.

[24] A. Chadha and N. Carolina. 2012. Comparative Study and Optimization of Feature- Extraction Techniques for Content based Image Retrieval. 52(20): 35-42.

[25] S. J. Ahn, J. H. Kim, S. M. Lee, S. J. Park and J. K. Han. 2019. CT reconstruction algorithms affect histogram and texture analysis: evidence for liver parenchyma, focal solid liver lesions, and renal cysts. Eur. Radiol. 29(8): 4008-4015.

[26] Y. Jiang, B. Bian and X. Wang. 2020. Identification of tomato maturity based on multinomial logistic regression with kernel clustering by integrating color moments and physicochemical indices. (May): 1-14.

[27] A. Kadir, L. E. Nugroho, A. Susanto and P. I. Santosa. 2011. Leaf Classification Using Shape, Color , and Texture Features. pp. 225-230.

[28] F. Liantoni and L. A. Hermanto. 2017. Adaptive Ant Colony Optimization on Mango Classification Using K-Nearest Neighbor and Support Vector Machine. 3(2): 75-79.

[29] O. R. Indriani, E. J. Kusuma, C. A. Sari, E. H. Rachmawanto, and D. R. I. M. Setiadi. 2017. Tomatoes classification using K-NN based on GLCM and HSV color space. in 2017 International Conference on Innovative and Creative Information Technology (ICITech). pp. 1-6.