



Classification of mango plants based on leaf shape using GLCM and K-nearest neighbor methods

Bona Joshua Hutasoit ^{a,1}, Herry Sofyan ^{a,2,*}, Frans Richard Kodong ^{a,3}

^a Department of Informatics, Universitas Pembangunan Nasional Veteran Yogyakarta

¹ bonajoshua98@gmail.com; ² herrysofyan@gmail.com*; ³ frkodong@gmail.com

* Corresponding Author

ARTICLE INFO

ABSTRACT



Article history

Received: November 1, 2021

Revised: November 7, 2021

Accepted: November 9, 2021

Keywords

Signature

Verification

Scale Invariant Feature

Transform

Objective: Apply the GLCM method to select mango leaf feature extraction and determine the accuracy level obtained from the K-Nearest Neighbor classification results.

Design/method/approach: Using GLCM and K-Nearest Neighbor(KNN) methods. System development using the Prototype method.

Results: The test results have been carried out using as many as 60 mango leaves compared to training data and 80:20 test data, with different accuracy. The highest accuracy is at K = 3 by 81% using 6 features, K = 6 by 78% using 5 features, and K = 7 by 74% using 4 features.

Authenticity/state of the art: The difference between this research and previous research is the pre-processing method, the type of features used, and the classification method. In this method, the mango leaf image is converted to grayscale, and a feature extraction process is carried out. Then the results of feature extraction will be classified using the K-Nearest Neighbor method. The output of this system is the result of the image classification of mango leaves, such as Kweni, Lalijowo, and Madu.

This is an open-access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



1. Introduction

Mango leaves have many similarities in shape, size, and color are almost the same as each other. It shows that there is quite a wide genetic diversity in mango leaves. So that from the public eye, it is difficult to distinguish between one type of mango tree with another even though it is a different type. For example, many people think that the type of mango planted is the sweet arum type, but it turns out to be a Manalagi mango resulting from a cross between the Gadung mango and the Golek mango. According to [1], this would undoubtedly be a problem in large-scale planting that would affect the quality of the fruit produced. Research conducted by [2] also said the same thing: in determining which mango trees to plant, many are often deceived because mango trees have many types.

Along with the rapid development of technology, it is widely used, including in the agricultural sector. The touch of technology is the main effort in solving problems in the world of agriculture. As with previous studies, mango leaves have differences that can be observed with the human eye, but

the accuracy is not perfect because they have similar leaf shapes and edges [3]. Therefore it is necessary to identify the type of mango based on the shape of the leaf. Research conducted by [4] said that the selection of leaves is an object of classification because each plant has different leaf features. In addition, leaves are easier to obtain because they do not depend on the season [5].

Research that has been related to the classification of mango species using the Backpropagation Artificial Neural Network (ANN) method on texture features uses five features, namely entropy, invariant, energy, contrast, and smoothness. In this study, the classification can distinguish two types of mango leaves using 30 leaf samples of each Gadung and Curut mango leaves. The result of classification with Artificial Neural Network (ANN) gives an accuracy of 54.24% [2].

Another study was conducted by [3] using the K-Nearest Neighbor method to determine the type of mango planted. This study uses a feature extraction process using the Sobel method. This study succeeded in obtaining an accuracy rate of 75%. Further research was carried out by [6] using the K-Means method. In this study, two leaf objects were used. RGB carries out the inputted image to grayscale processing. Then the edge detection process is carried out, then converted to a binary image, then the distance value is calculated using Euclidean distance. The last process in this study is to find the smallest distance used as a reference due to the classification. The study resulted in a rate of up to 83%. Another study was conducted by [1] using the Support Vector Machine (SVM) method. The study used 150 mango leaves, namely 50 Arumanis leaves, 50 Pancarasa leaves, and 50 Manalagi leaves, with an accuracy rate of 64.67%. The accuracy in this study is above the kernel threshold value of 50%.

Research on the classification of diseases using the K-Nearest Neighbor method was obtained with an accuracy value of 90% using the value of $k=3$, $k=7$, $k=9$, better than the Backpropagation method with an accuracy rate of 90%. Another study using the K-Nearest Neighbor method for classifying Arumanis and Manalagi mango trees in the image gives an average accuracy of 80% [7]. Then research using the K-Nearest Neighbor and Gray Level Co-Occurance Matrix method has been carried out by [8] to classify tomatoes using 100 data sets, and the highest accuracy is 100% with a value of $k = 3$. With a reasonably high accuracy value in previous studies, it is hoped that this study will obtain the same high or even higher results.

Based on previous research, this is the basis for conducting this research. This study combines the Gray Level Co-Occurance Matrix (GLCM) method to feature mango leaf species with the K-Nearest Neighbor method to classify previously obtained characteristics. This research is expected to produce an intelligent system that can classify mangoes with a high level of accuracy so that the results of this study can later help the public find out the type of mango based on the shape of the leaf.

2. Method

This research was conducted in several stages, which can be seen in Fig. 1. The detail of this process is shown in the description from the acquisition until classification.

2.1. Image Acquisition

The image acquisition stage was carried out by collecting mango leaves with 60 leaves from 3 types of leaves compared to training data and 80:20 test data. The image is taken using a Canon D90 camera and produces an image in .jpg format and 300 x 150 pixels. The system will use this cropped image for processing.

2.2. Image Preprocessing

The stages of image pre-processing in this study can be seen in Fig. 2. The resizing process is one part of the pre-processing process—resizing the image to 261×161. The purpose of the resizing process is to uniform the image used in this study [9].

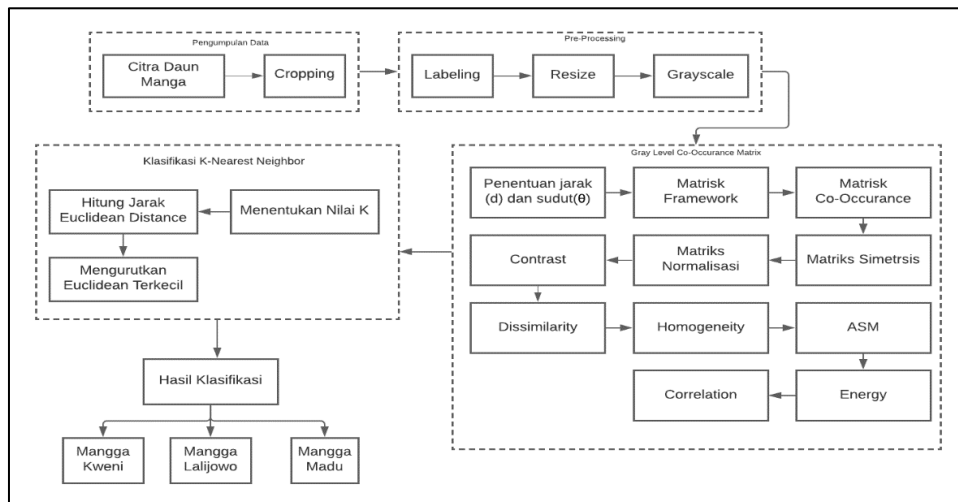


Fig. 1. Research Methodology

A grayscale image is an image whose pixel value represents the degree of gray or white intensity. To change the RGB image to grayscale (gray) can be done with the following equation [10]. Fig. 3 is a leaf image that has been converted to grayscale.

$$\text{Grayscale} = 0.299R + 0.587G + 0.144B \quad (1)$$

Table 1. Grayscale Results

Image	Grayscale	Types of leaf
		Kweni
		Lalijowo
		Madu

2.3. Gray Level Co-Occurance Matrix

Gray Level Co-Occurance Matrix is a matrix that describes the frequency of occurrence of pairs of two pixels with a certain intensity in a certain distance and direction in the image [11]. Distances in GLCM are represented in pixels, and angles are represented in degrees. This study uses four angle directions in the adjacency between pixels, namely 0°, 45°, 90°, and 135° where the distance between them is 1 pixel. In this study, the features used are Angular Second Moment, Contrast, Dissimilarity, Correlation, Inverse Difference Moment, and Energy [12].

1. Angular Second Moment

$$ASM = \sum_{i=1}^L \sum_{j=1}^L (GLCM(i, j))^2 \quad (2)$$

2. Contrast

$$\text{Contrast} = \sum_i k^2 = [\sum_i \sum_j GLCM(i, j)] \quad (3)$$

$$|i - j| = k \quad (4)$$

3. Dissimilarity

$$\text{Dissimilarity} = \sum_{i=1}^L \sum_{j=1}^L |i - j| \text{GLCM}(i, j) \quad (5)$$

4. Homogeneity

$$\text{Homogeneity} = \sum_{i=1}^L \sum_{j=1}^L \frac{\text{GLCM}(i, j)}{1 + (i - j)^2} \quad (6)$$

5. Energy

$$\text{Energy} = \sqrt{\sum_{i=1}^L \sum_{j=1}^L (\text{GLCM}(i, j))^2} \quad (7)$$

6. Correlation

$$\text{Correlation} = \sum_{i=1}^L \sum_{j=1}^L \frac{(i - \mu_i)(j - \mu_j) \text{GLCM}(i, j)}{(\sigma_i \sigma_j)} \quad (8)$$

2.4. K-Nearest Neighbor

After the image goes through the feature extraction process, then the image is used as input for the K-Nearest Neighbor process. The K-Nearest Neighbor (KNN) method for classifying objects based on data that has the closest distance to the object [13].

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (9)$$

Noted:

d: proximity

x: training data

y: data testing

n: the number of attributes 1 to m

i: individual attribute between 1 and n

Below are the steps for the K-Nearest Neighbor classification method [14].

1. Determine the parameter K (number of closest neighbors)
2. Calculating the square of the Euclidean Distance of each object to the given sample data
3. Sort objects into groups that have the smallest Euclidean distance.
4. Collect category Y (nearest neighbor classification)
5. Using the nearest neighbor category will produce a new data class (prediction). The best value of k for this algorithm depends on the data. In general, a high value of k will reduce noise in the classification.

3. Results and Discussion

The testing phase aims to determine the level of success of the system that has been built. In this study, system testing is done by implementing a confusion matrix. The confusion matrix shows the value of the evaluation metrics used in the tests in this study, namely Accuracy, Precision, and Recall.

Tests were carried out using the GLCM and K-Nearest Neighbor models. The model testing carried out in this study uses the k-fold cross-validation and confusion matrix methods. The data set used 60 data with each division of 80:20.

3.1. K-Fold Cross-Validation

K-fold cross-validation is a validation method that functions to calculate the average success rate of the model that has been created. This test uses k-fold cross-validation with a value of k = 5 or iterations five times, where each iteration produces a confusion matrix value.

The first test of the KNN model with a K = 7 using four types of features contained in GLCM. The types of features used are contrast, dissimilarity, homogeneity, ASM. The tests can be seen in Table 2.

Table 2. Testing the Confusion Matrix Model KNN K = 7 with 4 Features

Fold	Confusion Matrix											
	True Positive			false Positive			True negative			false negative		
	<i>Kwe</i>	<i>Lal</i>	<i>Mad</i>	<i>Kwe</i>	<i>Lal</i>	<i>Mad</i>	<i>Kwe</i>	<i>Lal</i>	<i>Mad</i>	<i>Kwe</i>	<i>Lal</i>	<i>Mad</i>
1	4	2	3	1	1	1	7	7	7	0	2	1
2	3	3	4	1	0	1	7	8	7	1	1	1
3	2	4	3	0	2	1	8	7	7	2	0	1
4	3	2	4	1	1	1	7	7	7	1	2	0
5	4	1	3	2	1	1	6	7	7	0	3	1

Based on the table above, the confusion matrix results are the values of TP, FP, TN, and FN. This calculation will be used to calculate the value of accuracy, Precision, and Recall of the KNN model using four features of GLCM. The following are test results from each iteration can be seen in Table 3.

Table 3. Accuracy, Precision, and Recall Results of the KNN K = 7 models with four features

Fold	Accuracy	Precision	Recall
1	75%	73.33%	75%
2	83%	85%	83%
3	75%	80%	75%
4	75%	73%	75%
5	66%	63%	66%
Average	74%	74.8%	74.8%

Table 3 is the result of the calculation of Accuracy, Precision, and Recall. The highest accuracy results are in iteration 2 with an accuracy value of 83%, Precision of 85%, and Recall of 83% in classifying types of Kweni mango, Lalijowo mango, and Madu mango using 4 GLCM features.

The second model test with KNN with a K = 6 using the five features found in GLCM. The types of features used are contrast, dissimilarity, homogeneity, ASM, and energy. Testing can be seen in Table 4.

Table 4. Testing the Confusion Matrix Model KNN K = 6 with 5 Features

Fold	Confusion Matrix											
	True Positive			false Positive			True negative			false negative		
	<i>Kwe</i>	<i>Lal</i>	<i>Mad</i>	<i>Kwe</i>	<i>Lal</i>	<i>Mad</i>	<i>Kwe</i>	<i>Lal</i>	<i>Mad</i>	<i>Kwe</i>	<i>Lal</i>	<i>Mad</i>
1	4	3	4	0	0	1	8	7	7	0	1	1
2	4	1	4	3	0	0	5	8	8	0	3	0
3	4	2	2	0	2	2	6	7	7	0	0	0
4	4	3	3	0	1	1	8	7	7	0	1	1
5	2	3	4	0	1	2	8	7	6	2	1	0

Based on the table above, the confusion matrix results are the values of TP, FP, TN, and FN. This calculation will be used to calculate the value of accuracy, Precision, and Recall of the KNN model using the five features of GLCM. The following are the test results from each iteration can be seen in Table 5.

Table 5. Accuracy, Precision, and Recall Results of the KNN K = 6 model with 5 features

Fold	Accuracy	Precision	Recall
1	91%	93%	91%
2	75%	85%	75%
3	66%	66%	66%

4	83%	83%	83%
5	75%	80%	75%
Average	78%	81.4%	78%

Table 5 is the result of the calculation of Accuracy, Precision, and Recall. The highest accuracy results are found in iteration 1 with an accuracy value of 91%, Precision of 93%, and Recall of 91% in classifying types of Kweni mango, Lalijowo, and Madu mango using 5 GLCM features.

Furthermore, testing the KNN model system with $K = 3$ using six features contained in GLCM. The types of features used are contrast, dissimilarity, homogeneity, ASM, energy, and correlation. Tests using six features can be seen in Table 6.

Table 6. Testing the Confusion Matrix Model KNN $K = 3$ with 6 Features

Fold	Confusion Matrix											
	True Positive			False Positive			True Negative			False Negative		
	Kwe	Lal	Mad	Kwe	Lal	Mad	Kwe	Lal	Mad	Kwe	Lal	Mad
1	3	3	4	1	1	0	7	7	8	1	1	0
2	4	3	3	1	1	0	7	7	8	0	1	1
3	4	1	3	1	1	2	7	7	6	0	3	1
4	4	3	4	0	0	1	8	8	7	0	3	0
5	4	3	3	1	1	0	7	7	8	0	1	0

Based on Table 5, the confusion matrix results are the values of TP, FP, TN, and FN. This calculation will be used to calculate the value of accuracy, Precision, and Recall of the KNN model using six features of GLCM. The following are test results from each iteration can be seen in Table 6.

Table 6 Accuracy, Precision, and Recall Results of the KNN $K = 3$ model with 6 features

Fold	Accuracy	Precision	Recall
1	83%	83%	83%
2	83%	85%	83%
3	66%	63%	66%
4	91%	93%	91%
5	83%	83%	81%
Average	81.2%	81.4%	80.8%

Based on Table 6, the highest accuracy is in iteration 4 with an accuracy rate of 91%, Precision of 93%, and Recall of 91% in classifying Kweni, Lalijowo, and Madu mango types using 6 GLCM features.

4. Conclusion

Based on the implementation of the tests carried out, it can be concluded that this system can classify types of Kweni mango, Lalijowo, and Madu mango well. It can be proven by the accuracy obtained from system testing. Based on system testing obtained from the confusion matrix, the high accuracy depends on the number of GLCM features used. The average accuracy of using four features is 74%, five features 78%, and six features 81%. So it can be concluded that the type of features and the number of GLCM features used significantly affect the high accuracy of the K-Nearest Neighbor method.

In this study, there are still some shortcomings that are expected to be developed for further research. Some things that can be done to develop this research are adding a pre-processing process to normalize the light because this research has not been able to normalize the light in the image. Then the thing that can be developed in further research is to add an edge detection method on leaves. Further research is also recommended to add more varied types of mango leaves because Indonesia has approximately 30 types of mangoes scattered in various regions [15].

References

- [1] S. Suhendri, F. Muhammad Muharam, and K. Aelani, "Implementasi Support Vector Machine (Svm) Untuk Klasifikasi Jenis Daun Mangga Menggunakan Metode Gray Level Co-Occurrence Matrix," *KOPERTIP J. Ilm. Manaj. Inform. dan Komput.*, vol. 1, no. 3, pp. 93–100, 2017, doi: 10.32485/kopertip.v1i03.22.
- [2] eko prasetyo soffiana agustin, "Klasifikasi jenis pohon mangga gadung dan curut berdasarkan tesktur daun," pp. 58–64, 2011.
- [3] M. Nasir, "Klasifikasi Jenis Mangga Berdasarkan Bentuk Daun Menggunakan Metode K-Nearest Neighbor," vol. 3, no. 2, pp. 87–91, 2018.
- [4] E. Budianita, T. Ulfadhayani, and F. Yanto, "Implementasi Algoritma Canny Dan Backpropagation Untuk Mengklasifikasi Jenis Tanaman Mangga," no. November, pp. 13–21, 2019.
- [5] A. Patriot, S. Pamungkas, N. Nafi, and N. Q. Nawafilah, "K-NN Klasifikasi Kematangan Buah Mangga Manalagi Menggunakan $L * A * B$ dan Fitur Statistik," vol. 4, pp. 1–8, 2019.
- [6] S. Wibowo, "Penentuan Jenis Buah Mangga Berdasarkan Bentuk Daun Menggunakan Metode K-Means," vol. 01, no. 02, pp. 16–24, 2017.
- [7] H. Judul, D. A. Kurniawan, P. Studi, T. Informatika, F. Teknik, and U. M. Jember, "KLASIFIKASI JENIS POHON MANGGA ARUMANIS DAN MANALAGI BERDASARKAN EKTRAKSI FITUR (GLCM)
KLASIFIKASI JENIS POHON MANGGA ARUMANIS DAN MANALAGI BERDASARKAN EKTRAKSI FITUR (GLCM)," 2016.
- [8] O. R. Indriani, E. J. Kusuma, C. A. Sari, E. H. Rachmawanto, and D. R. I. M. Setiadi, "Tomatoes classification using K-NN based on GLCM and HSV color space," *Proc. - 2017 Int. Conf. Innov. Creat. Inf. Technol. Comput. Intell. IoT, ICITech 2017*, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/INNOCIT.2017.8319133.
- [9] T. Pengenalan and P. Tanda, "Analisa kinerja jaringan syaraf tiruan back propagation terhadap pengenalan pola tanda tangan," 1993.
- [10] E. Susilowati, U. Gunadarma, K. Baru, and K. Clustering, "Konversi Citra RGB Ke Citra HSV Dan HCL Pada Citra Jeruk Medan," vol. 2, 2018.
- [11] E. Budianita, J. Jasril, and L. Handayani, "Implementasi Pengolahan Citra dan Klasifikasi K-Nearest Neighbour Untuk Membangun Aplikasi Pembeda Daging Sapi dan Babi Berbasis Web," *J. Sains dan Teknol. Ind.*, vol. 12, no. Vol 12, No 2 (2015): Juni 2015, pp. 242–247, 2015, [Online]. Available: <http://ejournal.uin-suska.ac.id/index.php/sitekin/article/view/1005>.
- [12] P. K. Mall, P. K. Singh, and D. Yadav, "GLCM based feature extraction and medical X-RAY image classification using machine learning techniques," *2019 IEEE Conf. Inf. Commun. Technol. CICT 2019*, 2019, doi: 10.1109/CICT48419.2019.9066263.
- [13] R. Rahmadiano, E. Mulyanto, and T. Sutojo, "Implementasi Pengolahan Citra dan Klasifikasi K-Nearest Neighbor untuk Mendeteksi Kualitas Telur Ayam," *J. VOI (Voice Informatics)*, vol. 8, no. 1, pp. 45–54, 2019, [Online]. Available: <http://voi.stmik-tasikmalaya.ac.id/index.php/voi/article/view/164>.
- [14] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019, doi: 10.30591/jpit.v4i1.1253.
- [15] G. F. Rohmi, W. B. Zulfikar, and Y. A. Gerhana, "Implementasi Citra Digital Berdasarkan Nilai HSV Untuk Mengidentifikasi Jenis Tanaman Mangga Menggunakan Algoritma K-Nearest Neighbor," *Insight*, vol. 1, no. 1, pp. 142–147, 2018.