Parijoto Fruits Classification using K-Nearest Neighbor Based on Gray Level Co-Occurrence Matrix Texture Extraction

I U W Mulyono¹, T C Lukita¹, C A Sari¹, D R I M Setiadi¹, E H Rachmawanto¹, A Susanto¹, M D M Putra² and D A Santoso¹

¹ Department of Informatics Engineering, Dian Nuswantoro University, 207 Imam Bonjol Street, Semarang 50131, Indonesia
² Faculty of Electronics and Information Engineering, South China University of Technology, Guangzhou, China

Abstract. Parijoto fruit is a typical fruit that grows around Muria Mountain, Kudus Regency and Mount Merapi in Yogyakarta, Indonesia. This fruit has many health benefits, especially for pregnant women. This fruit production is not much because of its limited growth around Mount Muria. So parijoto fruit is made into powder drink products and syrup, so that it can be consumed in a longer period of time and not only during the harvest. To get a good processed product requires good quality ingredients. Parijoto fruit needs to be sorted and classified. Current technology allows classification to be done by digital image processing. The Gray Level Co-occurrence Matrix (GLCM) method is proposed to extract the texture features from the parijoto fruit and then classify them using the K-Nearest Neighbor (KNN) method. GLCM can describe a spatial linear relationship of the frequency at which gray values are determined by other gray values in one area of investigation. It can simply use the statistical approach of appearance or histogram of the image matrix. In this way, information will be easily relative position of neighboring pixels that are suitable for the classification process using KNN. KNN was chosen because this method was proven to be used for relatively few datasets, but a normalization process was needed to increase accuracy. Based on the results of the implementation of the GLCM and KNN methods for parijoto fruit classification the classification accuracy was 80%.

1. Introduction
Parijoto Fruit is one of the typical fruits that grow around Muria Mountain, Kudus Regency and Mount Merapi in Yogyakarta, Indonesia. Parijoto fruit has a sour and bitter taste that has the benefits of fruit and leaves. Parijoto fruit contains nutrients that are good for pregnant women to consume for fertility. Parijoto fruit and leaves can also be used as a drug for thrush, sore throat, antibacterial, and diarrhea because they contain antimicrobial activity against Escherichia coli and Staphylococcus aureus[1,2]. So that Parijoto fruit can be consumed in a longer period of time, much Parijoto fruit is processed to be made into powder drink products or in the form of syrup. With the processing of Parijoto fruit and its limited number, the economic value of Parijoto fruit is quite high and is used as a source of income for the local community around Mount Muria. The quality of the Parijoto fruit will affect the quality of the product, so the Parijoto fruit needs to be sorted by quality to get good product quality.

Current technology can be used to help humans classify fruit quality[3–5] and of course, it can also be implemented on Parijoto fruit. Some classification methods that are widely applied in digital image processing such as K-Nearest Neighbor (KNN)[5–8], Support Vector Machine (SVM)[9–11], and Naïve Bayes (NB)[12,13]. Of the several classification methods, KNN is a relatively simplest method and has several advantages, which can overcome the probability density, consolidate the calculation results by
determining the number of neighbors (k) and still powerful for calculations with few parameters[5,6,14,15]. While SVM can work effectively in a high-dimensional space, but can efficiently memory because it uses a subset of training points and uses several kinds of kernels [9]. Whereas Naïve Bayes works with superior probability methods without the need for numerical optimization and can be used in a variety of binary, polynomial or other data set types. [12,15].

To carry out the classification process, image features that are extracted are needed first. One method of feature extraction that is widely used in various studies is the Gray Level Co-occurrence Matrix (GLCM)[7–10,16–19]. GLCM is an extraction feature of image textures that is very powerful for use in the image classification process. GLCM can describe a spatial linear relationship of the frequency at which gray values are determined by other gray values in one area of investigation. It can simply use the statistical approach of appearance or histogram of the image matrix. In this way, information will be easily relative position of neighboring pixels which is useful for the classification process later[16]. Classification using the KNN method by extracting GLCM features in various objects in digital images has been proven to have very good accuracy [4,5,7,8,20]. KNN is suitable because it is effective in the classification of a few data sets, but to produce optimal accuracy the data set must be preprocessed for data normalization[15]. So in this research, the KNN method will be applied to classify the GLCM features from the Parijoto fruit image data set. The results of the verification will be measured by a confusion matrix to determine the accuracy of the image classification.

2. Research Method

In this study, the image of parijoto fruit was classified into three different classes, namely good, defect, and rotten. The image is first acquired and then labeling is carried out based on three conditions.

Figure 1. Research method

The next step is preprocessing which consists of resizing and converting to grayscale to normalize. The image data is then divided into two, namely training data and test data. The next step is to convert all
images to grayscale. Then the image texture feature is extracted using the GLCM method. The results of the extraction of the training data feature are used to train the system, while the results of the extraction of the data testing feature are used to test the system using the KNN classifier. Finally, to calculate the level of accuracy used confusion matrix. Fig. 1 describes the method used in this study.

2.1. Image Acquisition and Labelling
At this stage, the image was acquired by taking the image of the Parijoto using the Samsung Galaxy J7 Prime's smartphone camera with a resolution of 13 megapixels. The number of images taken was 105 images which consisted of 35 images of Parijoto fruits each in good condition, defect condition, and rotten condition. Fig. 2 is a sample of Parijoto fruit in different conditions.

![Figure 2](image.png)

Figure 2. Parijoto fruit {(a) good condition (b) rotten condition (c) defects conditions}

2.2. Preprocessing
At this stage the image is resized using the imresize() function on Matlab, the image is resized to 480 × 480 pixels. The results of the resize process are then converted to a grayscale image with the rgb2gray() function. The sample image conversion results are in Fig. 3. After the entire image has been finished processing, the image is divided into two data groups, namely 75 training data and 30 testing data.

![Figure 3](image.png)

Figure 3. Sample of preprocessing results

2.3. GLCM Extraction
GLCM is one of the methods of feature extraction in texture and is included in the second-order statistics as a technique for obtaining values by calculating the probability of a neighboring relation of two pixels at a specified distance (d) and angle (θ) as shown in Fig. 4. The features taken are contrast, correlation, energy, and homogeneity[3,5,19].

![Figure 4](image.png)

Figure 4. The angle of GLCM extraction

Contrast is a calculation that relates to the amount of diversity of grayscale intensity in the image, contrast can be calculated with the formula (1).
\[ \text{corr} = \sum_{x} \sum_{y} (x - \mu_x) (y - \mu_y) \frac{p(x, y)}{\sigma_x \sigma_y} \]  

(2)

Energy can determine the intensity of gray with a certain concentration of pairs in the image matrix, energy can be calculated by the formula (3).

\[ \text{corr} = \sum_{x} \sum_{y} p(x, y)^2 \]  

(3)

Homogeneity is the number of gray levels that will get higher if the uniform and inverse GLCM are high, homogeneity can be calculated by the formula (4).

\[ \text{corr} = \sum_{x} \sum_{y} \frac{p(x, y)}{1 + |x - y|} \]  

(4)

Where \( x \) is the row value in the co-occurrence matrix; \( y \) is the column value in the co-occurrence matrix; \( p(x, y) \) are row (x) and column (y) values in the co-occurrence matrix; \( \mu_x \) is the average value of the line (x) calculated by \( \sum_{x} \sum_{y} x p(x, y) \); \( \mu_y \) is the average column value (y) calculated by \( \sum_{y} \sum_{x} y p(x, y) \); \( \sigma_x \) is the variance in the matrix (x) which can be calculated with \( \sqrt{\sum_{x} \sum_{y} (x - \mu_x)^2 p(x, y)} \); \( \sigma_y \) is the variance in the matrix (y) which can be calculated with \( \sqrt{\sum_{x} \sum_{y} (y - \mu_y)^2 p(x, y)} \). From the extraction, results are used as data to be classified at a later stage.

### 2.4. KNN Classification

Image classification is the stage where the process of grouping image types based on their class of training image data that has been known to extract texture results using GLCM with input from test sample image data. The KNN algorithm is used to classify the image/data ready for testing based on the training data image which is a reference of the KNN algorithm to determine the class of test data images using k values. The k value is a reference to the number of closest neighbors in determining the type of class in the test data image. The training image data and test image data are in the form of GLCM calculation results. After determining the k value, then do the distance calculation to determine the proximity of the sample training image data and test image data. Proximity distance calculation can use Euclidean Distance which can be calculated with the formula (5).

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

(5)

Where \( d \) = distance; \( x \) = training data; \( y \) = testing data; \( i \) = number of nearest neighbors; and \( n \) = number of images

### 2.5. Confusion Matrix Accuracy Measurement

The level of accuracy is the result of determining image classification data based on classifications. Calculation of the accuracy test can be done with the Confusion Matrix can be written using the formula (6).
\[ acc = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \]  

(6)

Where TP is a number of true positives; TN is a number of true negatives; FN is a number of false positives; FP is a number of false negatives.

3. Implementation and Results

At this stage, the image classification is carried out according to the proposed method where the image is extracted by the GLCM method. The resulting features are contrast, correlation, energy and homogeneity at angles 0°, 45°, 90°, 135°. As for the classification process, the KNN method is used with values K = 1, 3, 5, 7, and 9. From the classification results, the best accuracy value obtained is 80% at an angle of 90°. In detail, the results of the GLCM feature extraction for each angle with a value of K = 1 are presented in Fig. 5 to Fig. 8.

Figure 5. GLCM extraction results in 0° where K=1

Figure 6. GLCM extraction results in 45° where K=1
Figure 7. GLCM extraction results in 90° where K=1

Figure 8. GLCM extraction results in 135° where K=1

Actually, the GLCM method can produce five features, one feature that is not used is entropy, this feature is not used to speed up the computing process because in some studies these features are not used and in research conducted by Irawan et al.[6] evidenced the results of image classification without entropy features do not have a significant effect. In Fig.5 to Fig. 7 presented a GLCM feature value of 75 training images, where image no 1-25 image of parijoto fruit with good condition (T1), image no. 26-50 are parijoto fruit images with rotten conditions (T2) and image no 51-75 are parijoto fruit with a defective condition (T3). While the results of the classification test are presented in Table 1.

Table 1. Classification results

| Image   | Target results | Classification θ results |
|---------|----------------|--------------------------|
|         |                | 0° | 45° | 90° | 135° |
| T1 (1).jpg | Good          | Good | Good | Good | Good |
| T1 (2).jpg | Good          | Good | Good | Good | Good |
| T1 (3).jpg | Defect        | Defect | Defect | Defect | Defect |
| T1 (4).jpg | Good          | Defect | Good | Good | Good |
The results presented in Table 1 are then calculated using a confusion matrix. Confusion matrix calculations are performed at each angle 0˚, 45˚, 90˚, 135˚.

Table 2. Confusion matrix results on 0˚

| Target Results | Good | Rotten | Defect |
|----------------|------|--------|--------|
| Good           | 5    | 1      | 4      |
| Rotten         | 0    | 10     | 0      |
| Defect         | 0    | 4      | 6      |

Then the accuracy value obtained from Table 2 is $Acc = \frac{5+10+6}{30} \times 100\% = 70\%$

Table 3. Confusion matrix results on 45˚

| Target Results | Good | Rotten | Cacat |
|----------------|------|--------|-------|
| Good           | 6    | 1      | 3     |
| Rotten         | 0    | 10     | 0     |
| Defect         | 1    | 4      | 5     |

Then the accuracy value obtained from Table 3 is $Acc = \frac{6+10+5}{30} \times 100\% = 70\%$
Table 4. Confusion matrix results on 90°

| Target Results | Good | Rotten | Defect |
|----------------|------|--------|--------|
| Good           | 8    | 1      | 1      |
| Rotten         | 0    | 10     | 0      |
| Defect         | 0    | 4      | 6      |

Then the accuracy value obtained from Table 4 is \( Acc = \frac{8+10+6}{30} \times 100\% = 80\% \)

Table 5. Confusion matrix results on 135°

| Target Results | Good | Rotten | Defect |
|----------------|------|--------|--------|
| Good           | 6    | 3      | 1      |
| Rotten         | 0    | 10     | 0      |
| Defect         | 0    | 4      | 6      |

Then the accuracy value obtained from Table 4 is \( Acc = \frac{6+10+6}{30} \times 100\% = 73.33\% \)

4. Conclusions

This research implements the GLCM and KNN methods to classify parijoto fruit. Parijoto fruit with good quality has a smoother texture compared to parijoto fruit that is noted or rotten, therefore the GLCM method was chosen. While the KNN Method is sufficient with a little training data. Based on the results of implementation and testing, this method has an accuracy rate of 80%, this accuracy is good enough, but in subsequent studies, it can be combined with color feature extraction to optimize accuracy.

5. References

[1] Octaviani I 2016 Aktivitas Antibakteri Ekstrak Daun Parijoto (Medinilla Speciosa) Terhadap Escherichia Coli Dan Staphylococcus Aureus UAJY Repos. 111–12
[2] Sidiq Y and Mumpuni K E 2014 Identifikasi Variasi Genetik Parijoto (Medinilla Javanensis(Bl.)Bl. dan Medinilla Verrucosa (Bl.) Bl.) dengan Penanda Molekular Sebagai Sumber Belajar Proceeding Biology Education Conference pp 667–72
[3] Kumar C, Chauhan S, Alla R N and Mounica Gurram H 2015 Classifications of citrus fruit using image processing-GLCM parameters Int. Conf. Commun. Signal Process. 1743–7
[4] Sutojo T, Tirajani P S, Setiadi D R I M, Sari C A and Rachmawanto E H 2017 CBIR for classification of cow types using GLCM and color features extraction International conferences on Information Technology, Information Systems and Electrical Engineering (Yogyakarta: IEEE) pp 182–7
[5] Indriani O R, Kusuma E J, Sari C A, Rachmawanto E H and Setiadi D R I M 2017 Tomatoes classification using K-NN based on GLCM and HSV color space international Conference on Innovative and Creative Information Technology (Salatiga: IEEE) pp 1–6
[6] Irawan C, Ardyastiti E N, Setiadi D R I M, Rachmawanto E H and Sari C A 2018 A Survey: Effect of the Number of GLCM Features on Classification Accuracy of Lasem Batik Images using K-Nearest Neighbor International Seminar on Research of Information Technology and Intelligent Systems (Yogyakarta: IEEE) pp 33–8
[7] Htay T T and Maung S S 2018 Early Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbor (k-NN) on Mammography image International Symposium on Communications and Information Technologies (Bangkok: IEEE) pp 171–5
[8] Safira L, Irawan B, and Setianingsih C 2019 K-Nearest Neighbour Classification and Feature Extraction GLCM for Identification of Terry’s Nail IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (Bali: IEEE) pp 98–104

[9] Kavitha J C and Suruliandi A 2016 Texture and color feature extraction for classification of melanoma using SVM International Conference on Computing Technologies and Intelligent Data Engineering (Kovilpatti: IEEE) pp 1–6

[10] Solanki V, Patel V and Pati S 2018 Brain MRI Image Classification using Image Mining Algorithms International Conference on Computing Methodologies and Communication (IEEE) pp 516–9

[11] Rahmayuna N, Rahardwika D S, Sari C A, Setiadi D R I M and Rachmawanto E H 2018 Pathogenic Bacteria Genus Classification using Support Vector Machine 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (IEEE) pp 23–7

[12] Kusuma A, Setiadi D R I M and Putra M D M 2018 Tomato Maturity Classification using Naive Bayes Algorithm and Histogram Feature Extraction J. Appl. Intell. Syst. 3 39–48

[13] Wijaya E 2018 Implementation Analysis of GLCM and Naive Bayes Methods in Conducting Extractions on Dental Image IOP Conf. Ser. Mater. Sci. Eng. 407

[14] Enriko I K A, Suryanegara M and Gunawan D 2016 Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient’s Health Parameters J. Telecommunication, Electron. Comput. Eng. 8 59–65

[15] JadHAV S D and Channe H P 2016 Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques Int. J. Sci. Res. 5 1842–1842

[16] Liu Q and Liu X 2013 Feature Extraction of Human Viruses Microscopic Images Using Gray Level Co-occurrence Matrix International Conference on Computer Sciences and Applications (Wuhan: IEEE) pp 619–22

[17] Arabi P M, Joshi G and Vamsha Deepa N 2016 Performance evaluation of GLCM and pixel intensity matrix for skin texture analysis Perspect. Sci. 8 203–6

[18] Mukherjee G, Chatterjee A and Tudu B 2016 Study on the potential of combined GLCM features towards medicinal plant classification International Conference on Control, Instrumentation, Energy & Communication (IEEE) pp 98–102

[19] Seul A 2018 Monochromatic textures’ features extraction using extended GLCM approach for classification of autonomous cleaning robot work area International Interdisciplinary PhD Workshop (IEEE) pp 220–5

[20] Riandini and Delimayanti M K 2017 Feature extraction and classification of thorax x-ray image in the assessment of osteoporosis International Conference on Electrical Engineering, Computer Science and Informatics (IEEE) pp 1–5