

Pears Classification Using Principal Component Analysis and K-Nearest Neighbor

Moh. Arie Hasan
Sekolah Tinggi Mamajemen Informatika dan Komputer
Nusa Mandiri
Jakarta, Indonesia
14002250@nusamandiri.ac.id

Arief Setya Budi
Sekolah Tinggi Mamajemen Informatika dan
Komputer Nusa Mandiri
Jakarta, Indonesia
arief.aeg@bsi.ac.id

Submitted: Feb 20, 2020

Accepted: Mar 7, 2020

Published: Apr 1, 2020

Abstract— Pears is a fruit that is widely available in tropical climates such as in Western Europe, Asia, Africa and one of them is Indonesia. There are many types of pears found in Indonesia. Types of pears can be distinguished from the color, size, and shape. But it is still difficult for ordinary people to get to know the types of pears. This is what gave rise to the idea to make a study related to image processing to classify three types of pears, namely abate, red and William pears in order to help determine the differences in the three types of pears. The dataset used is 99 pears. The pear type classification process is carried out by testing the pear image based on existing training data. Stages of training and testing used consisted of image segmentation in the form of RGB and HSV conversion for feature extraction. Furthermore, by using Principal Component Analysis (PCA) data is grouped and K-Nearest Neighbor (KNN) is used to determine data classification. The use of adequate training data will further improve the accuracy of the classification of pears. The final results of this study indicate the accuracy of the classification of pears for all three types of pears by 87.5%.

Keywords— Pears, Principal Component Analysis, Image Processing, K-Nearest Neighbor

I. INTRODUCTION

Pears are one of the many imported fruits found in the Indonesian market both in traditional markets and in supermarkets. This fruit has approximately 30 types, but in general there are only 3 types if distinguished by their skin color, namely green, yellow and red. In terms of shape, pears have a variety of shapes. Some pears are round like apples and some are shaped like bells (Octavia, Jesslyn, & Gasim, 2016). Because it has a shape that is almost similar to other fruits, many ordinary people find it difficult to classify a type of pear.

In image processing, computer graphics, and computer vision can be considered as "translating" input images into corresponding output images (Isola, Zhu, ..., & 2017, n.d.). An important part of image processing is color. Besides being able to be seen visually, the image also has important information in the presentation of the image quality. External color

features and firmness of internal features are the most important factors observed by consumers (wholesalers or retailers) to determine fruit quality (Sehgal & Goel, 2016).

In a previous study, backpropagation neural networks were used with images incorporating grayscale, HSV, and $L * a * b *$ to identify pear (Octavia et al., 2016). Research on Principal Component Analysis (PCA) has been carried out to identify floral patterns (Herfina, 2013) and application of distance transform method for identify handwriting pattern using PCA (Husein, 2019). Research using the K-Nearest Neighbor Method to perform Leaf Classification with Enhanced Image Features (Liantoni, 2016) and Fruits Recognition based on Texture Features and K-Nearest Neighbor (Ariffin, Mustaffa, Abdullah, & Nasharuddin, 2018).

Based on these problems, we need a way to classify three types of pears namely abate, red and william pears using image processing. The process of classifying pears using PCA and KNN methods. The purpose of this study is to classify the types of abate, red and william pears from existing images. Furthermore, the image is processed using the matlab application to get the results of image classification of abate, red and william pears.

II. LITERATURE REVIEW

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a well-established method for feature extraction and dimensionality reduction (Subasi & Gursoy, 2010). PCA is a method that involves a mathematical procedure that changes and transforms a large number of correlated variables into a small number of uncorrelated variables, without losing important information in them. A number of two-dimensional images of each three-dimensional object that will be recognized, collected to represent the object as a reference image. From the set of reference images, feature extraction will then be performed to obtain characteristic information (characteristics) of the object. The result of feature extraction is used for multi object orientation recognition process. Principal Component Analysis is widely used to project or convert a large data set into a form of data presentation with a smaller size. PCA transformation to a large data space will produce a number of orthonormal basis vectors in the form of a collection of eigenvectors from a particular covariant matrix that can optimally present the data distribution (Muhammad & Isnanto Riza, n.d.). The concept of using PCA includes the calculation of standard deviation values, covariance matrices, eigenvalue values and eigen vectors. PCA can use the method of warranty or correlation. If needed, the data is standardized first so it approaches the standard normal distribution. In this case the covariance method is used with the following algorithm:

1. Collecting data in the form of gray-level matrix X of size M x N. Suppose that is a vector N x 1.:

2. Calculate average:

$$x = \frac{1}{M} \sum_{i=1}^M x_i \tag{1}$$

3. Calculate the average difference:

$$\Phi_i = x_i - x \tag{2}$$

4. Determine the covariance matrix From matrix $X=[\Phi_1 \ \Phi_2 \ \dots \ \Phi_M]$ (matriks NxM), Hitung kovarian:

$$C = \frac{1}{M} \sum_{n=1}^M \phi_n \phi_n^T = XX^T \tag{3}$$

5. Determine the characteristic values and characteristic vectors of the covariance matrices

$$C : \lambda_1 > \lambda_2 > \dots > \lambda_N \tag{4}$$

and

$$C : u_1, u_2, \dots, u_n \tag{5}$$

6. Sort the characteristic vector u and the characteristic value λ in the diagonal matrix in descending order according to the greatest cumulative probability value for each characteristic vector so that the dominant characteristic values are obtained (Herfina, 2013).

The use of the PCA method aims to group data into several classes which are then grouped so that they can classify images of abate, red and william pears.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) algorithm is a method for classifying objects based on learning data that is the closest distance to the object. Learning data is projected into multi-dimensional space, where each dimension is representing features of data. The purpose of the KNN algorithm is to classify new objects based on attributes and training samples where the results of the new test samples are classified based on the majority of the categories in the KNN. In the classification process, this algorithm does not use any model to be matched and only based on memory (Liantoni, 2016).

The working principle of the KNN is to find the closest distance between the data to be evaluated with its closest neighbor K in the training data. The training data is projected into a multi-dimensional space, where each dimension represents the features of the data. This space is divided into sections based on the classification of training data. A point on this space is marked by class c, if class c is the most common classification found in the nearest k of the point. Near or far neighbors are usually calculated

based on Euclidean distances with the following formula:

$$d_i = \sqrt{\sum_{i=1}^p (x_{2i} - x_{1i})^2} \tag{6}$$

With x_1 = sample data, x_2 = test data, i = data variable, $dist$ = distance, p = data dimension. In the learning phase, this algorithm only stores feature vectors and classifications of learning data. In the classification phase, the same features are calculated for the test data. The distance of this new vector to all the learning data vectors is calculated, and the closest number of k is taken. The new points are predicted to be included in the most classifications of these points. The best k value for this algorithm depends on the data. Generally, high k values reduce the effect of noisation on classification, but make the boundaries between each classification more blurred. A good k value can be chosen with parameter optimization, for example by using cross-validation. A special case where classification is predicted based on the closest learning data (in other words, $k = 1$) is called the nearest neighbor algorithm (Whidhiasih, Wahanani, & Supriyanto, 2013).

Evaluation of KNN Performance

Confusion matrix is used to evaluate the performance of an algorithm. Confusion matrix has information about the actual data and the results of the prediction of a classification into matrix form (Pulungan, Zarlis, & Suwilo, 2019).

III. PROPOSED METHOD

The data used in this study are 99 images with 75% of the dataset used for training, and 25% for testing. Dataset consisted of 75 images of pear images consisting of 25 training data of abate pear images, 25 training data of red pear images, and 25 training data of pear william images. The test data consisted of 8 images of pear abate, 8 images of pear red and 8 images of pear william (“Fruits 360 | Kaggle,” n.d.). Examples of images of these three types of fruit can be seen in Figure 1.

(a) (b) (c)

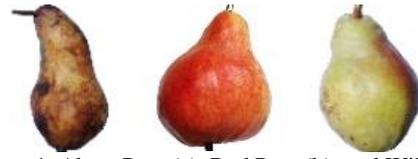


Figure 1. Abate Pear (a), Red Pear (b), and William Pear (c)

TABLE I. IMAGE DATASET OF PEAR

Class	Number of Images	Training Images	Testing Images
Abate	33	25	8
Red	33	25	8
William	33	25	8

The classification process of pear image types can be seen in Figure 2.

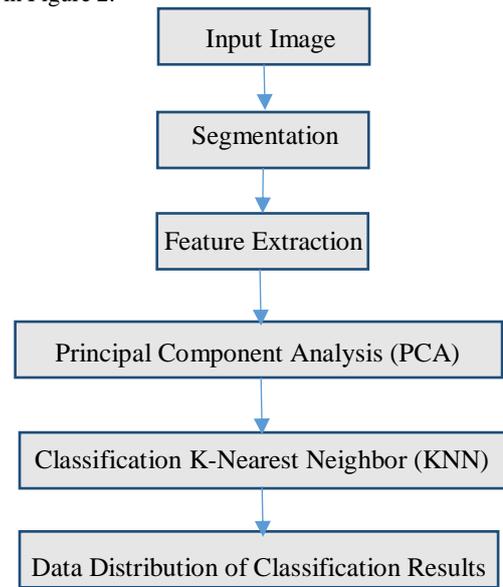


Figure 2. Classification Design of Pears

The picture shows that the method used for the classification of pear image types starts from the input of pear image then image segmentation is carried out to get the results of segmentation. Then the feature extraction process becomes red, green, blue, hue, saturation, and value. The feature extraction results that have been obtained are converted into principal components. The next step is to classify the KNN algorithm to determine the three types of pear images.

Input RGB Image

RGB color space is widely used and is usually the default color space for storing and representing digital images. We can get other color spaces from RGB or non-linear transformations. RGB color space is the color space used by computers, graphics cards and monitors or LCDs (Kolkur, Kalbande, Shimpi, ..., & 2017, n.d.). This process

aims to display the RGB (Red, Green, Blue) color space of the managed pear image.

RGB formula:

$$r = \frac{R}{R+G+B} \quad (7)$$

$$g = \frac{G}{R+G+B} \quad (8)$$

$$b = \frac{B}{R+G+B} \quad (9)$$

Image Segmentation

Image segmentation is part of the image processing process to divide an image into homogeneous regions based on certain similarity criteria between the gray level of a pixel and the gray level of its neighboring pixels, then the results of this segmentation process will be used for further processing. The Otsu method is a method for segmenting digital images using threshold values automatically, i.e. changing gray digital images to black and white based on comparison of threshold values with pixel color values of digital images. To get the threshold value there is a calculation that must be done. The first step that must be done is to make a histogram. From the histogram we can know the number of pixels for each gray level. The gray level of the image is expressed as i through L . The level i starts with 1, which is pixel 0. For L , the maximum level is 256 with pixels worth 255 (Syafi'i, Wahyuningrum, & Muntasa, 2016).

The threshold value to look for in a grayscale image is expressed as k . The value of k ranges from 0 to $L-1$, with a value of $L = 256$. So the probability of each pixel at level i is expressed by the equation :

$$P_i = \frac{n_i}{N} \quad (10)$$

The cumulative number formula of $\omega(k)$, for $L = 0, 1, 2, \dots, L-1$:

$$\omega(k) = \sum_{i=0}^k p_i \quad (11)$$

The cumulative average formula of $\mu(k)$, for $L = 0, 1, 2, \dots, L-1$:

$$\mu(k) = \sum_{i=0}^k i \cdot p_i \quad (12)$$

Formula for calculating the mean global intensity (k) μ_T :

$$\mu_T(k) = \sum_{i=0}^{L-1} i \cdot p_i \quad (13)$$

The equation for between class variance :

$$\sigma_B^2(k) = \frac{[\mu_T \omega(k) - \mu(k)]^2}{[\omega(k)[1 - \omega(k)]]} \quad (14)$$

The results of the calculation between the variance class look for the maximum value. The largest value is used as the threshold or the value of (k), with the equation

$$\sigma_B^2(k^*) = \max_{1 \leq x \leq L} \sigma_B^2(k) \quad (15)$$

Between class variance aims to find the threshold value of a grayscale image, the threshold value is used as a reference value to convert a grayscale image to a binary image. Each image has a different threshold value [6].

Hue Saturation Value (HSV)

Input images in the RGB color space are converted to HSV color space using transformations. HSV images are collections of three different images as hue, saturation, and value (Shaik, Ganesan, Kalist, ..., & 2015, n.d.). HSV has a closeness to the RGB system in describing colors that humans can see. HSV serves to reduce the intensity of light from outside and be able to detect certain objects. Here's the formula from RGB to HSV:

$$\begin{aligned} C_{\max} &= \max(R' \ G' \ B') \\ C_{\min} &= \min(R' \ G' \ B') \\ \Delta &= C_{\max} - C_{\min} \end{aligned} \quad (16)$$

C_{\max} functions to determine the largest constant value in the RGB value, while C_{\min} determines the smallest value in the RGB value.

Plotting Data Distribution

Plotting the data distribution is done with the aim of testing and viewing the image data distribution graph which is processed based on hue and saturation values to see the results of testing accuracy of pear type image processing. Plotting the data distribution that will be displayed is the distribution of training data in each class, the distribution of training data for each class along with the boundary lines and and the distribution of test data in each class.

IV. RESULT AND DISCUSSION

Research tool specifications

The tools used on this research is as follows :

1. Laptop with processor : AMD E-350 1.60GHz
2. Operating system : Windows 8 32-bit
3. Software Matlab R2013a used for making programs classification of determining the type of pears

The initial process of pear type classification is input of pear image from the training data in this study. Training data can be seen in Figure 3.

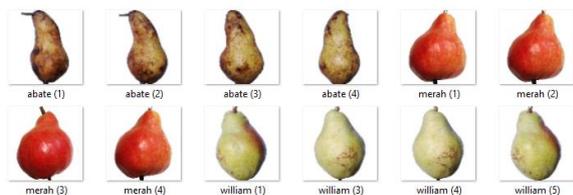


Figure 3. Image of Pear Train Data

Furthermore, the data that has been inputted is performed image segmentation using morphological operation methods to perfect the results of segmentation. Convert grayscale image that aims to determine the foreground area and background area with the value of a binary image, as shown in figure 4.

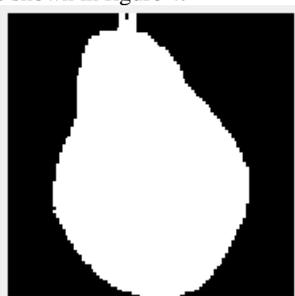


Figure 4. Binary Image

Then the transformation of the color space from the RGB image to the HSV image (Hue, Saturation, Value) is used as a reference to recognize the color of an object in a digital image and reduce the intensity of light from outside which can be seen in Figure 5.



Figure 5. Image of the Segmentation Results

Based on the image results of segmentation that has been obtained, then feature extraction is performed to obtain the value of RGB, hue, saturation, value, and area of the pear image. After that the reduction is done using the PCA algorithm to get the results of the classification of images of abate, red and william pears.

The test results can be seen from plotting graphs of processed image data. Plotting the distribution of training data in each class is shown in Figure 6.

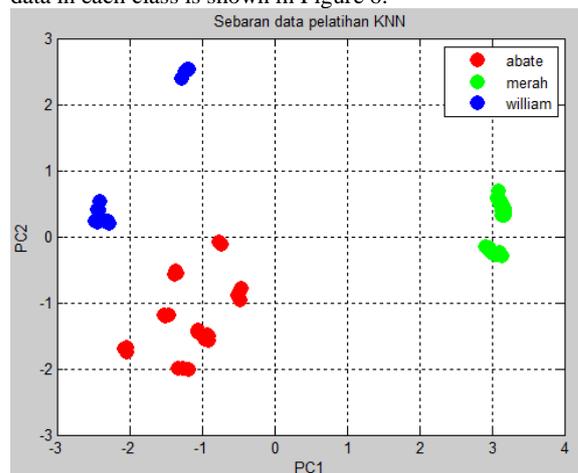


Figure 6. Distribution of Training Data

Based on the training data that has been obtained, testing is done using test data. The following is a display of the distribution of training data and test data based on the boundary line using the PCA and KNN algorithms seen in Figure 7.

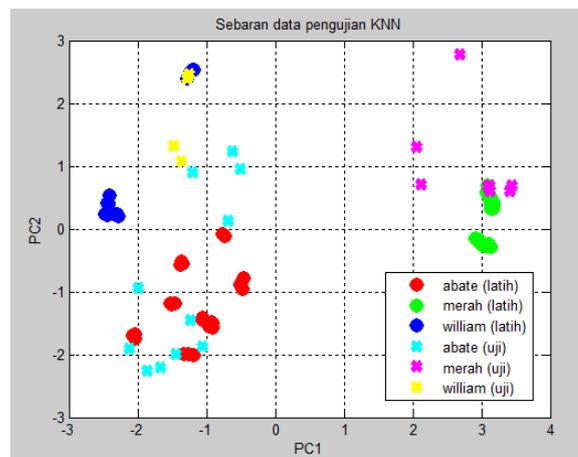


Figure 7. Distribution of Test Data and Training Data

Furthermore, to get the accuracy of the classification of pears, testing is done using the GUI application using the Matlab application. The designed application consists of several functions, namely image input, image segmentation process, feature extraction, and the process of determining the results of classification. Following is the appearance of the Matlab GUI Application that has been made:

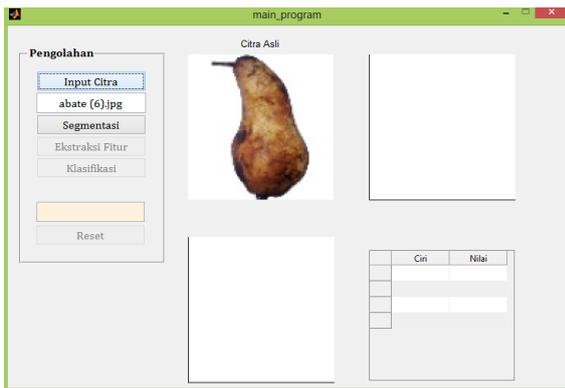


Figure 8. Display of Image Input Matlab GUI Application

Following is the appearance of the Matlab GUI Application segmented images, which consist of binary images and segmented images:

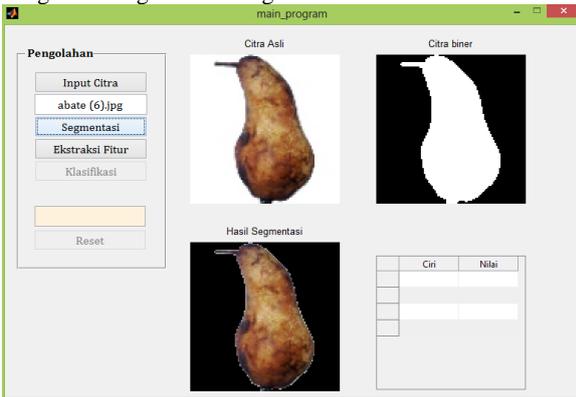


Figure 9. Display of Matlab GUI Application Image segmentation results

The following is the display of the extracted image features consisting of Red, Green, Blue, Hue, Saturation, Value, and Area features in the Matlab GUI Application:

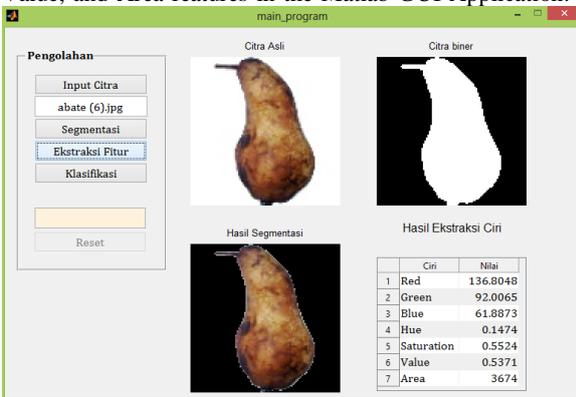


Figure 10. Display of Matlab GUI Application Extraction Results

Based on the feature extraction results that have been obtained, a process is carried out to determine the

classification using the PCA and KNN algorithms. The following is the appearance of the Matlab GUI classification results of the three types of pears:

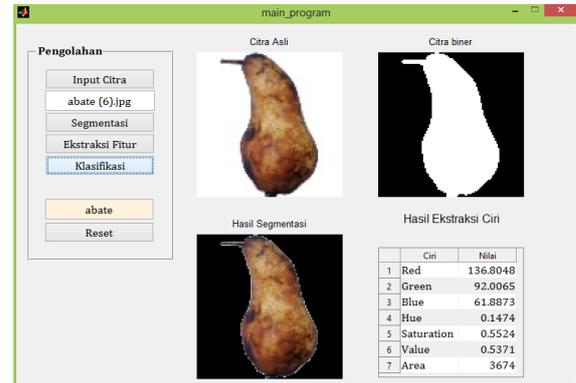


Figure 11. Display of Matlab GUI Application Results of Abate Pear Classification Results

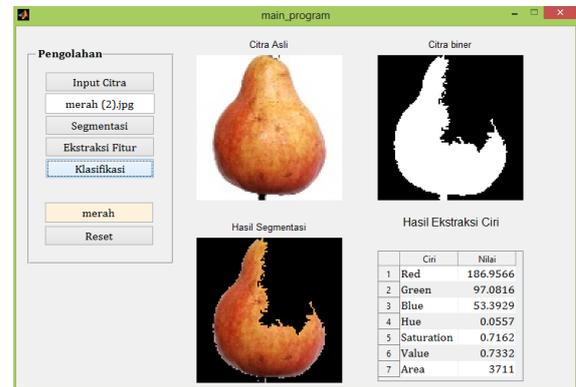


Figure 12. Display of Matlab GUI Application Results of Classification of Red Pears

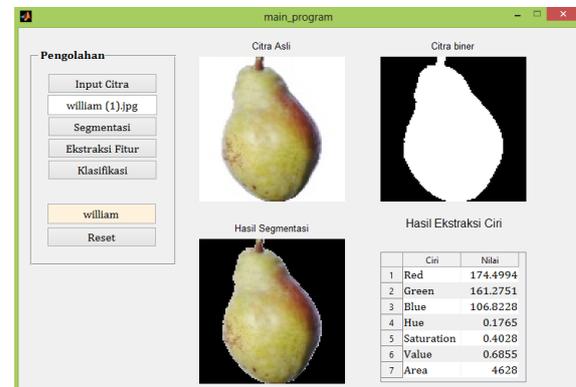


Figure 13. Display of Matlab GUI Application Results of William Pear Classification Results

The test results carried out on a test data of 24 pears using the Matlab GUI application can be seen in table 1.

TABLE 2. PEARS IMAGE PROCESSING RESULTS

No.	Test Image	Original Class	Output Class	Result
1		Abate	Abate	True
2		Abate	Abate	True
3		Abate	Abate	True
4		Abate	Abate	True
5		Abate	Abate	True
6		Abate	Abate	True
7		Abate	Abate	True
8		Abate	Abate	True
9		Red	Red	True
10		Red	Red	True
11		Red	Red	True
12		Red	Red	True
13		Red	Red	True
14		Red	Red	True
15		Red	Red	True
16		Red	Red	True
17		William	William	True

No.	Test Image	Original Class	Output Class	Result
18		William	Abate	False
19		William	Abate	False
20		William	Abate	False
21		William	William	True
22		William	William	True
23		William	William	True
24		William	William	True

Based on the table above, 21 pears were successfully classified according to their type, but there were 3 William Pears classified as Pear Abate. This can be seen in the GUI display Figure 14

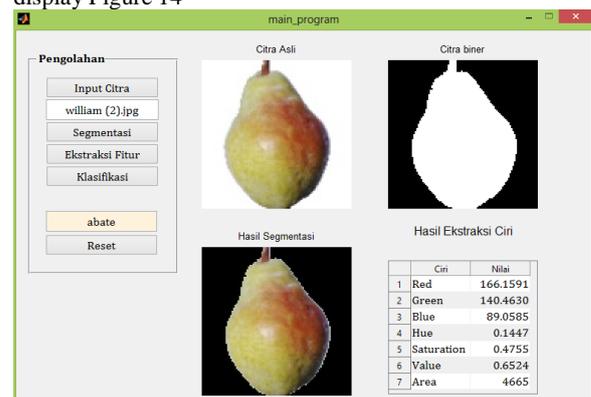


Figure 14. Display of the Matlab GUI Application Results of William Pear Classification Classified into Abate Pear

Image data used for testing are 24 consisting of 8 images of pear abate, 8 images of red pear, and 8 images of pear william. The test results show that the number of pears that were successfully classified according to their class amounted to 21 images while 3 images did not match. Classification results will be presented in the form of confusion matrix. This table consists of predict classes and actual classes. The 3x3 confusion matrix model is shown in Table 3 while the accuracy value of the model is obtained

from equation (17), the exact amount of data classified divided by the total data. Based on the calculation results obtained an accuracy value of 87.5%. The resulting accuracy indicates that the PCA and KNN algorithms are very well applied in the classification of fruit types.

TABLE 3. CONFUSION MATRIX
ACCURACY = 87.5%

Actual Class	Predict Class			Total
	Abate (A)	Red (B)	William (C)	
Abate (A)	8	0	0	8
Red (B)	0	8	0	8
William (C)	3	0	5	8
	11	8	5	24

$$\text{Akurasi} = \frac{AA+BB+CC}{AA+AB+AC+BA+BB+BC+CA+CB+CC} \quad (17)$$

V. CONCLUSION AND SUGGESTION

This study found the results of how ordinary people can easily determine the type of pear only from an image processing. From the results of the classification process of image processing of abate, red and william pears, an accuracy of 87.5% was obtained. Using the Principal Component Analysis and K-Nearest Neighbor algorithm is very suitable in the classification process of pears. Image quality is very influential on the results of classification as well as the amount of training data used to obtain classification results. The more training data used, the better the accuracy of the classification of pears. It is recommended to develop further research using more than three types of pears.

VI. REFERENCES

- Ariffin, N. I. K., Mustaffa, M. R., Abdullah, L. N., & Nasharuddin, N. A. (2018). Fruits Recognition based on Texture Features and K-Nearest Neighbor. *International Journal of Engineering & Technology*, 7(4.31), 452–458.
- Fruits 360 | Kaggle. (n.d.). Retrieved August 22, 2019, from <https://www.kaggle.com/moltean/fruits>
- Herfina. (2013). Pengenalan Pola Bentuk Bunga Menggunakan Principle Component Analysis. *Seminar Nasional Teknologi Informasi Dan Multimedia*, (7), 25–30.
- Husein, A. M. (2019). Penerapan Metode Distance Transform Pada Kernel Discriminant Analysis Untuk Pengenalan Pola Tulisan Tangan Angka Berbasis Principal Component Analysis . *Sinkron*, 2(Sinkron), 31–36.
- Isola, P., Zhu, J., ... T. Z.-P. of the I., & 2017, undefined. (n.d.). Image-to-image translation with conditional adversarial networks. *Openaccess.Thecvf.Com*.
- Kolkur, S., Kalbande, D., Shimpi, P., ... C. B. preprint arXiv, & 2017, undefined. (n.d.). Human skin detection using RGB, HSV and YCbCr color models. *Arxiv.Org*.
- Liantoni, F. (2016). Klasifikasi Daun Dengan Perbaikan Fitur Citra Menggunakan Metode K-Nearest Neighbor. *Jurnal ULTIMATICS*, 7(2), 98–104. <https://doi.org/10.31937/ti.v7i2.356>
- Muhammad, J., & Isnanto Riza, S. I. (n.d.). Identifikasi iris mata menggunakan metode analisis komponen utama dan perhitungan jarak euclidean, 1–9.
- Octavia, M., Jesslyn, K., & Gasim. (2016). Perbandingan Tingkat Akurasi Jenis Citra Keabuan , HSV, Dan L*a*b* Pada Identifikasi Jenis Buah Pir. *Ilmiah Informatika Global*, 7(1), 7–11.
- Pulungan, A. F., Zarlis, M., & Suwilo, S. (2019). Analysis of Braycurtis, Canberra and Euclidean Distance in KNN Algorithm. *Sinkron*, 4(1), 74. <https://doi.org/10.33395/sinkron.v4i1.10207>
- Sehgal, P., & Goel, N. (2016). Auto-annotation of tomato images based on ripeness and firmness classification for multimodal retrieval. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1084–1091). IEEE. <https://doi.org/10.1109/ICACCI.2016.7732189>
- Shaik, K., Ganesan, P., Kalist, V., ... B. S.-P. C., & 2015, undefined. (n.d.). Comparative study of skin color detection and segmentation in HSV and YCbCr color space. *Elsevier*.
- Subasi, A., & Gursoy, M. I. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, 37(12), 8659–8666. <https://doi.org/10.1016/j.eswa.2010.06.065>
- Syafi'i, S. I., Wahyuningrum, R. T., & Muntasa, A. (2016). Segmentasi Obyek Pada Citra Digital Menggunakan Metode Otsu Thresholding. *Jurnal Informatika*, 13(1), 1–8. <https://doi.org/10.9744/informatika.13.1.1-8>
- Whidhiasih, R. N., Wahanani, N. A., & Supriyanto, S. (2013). Klasifikasi Buah Belimbing Berdasarkan Citra Red-Green-Blue Menggunakan Knn Dan Lda. *Penelitian Ilmu Komputer Sistem Embedded Dan Logic*, 1(1), 29–35.

