Emerging Information Science and Technology
Vol. 1, No. 1, (2020), pp. 35-43
DOI: 10.18196/eist.115

Classification of Mangosteen Surface Quality Using Principal Component Analysis

Slamet Riyadi[1], Amelia Mutiara Ayu Pratiwi[2], Cahya Damarjati[3], Tony K. Hariadi[4], Indira Prabasari[5], Nafi Ananda Utama[6]

[1,2,3,4,5,6]Universitas Muhammadiyah Yogyakarta, Jln.Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta 55183, Indonesia
[1]riyadi@umy.ac.id, [2]amelia.mutiara.2013@ft.umy.ac.id, [3]cahya.damarjati@umy.ac.id, [4]tonykhariadi@umy.ac.id, [5]i.prabasari@umy.ac.id
[6]nafi@umy.ac.id

Abstract

Mangosteen (Garcinia mangostana L) is one of the primary contributor for Indonesia export. For export commodity, the fruit should comply the quality requirement including its surface. Presently, the surface is evaluated by human visual to classify between defect and non- defect surface. This conventional method is less accurate and takes time, especially in high volume harvest. In order to overcome this problem, this research proposed images processing based classification method using principal component analysis (PCA). The method involved pre-processing task, PCA decomposition, and statistical features extraction and classification task using linear discriminant analysis. The method has been tested on 120 images by applying 4-fold cross validation method and achieve classification accuracy of 96.67%, 90.00%, 90.00% and 100.00% for fold-1, fold-2, fold-3 and fold-4, respectively. In conclusion, the proposed method succeeded to classify between defect and non-defect mangosteen surface with 94.16% accuracy.

Keywords: *Mangosteen, principal component analysis, Statistical features extraction, Linear discriminant analysis, K-Fold cross validation*

1. Introduction

Mangosteen (Garcinia mangostana L) becomes a biggest contributor for fruit export in Indonesia. As an export commodity, the requirements quality of mangosteen should be maintained so that the exported fruit can be accepted by the consumer. There are a lot of problem faced by the farmers such as mangosteen is a tropical humid plant which requires years to produce fruit, the relatively long juvenile period to produce fruit. The other problem in encountering the competition with mangosteen export production from the other countries is not only the quality of the mangosteen but also the marketing due to the distance of the export destination from Indonesia. Fruit quality will be reduced if the marketing process takes a long time. Because in every marketing area whether it is domestic or international has different quality demand. Therefore, mangosteen which has high quality is needed in order to export and compete with the other countries.

The familiar quality investigation method of mangosteen nowadays is conventional method which is manual investigation using direct visual to the classified fruit. This method is less effective because it takes a relatively long time for the sorter to select the mangosteen which resulting different perception toward the quality of the mangosteen, also the cost is considerably high.

Based on those problems, a study was required in order to find a new method which is more reliable and have higher accuracy and affectivity in order to replace the conventional method. Many researches about fruit evaluation and it founds different result. Principal Component Analysis for fruit quality evaluation of cooled banana 'nanicao' was studied by Sanches et al.

[1], this work purpose is to evaluate the quality of banana based on two conditions, storage temperature and three different packages. This research was using multivariate analysis in order to achieve the goal, and the result prove that PCA is necessary to decrease mechanical injuries. Wang et al. [2] tried to improve the sensor quality of Yai perry and some other fruits, this research combined between sensory evaluation and PCA method. The result of this combination is hawthorn-Yali perry and plum-Yali perry got the highest accuracy. Effects of irrigation and fertilization on tomato fruit was studied by Wang and Xing [3], the experiment was carried in a solar greenhouse spanning three consecutive growing season using Water Use Efficiency (WUE) and Fertilizer Partial Factor Productivity (PFP) parameter. The proper application of drip fertigation may be a good compromise for solar greenhouse-grown tomatoes. An evaluation of apricot fruit quality and correlation between physical and chemical attributes was studied by Mratinic et al. [4]. This research used the fruit of apricot (Prunus armeniaca L.., Rosaceae) as object. The chemical organic matters from fruit has potential nutritional, medical, and commercial values. The results based on these attributes showed that analysis of different apricot genotypes were separated into groups with similar physical and chemical attributes. Seven kind of grapes were used as materials in order to evaluate fruit quality of wine grape using Principal Component Analysis. This research was performed by Chunyan et al. [5] using several indicators including single grain weight, vertical diameter, transverse diameter, berry size, fruit shape index, soluble solid, total sugar, reducing sugar, titratable acid, sugar acid ration, tannin, total phenolic. These attributes were analysed using by difference analysis, correlation analysis, and component analysis. The highest accuracy was showed by PCA method, it reached 90.7% compared to other methods. Study about Principal Component Analysis also performed by Bo et al. [6] in Principal Component Analysis and Comprehensive Evaluation of the Fruit Quality of Lycium Barbarium L. from Different Regions. Three kind of lyceum barbarium from three differences provinces were used as the materials. This research aim is to analyse the external quality such as kernel mass and fruit shape index, and intrinsic factor such as carotene, betaine, and polysaccharide. The results showed that compositions were extracted by PCA and their accumulative variance contributions was 91.71%. Evaluation of apple quality based on Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) was studied by Liyan et al. [7]. The purpose of this study was to investigate the variations in physical and chemical characteristic of apple fruits. This study used twenty quality parameters of apple such as weight, volume, density, colour, hardness, sugar-acid ratio etc. The results showed that HCA classified 30 varieties into five main class of the measured parameters, which was consistent with the results of PCA.

In this research, a method which was used to detect the mangosteen surface defect is based-on mangosteen surface images using principal component analysis (PCA). In this study, texture measurement based on score value, latent, and coeff with the characteristic extraction such as mean, energy, standard deviation, and variance were used to characterize the fruit surface texture which later used in the classification used to differentiate defect and non-defect mangosteen surface using Linear Discriminant Analysis (LDA) classification method.

2. Method

The detection of surface defect engages the procedure as shown in the Figure 1. First, the images of mangosteen surface were resized and converted to grayscale. Afterward, we processed the image using PCA decomposition and extract the features using statistical extraction. Lastly, linear discriminant analysis was applied to classify it as “defect” or “no defect” image.

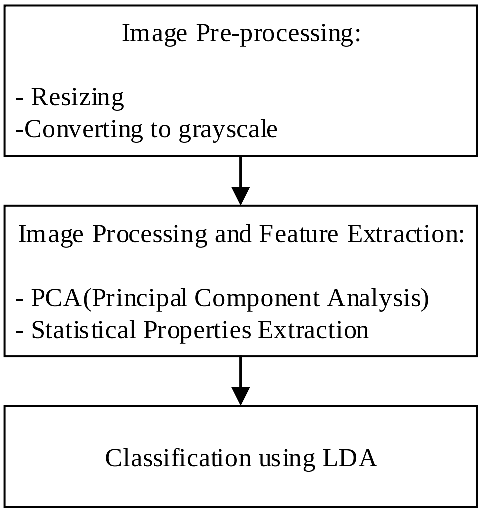


Image Pre-processing:
- Resizing
-Converting to grayscale
Image Processing and Feature Extraction:
- PCA(Principal Component Analysis)
- Statistical Properties Extraction
Classification using LDA

Figure 1. Classification Process

2.1. Image Pre-processing

Image process in the pre-processing stage was aimed to concentrate input image before it was processed using PCA method. First, the size of RGB image was changed to 512x512 dimensions. Second, the image was changed into grayscale mode to simplify the images and decrease the computation time.

2.2. Principal Component Analysis

PCA is a technique which is used to analyze an observation data table into a new table which has the same correlation. The purpose of conducting PCA was to simplify complex observation data in order to ease the data processing and analyzing. According to the researcher, PCA is a statistical technique which linearly transforms a set of original variable into smaller or simpler variable which is uncorrelated and can represent the information from the set of original variable [8].

PCA can also be used to reduce the dimension of a data without significantly reducing the characteristic of those data [9]. Frequently, some Principal Component (PC) is enough to explain the structure of an original data. In case the data in the original dimension is hard to present using graphic, with two PC or one PC the data can be imaged using graphic.

The purpose of PCA is to explain the part of variation in a set of observed variable on the basis of several dimensions, from various variable changes into fewer variables. PCA itself used to:

1. Simplify correlation pattern between observed variables.
2. Reduce a set of variable into smaller factor.
3. Give operational definition along with key dimension regression of the used of observed variable.
4. Test the underlying theories [10].

The value contained in PCA is the data analysis component which comes from the data input which has been sorted from the first line to the last line, where the first line contains the most important information of the data (principle component), then the second information is in the second line. It goes until the last line, latent is a certain value of a variable which is used as an alternative to save the component analysis, whereas coeff is a constants value in the processed image data, coeff show the differentiation of values from each matrix pixels.

2.3. Statistical Properties Extraction

Feature extraction is the process which increase unique characteristic of a thing in the form of value which is going to use for analysis. Feature extraction used in this study was mean, energy, standard deviation, and variance. Each feature extraction was calculated from matrix PCA.

Statistic property which was calculated was; mean was calculated using equation (1), energy was calculated using equation (2), standard deviation was calculated using equation (3), and variance was calculated using equation (4)

$$\mu = \frac{1}{N}\sum_{t=1}^{N} A_i \tag{1}$$

$$E_k = \frac{1}{MxN}\sum_{i=1}^{M} \sum_{i=1}^{M} |x_k(i,j)| \tag{2}$$

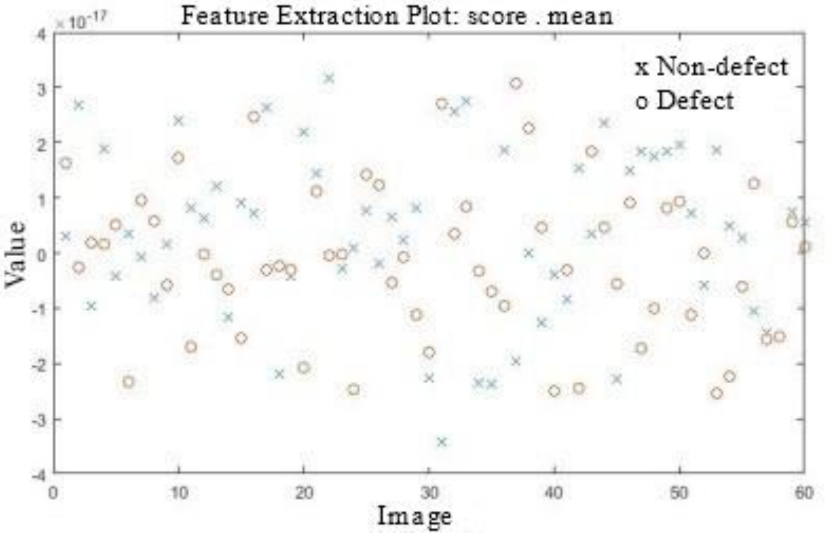
$$S = \sqrt{\frac{1}{N-1}\sum_{t=i}^{n} |A_i - \mu|^2} \tag{3}$$

$$y = \sigma^2 = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}|u_{ij}|^2 - \frac{|\sum_{i=1}^{M}\sum_{j=1}^{N}u_{ij}|^2}{M*N}}{M*N-1} \tag{4}$$

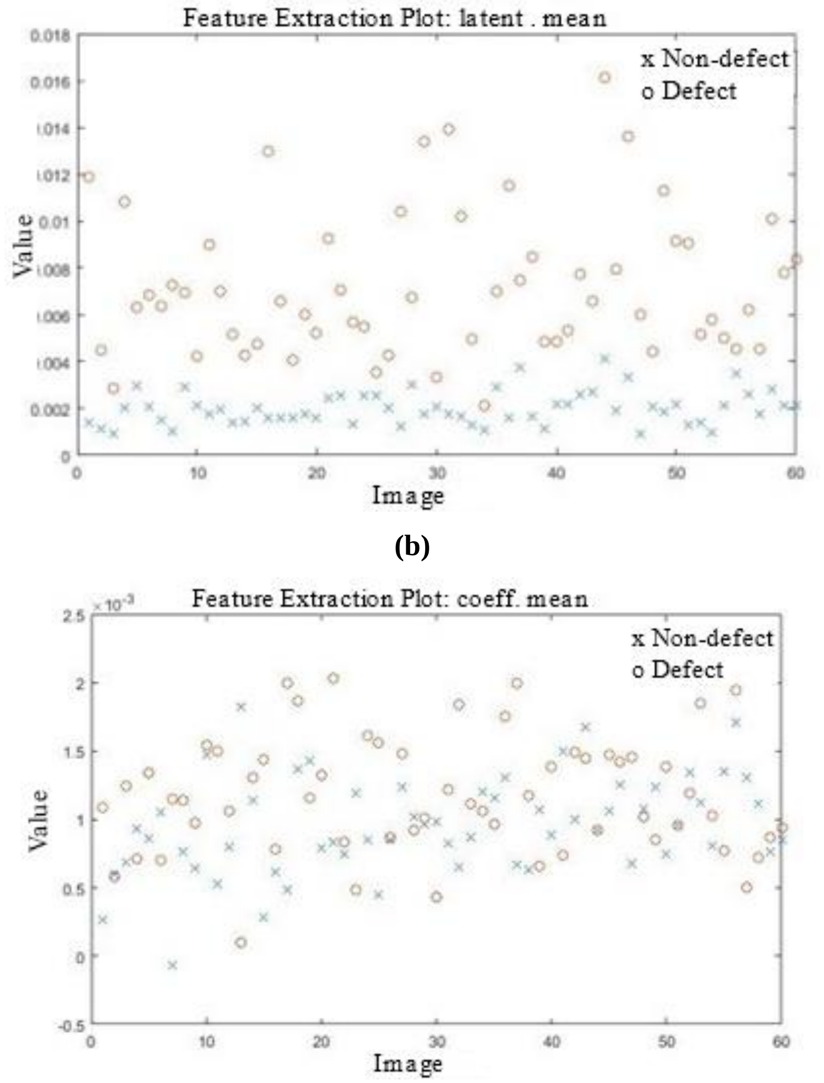
2.4. Linier Discriminant Analysis

Linear discriminant analysis is a scheme which is famous for feature extraction and dimension reduction. Discriminant analysis is dependencies statistical analysis technique which is used to classify several groups of object. The classification was divided into two classes, “defect” and “no defect”. This LDA classification method uses score, latent, coeff and input which were computed by PCA. Those derived from characteristic extraction value, where the input value used is mean, energy, standard deviation, and variance. The scatter plot result of the tests are shown in the Figure 2, Figure 3, Figure 4, and Figure 5.

2.4.1. Mean



(a)

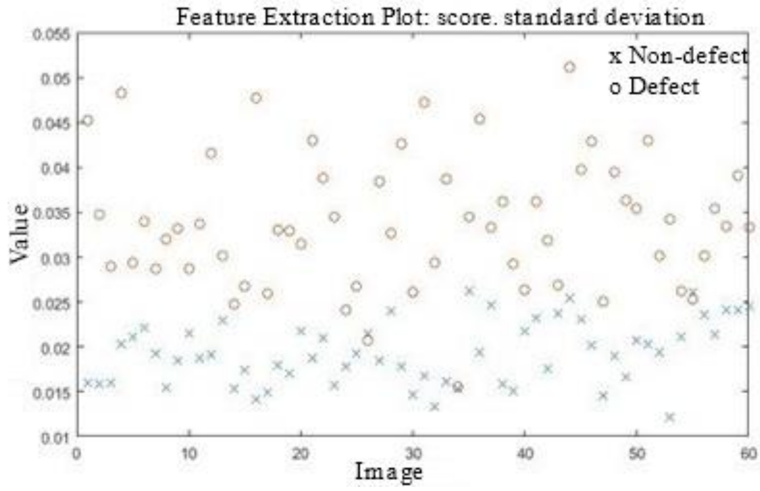


(b)

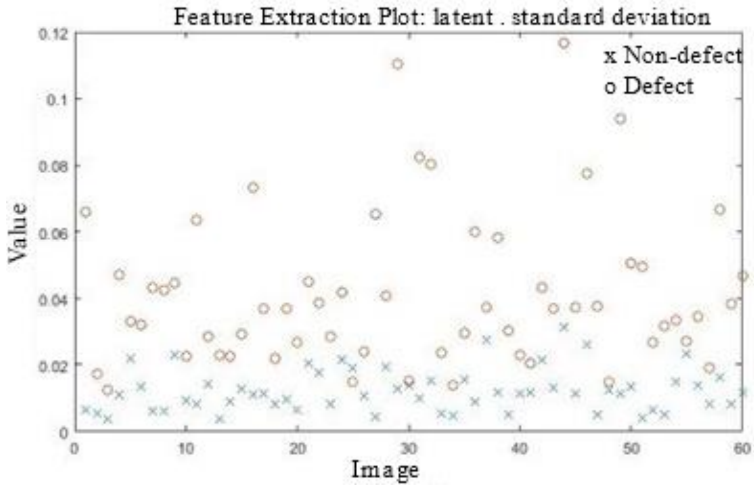
(c)

Figure 2. Feature Extraction (a) Score-Mean (b) Latent-Mean (c) Coeff-Mean

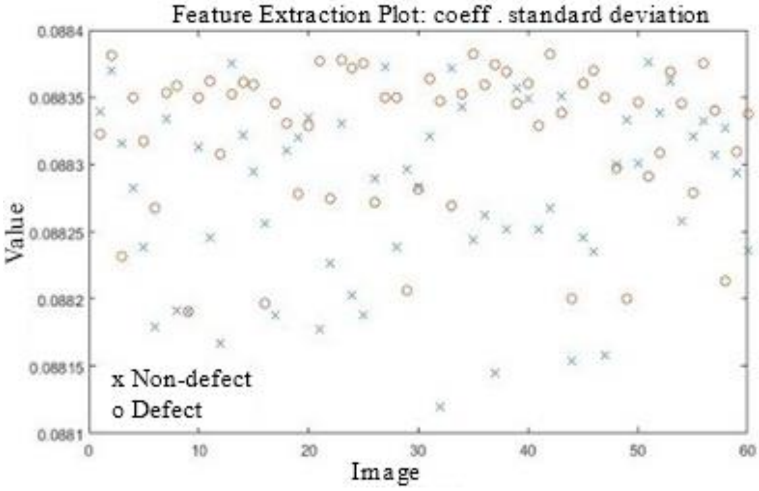
2.4.2. Standard Deviation



(a)



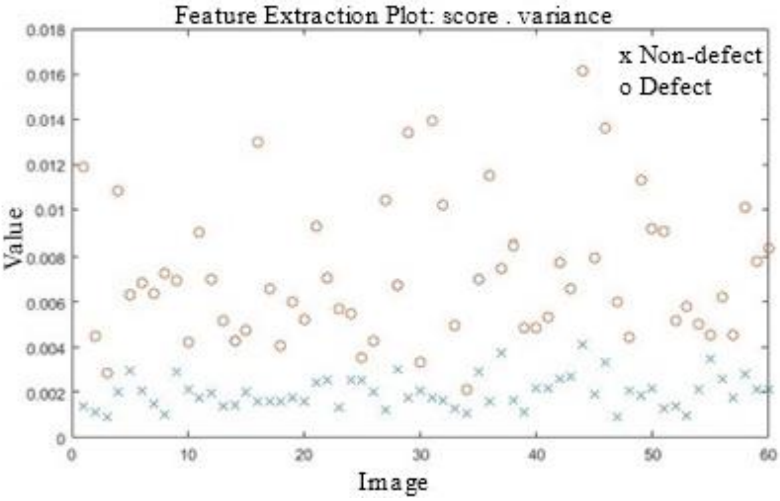
(b)



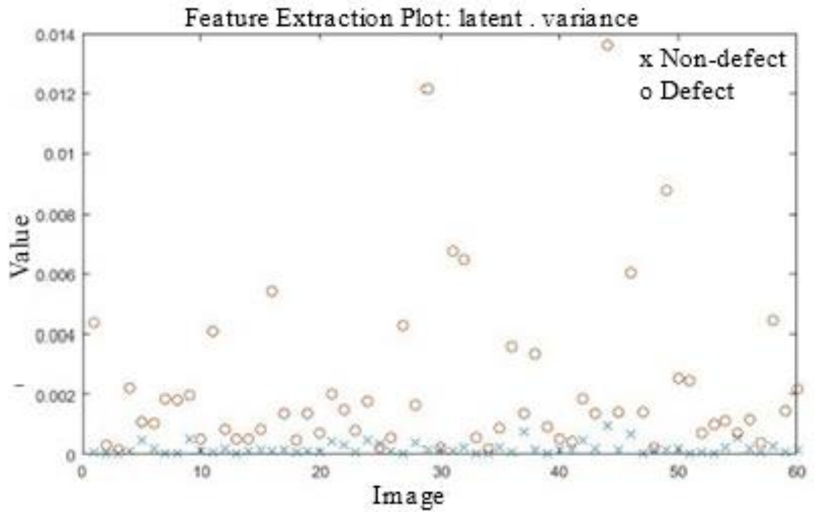
(c)

Figure 3. Feature Extraction (a) Score-Std (b) Latent-Std (c) Coeff-Std

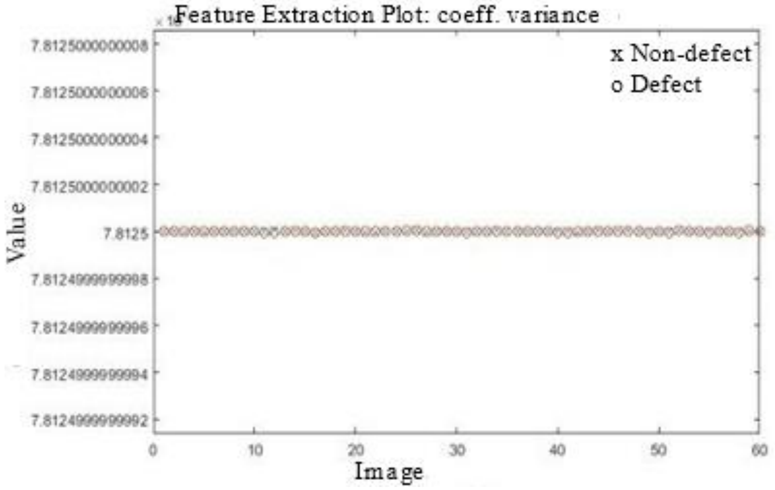
2.4.3. Variance



(a)



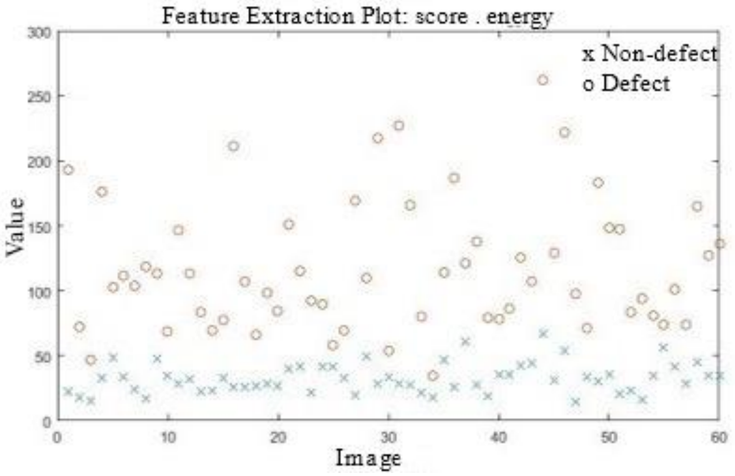
(b)



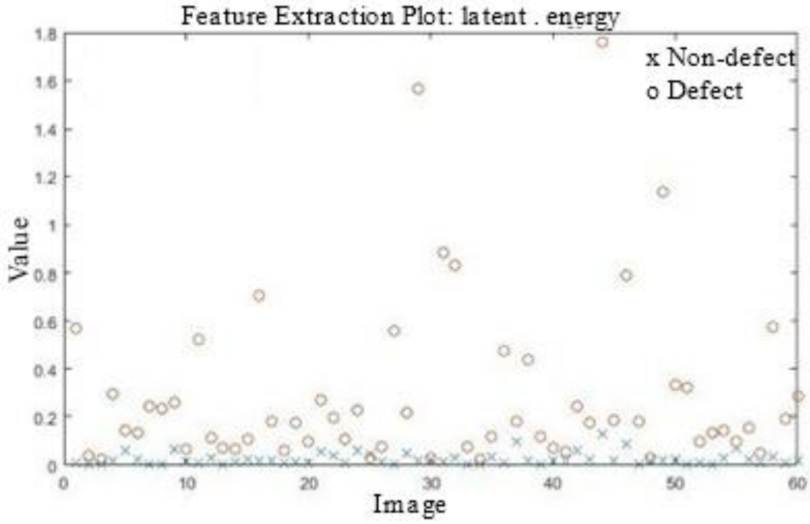
(c)

Figure 4. Feature Extraction (a) Score-Variance (b) Latent-Variance (c) Coeff-Variance

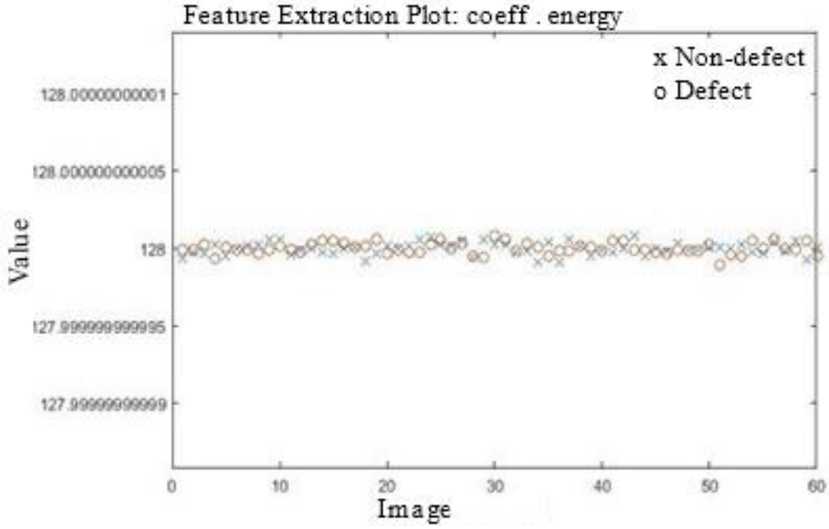
2.4.4. Energy



(a)



(b)



(c)

Figure 5. Feature Extraction (a) Score-Energy (b) Latent-Energy (c) Coeff-Energy

3. Result

There were 120 images used in this study, the image consisted of defect image and no defect image which were used for training and each group consisted of 30 images which were used for testing. All of the images in the database were taken using the same set of data acquisition. During experiment we used four features such as mean, standard deviation,variance, and energy which were combined using three statistical measurement, there were score, latent, and coeff.

Non-defect images based on Mean-latent combination has lower value than defect images and other statistical value mean-score and mean-coeff have spreaded out pattern. Next combination is between std-score and std-latent, this mixture shows that non-defect images have lower value than defect images. Other statistical combination between variance-energy and variance-latent show that non-defect images have lower value than defect images. The last combination between energy-score and energy-latent also show that non-defect images have lower value compare to defect images. These features could be properly used as the input value for Principal Component Analysis.

3.1. The Result of Classification and Validation in the Fold-1

From the Table 1, the result of accuracy test of the image sample in the fold-1 was obtained. Error was occur in the image testing when there is an image, def11.jpg as shown in Figure 6 which is classified as defect image, in the other hand the test result using the program was classified as no defect image. In the manual classification, that image was classified as defect image because the image condition was unclear so it caused errors in the classification.

Table 1. Accuracy Result in the Fold-1

Result	Images			Total Error	Accuracy
	Defect	Non-defect	Total		
Fold-1	15	15	30	1	96.67%

Figure 6. def1.jpg

3.2. The Result of Classification and Validation in the Fold-2

From the data in the Table 2, the result of accuracy test was obtained from image test in the fold-2. There were three images which error was found, they were def25.jpg, def26.jpg, and def30.jpg shown in Figure 7 which was classified as defect image, meanwhile, the result obtained from testing using the program was classified as no defect image. In the manual classification, the images was classify as defect images because the image was not clear which resulting the error in the classification.

Table 2. Accuracy Result in the Fold-2

Result	Images			Total Error	Accuracy
	Defect	Non-defect	Total		
Fold-2	15	15	30	3	90.00%

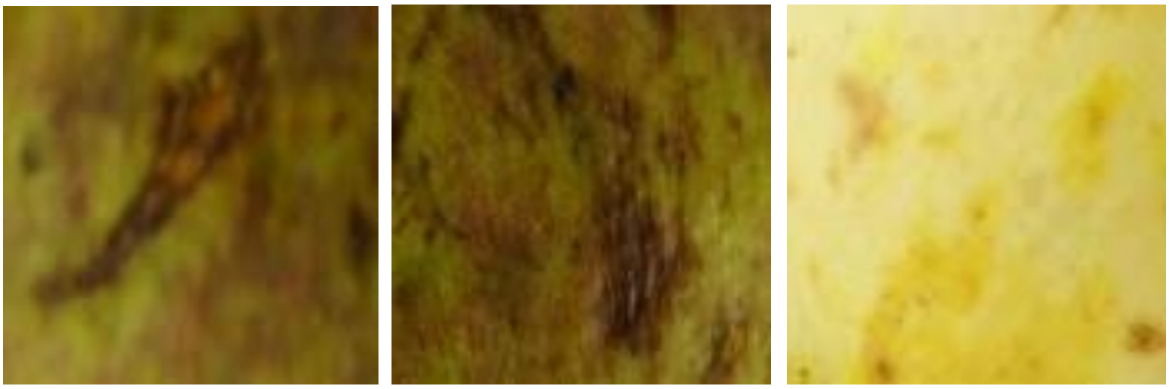


Figure 7. def25.jpg, def26.jpg, and def3.jpg

3.3. The Result of Classification and Validation in the Fold-3

The result of accuracy test of image test in the fold-3 was obtained from the Table 3. The error that occur in the image test of the three image def34.jpg, def37.jpg and fin44.jpg shown in Figure 8 which were classified as defect image (def34.jpg and def37.jpg) and no defect (fin44.jpg), in the other hand the result of the test using the program was classified as no defect image (def34.jpg and def37.jpg) and defect (fin44.jpg). In the manual classification, those images was included in defect image because the image condition of images def34.jpg and def37.jpg was not clear. Whereas the image fin44.jpg was in the good image condition under the manual classification, however the presence of brown spots and small holes made the image classified as defect image. This problem caused classification error.

Table 3. Accuracy Result in the Fold-3

Result	Images			Total Error	Accuracy
	Defect	Non-defect	Total		
Fold-3	15	15	30	0	100.00%

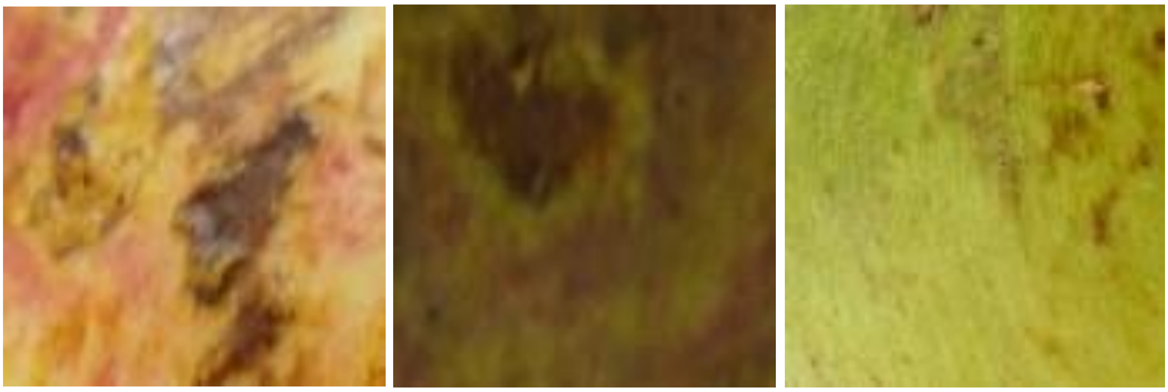


Figure 8. def34.jpg, def37.jpg, and fin37.jpg

3.4. The Result of Classification and Validation in the Fold-4

From the Table 4, accuracy test from image test in fold-4 was obtained. There were no error found in the image test therefore it resulted perfect accuracy of 100% in fold-4.

Table 4. Accuracy Result in the Fold-4

Result	Images			Total Error	Accuracy
	Defect	Non-defect	Total		
Fold-4	15	15	30	0	100.00%

Total accuracy of 120 data was obtained by averaging four accuracies of every fold which is summarized in Table 5.

Table 5. Total Accuracy Percentage Testing Result 4-Fold

Accuracy				Average
Fold-1	Fold-2	Fold-3	Fold-4	
96.67%	90%	90%	100%	94.16%

4. Conclusion

Based on the result of this study, it can be concluded that the Principal Component Analysis had successfully used to classify between defect and non-defect of mangosteen surface. From

PCA values in the form of score and latent using the standard deviation and variation extraction methods for the whole 4- fold, it was obtained accuracy of 94.16%.

Acknowledgement

Authors would like to acknowledge Ministry of Research, Technology and Higher Education, Republic of Indonesia for the grant Penelitian Unggulan Perguruan Tinggi 2017.

References

[1] J. Sanches, P. A. M. Leal, J. H. Saravali, and S. Antoniali, “Principal components analysis for quality evaluation of cooled banana ‘Nanicao’ in different packing,” *Rev. Bras. Frutic, Jabotical*, vol. 25, no. 2, pp. 220–223, 2003.

[2] W. Yanhui, Y. Liu, Y. Zhang, and Z. Xu, “Quality evaluation of mixed brewed perries based on PCA and sensory evaluation,” *Front. Agric. China*, vol. 5, no. 4, pp. 529–533, 2011.

[3] X. Wang and Y. Xing, “Evaluation of the effects of irrigation and fertilization on tomato fruit yield and quality: a principal component analysis,” *Scientific Reports*, vol. 7, no. 1, pp. 1-13, 2017.

[4] E. Mratinic, B. Popovski, T. Milosevic, and M. Popovska, “Evaluation of Apricot Fruit Quality and Correlations Between Physical and Chemical Attributes,” *Czech J. Food Sci.*, vol. 29, no. 2, pp. 161–170, 2011.

[5] L. Chunyan, X. Yue, L. Dongmei, Z. Jing, and W. Zhenping, “Evaluation on Fruit Quality of Wine Grape Based on Principal Component Analysis,” *Northern Horticulture*, vol. 11, no. 3, 2017.

[6] Z. Bo, D. Guoli, and H. Ting, “Principal Component Analysis and Comprehensive Evaluation of the Fruit Quality of Lycium barbarum L. from Different Regions,” *Acta Agriculturae Boreali-Occidentalis Sin.*, vol. 23, no. 8, pp. 155-159, 2014.

[7] G. Liyan, M. Xianjun, L. Naiqiao, and B. Jinfeng, “Evaluation of apple quality based on principal component and hierarchical cluster analysis,” *Trans. Chinese Soc. Agric. Eng.*, vol. 30, no. 13, pp. 276-285, 2014.

[8] G. H. Dunteman, Principal Component Analysis, Newbury Park London: Sage Publications, 1989.

[9] L. I. Smith, A Tutorial on Principal Component Analysis, John Wiley & Sons Inc., 2002.

[10] B. G. Tabachnick and L.S. Fidell, Using Multivariate Statistics 7th ed., Boston, MA: Pearson, 2007.