Analysis of Cantaloupe Fruit Maturity Based on Fruit Skin Color Using Naive Bayes Classifier

M A Bustomi* and M F Asy’ari
Department of Physics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia
*Email: a_bustomi@physics.its.ac.id

Abstract. The traditional sorting of fruit maturity can be done by seeing the color of the fruit's skin. Manual sorting will take a long time and the results are subjective. This paper presents the results of maturing cantaloupe fruit based on the color of the fruit skin using a digital image of the fruit skin. The research objective is to classify the maturity of cantaloupe fruit using the Naive Bayes Classifier method and compare the results with similar studies using the Learning Vector Quantization (LVQ) Artificial Neural Network method. This study used the image of a raw and mature cantaloupe rind of 15 images each. A total of 16 images are grouped into training data for the training process and 14 other images are grouped into test data for the testing process. The results showed that the accuracy of training and testing using the Naive Bayes Classifier method was 68.75% and 57.14%, respectively. The accuracy of the training and testing of the Naive Bayes Classifier method turns out to be lower compared to the LVQ Artificial Neural Network method.

1. Introduction
Cantaloupe is a fruit that is quite popular with people in Indonesia. This fruit contains lots of fiber, vitamin A, vitamin C, vitamin B6, iron, and others. Knowing the level of fruit maturity is important for different purposes such as priority distribution to market and storage based on the level of maturity [1]. Traditionally, the level of maturity of vegetables and fruit is classified by color and physical characteristics by manual sorting. However, manual sorting of vegetables and fruit is a time-consuming procedure and depends on the person performing it. The skill of sorting vegetables and fruit varies from person to person. Therefore the results are not necessarily accurate because of everyone's subjectivity [2]. Color is one of the important characteristics in determining the level of maturity and quality of vegetables and fruits. The color feature has been widely applied to evaluate the level of fruit maturity. Color features composed of basic colors red, green, and blue (RGB) can be used to classify the level of maturity in tomatoes [3], passion fruit [4], and lemon [5].

Efforts to identify, classify, and sort fruit ripeness levels automatically are based on digital images. Both on static images (photos) and on moving images (videos) [6-10]. Various choices of methods for automatic classification of fruit maturity levels have also been studied and carried out by many people, for example using machine vision [11] and artificial neural network methods [12, 13]. The machine vision method is developed based on the ability of computer hardware and software to detect, process, and recognize an image (vision). The neural network method is developed based on the ability of computer software to identify and recognize an object mimicking the workings of human neural networks. In previous studies, we have used the Learning Vector Quantization (LVQ) Artificial Neural Network method to classify the maturity level of cantaloupe fruit based on fruit skin color features [14].
Another method that can be used to classify objects is the Naive Bayes Classifier (NBC) method. Unlike the neural network method, the NBC method uses a statistical approach in the object classification process. This method is developed based on the Bayes theorem [15]. The NBC method has yielded reasonably good results in the classification of X-ray images of the lungs into lung disease classes [16]. In agriculture, the NBC method has also been applied to the selection of apple fruit varieties [17].

This paper will discuss the results of research on the application of the NBC method to classify the ripeness of cantaloupe based on the texture of the fruit skin color. This study uses image objects that are the same as image objects in previous studies. We did this research because there were no research results that explicitly reported the comparison between the two classification methods, namely the LVQ neural network method and the Naive Bayes Classifier (NBC) method for the same image object. Thus the novelty of the research is the comparison of the results of the LVQ neural network method and the Naive Bayes Classifier (NBC) method in classifying the same image object.

2. Methods

2.1. Object of Research

The object used in the study was data in the form of images of 30 cantaloupe fruit skins. The image of cantaloupe rind consists of 15 images of raw cantaloupe rind and 15 ripe images of cantaloupe fruit skin. Figure 1 shows one of 15 images of raw cantaloupe rind skin and one of 15 images of raw cantaloupe rind skin, one of 15 images of ripe cantaloupe rind.

![Examples of Images of Cantaloupe Rind Used](image1.png)

**Figure 1.** Examples of Images of Cantaloupe Rind Used are (a) Raw and (b) Ripe

The image of cantaloupe rind used in the study was divided into 2 classes, namely class 1 which was an image of 15 raw cantaloupe fruit skin, and class 2 which was an image of 15 ripe cantaloupe fruit skin. Each of these classes is further divided into 8 images as training data and 7 images as test data.

2.2. Pre-Processing and Feature Extraction

Each image used in the study will be pre-processed. Image pre-processing aims to improve image quality and highlight the information contained in the image for further processing purposes. The stages of initial image processing are as follows:

1. Convert RGB image to greyscale
   Conversion is performed to convert a color image (RGB image) into a gray image (greyscale image). The purpose of conversion is to simplify the process of digital image processing. Grayscale images are simpler in processing than RGB images.

2. Median Filter
   The median filter used in this study is $9 \times 9$ matrices. The purpose of the filtering process is to reduce noise in the image.
3. Adaptive Histogram Equalization

Histogram equalization is a method to improve image quality by changing the value of the image's gray level distribution. In this process, the image is divided into several parts and each part of the image is subjected to a histogram equalization. The purpose of the histogram equalization is to produce a uniform image histogram.

After the initial image processing, next is the image feature extraction process. This process aims to bring out the characteristics of the image and bring up important information from the image. The image that has gone through the adaptive histogram equalization process is then extracted from the histogram feature values. The histogram features used in this study are mean, standard deviation, kurtosis, skewness, and entropy. The results of this histogram feature extraction are used as a feature of each image of cantaloupe rind in the classification process of cantaloupe fruit maturity.

2.3. Classification with Naive Bayes Classifier

The classification method used in this research is the Naive Bayes Classifier (NBC) method. NBC is a probabilistic based prediction technique based on the application of Bayes' Theorem. There are two processes in classification, namely the training process and the testing process.

1. Training Process

In this process, training is carried out on the training data by calculating the mean and standard deviation of each feature for each class. Mean (\(\mu\)) is calculated using the equation:

\[
\mu = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

Where \(n\) is the amount of data and \(x\) is the \(i\)-th image of each feature and each class. The standard deviation (\(\sigma\)) is calculated using the equation:

\[
\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2}
\]

The training process is continued by looking for the probability of each image on feature \(X\) against class \(Y = y\) using the Gaussian distribution expressed by the equation:

\[
P(X_i|Y = y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \left( \frac{X_i - \mu}{\sigma} \right)^2 \right]
\]

After obtaining the independent probability \(P(X_i|Y = y)\), the next process is to get the posterior probability or the final probability of Naive Bayes. From the two Naive Bayes probability values for class \(y = 1\) and \(y = 2\), a comparison will be made to get the largest value which will be classified into the result class. The resulting class is the output of the classification result. The resulting class that corresponds to the original class is the correct training data, while the result class that does not match the original class is incorrect training data. Next is the determination of the accuracy value in the training process. The accuracy value is the level of accuracy of the method in recognizing the input of a given object so that it can produce the correct output. The accuracy of the training process is expressed in the equation:

\[
\text{training accuracy} = \frac{\text{the correct amount of training data}}{\text{total training data}} \times 100
\]

2. Testing Process

The testing process is carried out on the test data based on the mean (\(\mu\)) and standard deviation (\(\sigma\)) values of the training data. The probability of each image on feature \(X\) against class \(Y = y\) or \(P(X_i|Y = y)\) searched using the probability Gaussian distribution. An important process in testing is the
determination of the posterior probability or the final probability of Naive Bayes. From the two Naive Bayes probability values for class $y = 1$ and $y = 2$, a comparison will be made to get the largest value which will be classified into the result class. The level of accuracy of the testing process is calculated using the equation:

$$\text{testing accuracy} = \frac{\text{the correct amount of test data}}{\text{total test data amount}} \times 100\% \quad (5)$$

3. Results and Discussion

3.1. Results of Training

The training process uses training data groups. The training data consisted of 8 images of raw cantaloupe rind (class 1) and 8 images of ripe cantaloupe rind (class 2). In each class and each feature on the training data, the average value (mean) is calculated using Eq. (1). Also, the standard deviation for each class and each feature is calculated using Eq. (2). Table 1 shows the mean ($\mu$) and standard deviation ($\sigma$) of each class and each feature.

| Class | Feature          | $\mu$  | $\sigma$ |
|-------|------------------|--------|----------|
| Mean  | 132.2219         | 9.8965 |
| Deviation Standard | 41.7873 | 9.2725  |
| Entropy | 7.3535           | 0.3195 |
| Skewness | -0.2082         | 1.0761 |
| Kurtosis | 6.7725           | 5.9938 |
| Mean  | 135.1843         | 8.5358 |
| Deviation Standard | 42.8205 | 12.8668 |
| Entropy | 7.3530           | 0.4194 |
| Skewness | 0.0065           | 0.2858 |
| Kurtosis | 5.5125           | 5.2855 |

The mean and standard deviation values are used as input parameters to determine the probability $p(X|y = 1, 2)$ of each image on a feature $X$ against class $Y = y$. Determination of probability using the Gaussian distribution of Eq. (3). After this probability is calculated for each image in each class and each feature, the next step is to calculate the final probability of Naive Bayes $P(y|X)$ with the initial probability $P(X) = 0.5$ because the amount of data for the two classes is the same.

The classification results were obtained after the final Naive Bayes probabilities $P(y = 1|X)$ were compared with the results $P(y = 2|X)$. In other words, the final Naive Bayes probability values obtained in the two classes were compared. The greatest value of the two classes in each image that will be used as the resulting class of the image. The following is the classification result of the training process on the training data.

The results of the training process can be expressed in terms of the level of system accuracy or performance. The level of accuracy of this system is used to see how good the Naive Bayes Classifier method is in recognizing data classes. Based on the results in Table 2, the accuracy of the training process can be calculated using Eq. (4). The result is a training accuracy of $11/16 \times 100\% = 68.75$
Table 2. Classification Results in The Training Process

| Image Number | Information | Class | Result |
|--------------|-------------|-------|--------|
| 1            | Raw         | 1     | 1      |
| 2            | Ripe        | 2     | 2      |
| 3            | Raw         | 1     | 2      |
| 4            | Ripe        | 2     | 2      |
| 5            | Raw         | 1     | 1      |
| 6            | Ripe        | 2     | 2      |
| 7            | Raw         | 1     | 2      |
| 8            | Ripe        | 2     | 2      |
| 9            | Raw         | 1     | 2      |
| 10           | Ripe        | 2     | 2      |
| 11           | Raw         | 1     | 1      |
| 12           | Ripe        | 2     | 2      |
| 13           | Raw         | 1     | 2      |
| 14           | Ripe        | 2     | 2      |
| 15           | Raw         | 1     | 2      |
| 16           | Ripe        | 2     | 2      |

Amount of training data 16
The correct amount of training data 11
The wrong amount of training data 5

It can be seen that the accuracy obtained in the training process using the Naive Bayes Classifier method is lower than the LVQ Artificial Neural Network method which has an accuracy of 75% [14]. In contrast to the results of this study, another study using the Naive Bayes Classifier method for the classification of apples has an accuracy of 91% in the training process [17].

3.2. Results of Testing

The testing process uses the test data group. The test data consisted of 14 images of cantaloupe rind which were divided into 7 images of raw cantaloupe rind (class 1) and 7 images of ripe cantaloupe rind (class 2). The testing process depends on the mean ($\mu$) and standard deviation ($\sigma$) of the training data in Table 1. Testing is done by calculating the probability of each test data image on feature X against class Y = 1, 2 or $P(X|Y=1,2)$ that is searched using the Gaussian probability distribution (Equation 3).

The Naive Bayes Classifier is an application of the Bayes Theorem which has a strong assumption of independence. In this study, the assumption in question is that the features are not interdependent (independent). After the probabilities for each image in each class and every feature are calculated, the next step is to calculate the final probability of Naive Bayes $P(Y|X)$.

The Naive Bayes classification is done by comparing the final Naive Bayes probability (posterior probability) for class 1 $P(Y=1|X)$ with the Naive Bayes final probability (posterior probability) for class 2 $P(Y=2|X)$. The greatest value of the two classes in each image that will be used as the class of the image test results. The results of the classification of the testing process can be seen in Table 3. Based on Table 3, the accuracy of the Naive Bayes Classifier testing process is 8/14 x 100% = 57.14%.
Table 3. Classification Results in The Testing Process

| Image Number | Information | Class | Result |
|--------------|-------------|-------|--------|
| 1            | Raw         | 1     | 2      |
| 2            | Ripe        | 2     | 1      |
| 3            | Raw         | 1     | 1      |
| 4            | Ripe        | 2     | 2      |
| 5            | Raw         | 1     | 2      |
| 6            | Ripe        | 2     | 2      |
| 7            | Raw         | 1     | 1      |
| 8            | Ripe        | 2     | 2      |
| 9            | Raw         | 1     | 2      |
| 10           | Ripe        | 2     | 2      |
| 11           | Raw         | 1     | 2      |
| 12           | Ripe        | 2     | 2      |
| 13           | Raw         | 1     | 2      |
| 14           | Ripe        | 2     | 2      |

Amount of testing data 14
The correct amount of testing data 8
The wrong amount of testing data 6

3.3. Discussion

In this study, the same image data of cantaloupe rind used in previous studies, but using a different method. The classification results of the Naive Bayes Classifier method only produce a test accuracy of 57.14%. The test accuracy is smaller when compared to the LVQ Neural Network method which has a testing accuracy of 78.6% [14]. Different results were obtained because the LVQ Neural Network method had parameters that could be set in value, for example, iteration (epoch), learning rate (α), and minimal error. Meanwhile, the Naive Bayes Classifier method only relies on input parameters which are absolute and cannot be changed in value. However, the Naive Bayes Classifier method has advantages over the LVQ Neural Network method, including that it does not require the initialization of input parameters.

Although the accuracy of the results obtained in this study, both training accuracy and testing accuracy are lower than the LVQ Neural Network method, other studies are using the Naive Bayes Classifier method with various research objects that have higher accuracy results. For example, research on the classification of apples uses the Naive Bayes Classifier method which has an average testing accuracy of 91%. The data used in this study were quite a lot, there were 3 varieties of apples classified with 50 images of each apple variety [17]. The results of these studies indicate that the Naive Bayes Classifier method can also produce fairly good accuracy.

The results of the testing accuracy of the Naive Bayes Classifier method which is lower than the LVQ Artificial Neural Network method in this study could also be caused by the limited amount of image data, which is only 15 image data for each group. Different results may be obtained if the number of image data is multiplied, for example up to 50 image data such as in the Misigo study on apples [17]. Therefore, further research needs to be carried out to test the classification accuracy using the Naive Bayes Classifier method compared to the LVQ Neural Network method on a larger amount of image data.

4. Conclusion

The classification results of cantaloupe fruit maturity using the Naive Bayes Classifier (NBC) method based on the image samples of cantaloupe fruit skin color used in this study have accuracy in training and testing of 68.75% and 57.14%, respectively. Comparing the results of the Naive Bayes Classifier
(NBC) method with the Learning Vector Quantization (LVQ) Artificial Neural Network method for the same cases, it is concluded that the classification accuracy of cantaloupe fruit maturity using the Naive Bayes Classifier method is lower than the LVQ Artificial Neural Network method. These results apply to both training and testing.

Further research needs to be carried out to test whether it is true that the classification accuracy using the Naive Bayes Classifier method is lower than the LVQ Artificial Neural Network method generally applies in various other cases.

**Acknowledgments**
The author would like to thank the Department of Physics Faculty of Science Institut Teknologi Sepuluh Nopember (ITS) Surabaya for research facilities support so that this research can be carried out well.

**Reference**

[1] Xiao Q, Niu W, and Zhang H 2015 *6-th IEEE: Intern. Conf. Softw. Eng. Serv. Sci. (ICSESS)* 944 DOI: https://doi.org/10.1109/ICSESS.2015.7339210.

[2] M R Satpute and S M Jagdale 2016 *Int. Res. J. Eng. Tech. (IRJET)* 3 703.

[3] Tu S, Xue Y, Zheng C, Qi Y, Wan H, and Mao L 2018 *Biosyst. Eng.* 175 156.

[4] Khojastehnazhand M, Omid M, and Tabatabaeefar 2010 *Afr. J. Plant Sci.* 4 122.

[5] Pu Y Y, Feng Y Z, and Sun D A 2015 *Comp. Rev. Food Sci, Food Saf.* 14 176.

[6] Chaudhry N M A 2019 *Innovative and Non-Destructive Technologies to Evaluate Quality of Rocket Leaves for Ready to Eat Salads* Doctoral Thesis (Foggia: Universita di Foggia)

[7] Naik S and Patel B 2017 *Int. J. Comput. Appl.* 170 22.

[8] A Alipasandi, Ghaffari H, and Alibeygly S Z 2013 *Int. J. Agron. Plant Prod.* 4 2179.

[9] Kavitakomal and Sonia 2019 *Int. J. Sci. Tech. Res.* 8 463.

[10] Bustomi M A 2019 *J. Phys. Res. Appl.* 15 119.

[11] Bustomi M A, Faricha A, Ramdhan A, and Faridawati 2018 *ARPN J. Eng. and Appl. Sci.* 13 718.

[12] Misigo R 2016 *Classification of Selected Apple Fruit Varieties Using Na`ive Bayes* Master Thesis (Nairobi: Nairobi University)