Identification system for beeswax on Malang apple using VNIR imaging

Naufal Praditya and Adhi Harmoko Saputro*

Department of Physics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Indonesia

*adhi@sci.ui.ac.id

Abstract. Wax coating identification on fruits is very difficult without a non-destructive method. In general, destructive methods were used to identify wax or coatings by soaking the fruit in hot water or using a mixture of vinegar and water. There are also destructive systems that was used such as gas chromatography linked with mass spectrometry, but this method takes much time and difficult to operate. Visible Near Infrared Imaging (VNIR) becomes the alternate solution to identify wax on the surface of the fruit without spoiling the quality of the fruit. In this study, identification system for beeswax application on apples has been made successfully. The process starts through image acquisition, image correction, object detection, window averaging, classification model, and the coating status. The VNIR image was acquired on a wavelength range from 400 to 1000 nm. The data was divided for training and testing the classification model using cross-validation method, then the model was evaluated using confusion matrix. Several classification models were used to compare the result and to conclude which model gives the best accuracy for identification and classification problems. The accuracy of the three models were 72.92% for PCA-SVM model, 81.25% for DT model, and 91.67% for RF model.

1. Introduction
Coating of apples with wax has been done since the beginning of the 19th century to replace some of the natural waxes that was removed in the washing and cleaning process, thus reduces water loss during handling and marketing [1]. To distinguish wax-coated apples and non-waxed apples, several methods have been applied such as gas chromatography linked with mass spectroscopy, high-performance liquid chromatography, etc., but this method is very time-consuming and requires high expertise for the operator [2]. Identifying wax on apples was not an easy task. It cannot be distinguished using human eyes, since we don’t have the dataset of the fresh and the waxed apples. Machine Learning becomes the solution for identification problems using pattern recognition. In this case, classifying the fresh and the waxed apples requires supervised learning algorithms such as Support Vector Machines, Decision Tree, Random Forest, etc.

Visible Near Infrared image, or often called as hyperspectral imaging is a fast and easy method to operate, it is a dominant technique which acquires high quality spatial and spectral information in a certain range of electromagnetic wavelengths to determine the external and internal quality of a product [3]. Therefore, hyperspectral imaging technique can be an alternate solution to identify wax coating on apples. Some studies on the identification or classification of fruits and vegetables has been made using...
hyperspectral imagery, such as application of edible coatings on fruits and vegetables [4], identification of bruise and fungi contamination in strawberries [5], early identification of charcoal rot disease in soybean stems [6], and classification of fungal infected date fruits [7].

In this study, waxed and non-waxed Malang apple was used as an object to be classified using classification models such as Support Vector Machine, Decision Tree, and Random Forest. Then, the three model was compared to find out which one has the highest accuracy for classification problems. Confusion Matrix was used as an evaluation method for the models.

2. Design and Implementation

2.1. Sample Preparation
The samples were selected from 20 Malang apples, which were supplied directly from Batu, Malang, Indonesia. The selection process was based on the surface of the apples, where there should not be any bruises. Assuming these samples were fresh and was not waxed, all of them were waxed manually using the wax emulsion after acquiring their unwaxed hyperspectral image. After the apples were waxed by brushing method, the waxed hyperspectral image was acquired. The position of each unwaxed and waxed apples sides are the same.

The wax emulsion was created using 25 grams of beeswax, 60 mL of coconut oil, and 50 mL of sunflower oil [8]. The beeswax was melted first until 90°C, then the coconut oil and the sunflower oil was poured and stirred altogether. After done mixing, the emulsion was let sit for 2 hours until it becomes a viscous fluid and the emulsion was ready to be applied.

2.2. Measurement System
Some hardware and software were used in this study to acquire the result. The hardware is Specim FX-10 as the hyperspectral camera, halogen light as light sources, Teflon board and the slider, and the last is the computer. For the software, CNC USB Controller was used to move the slider and Lumo Recorder was used to capture the hyperspectral image. The system was designed as shown in Figure 1.

![Figure 1. Design of the Measurement System.](image)

2.3. VNIR Image Acquisition and Processing
In this study, there are several processing steps that starts from image acquisition, image correction, object detection, window averaging, and classification modelling. There were three images that was acquired using the hyperspectral camera, it is the white reference, the dark reference, and the sample. The white reference was taken with the halogen lights turned on to reflect the light onto the camera from the Teflon board. The dark reference was taken with the lights turned off and the camera lens was closed using the lid. After calibrating the camera using both references, the sample image was acquired. Later, the acquired image was corrected to be normalized using the same white and dark reference through
Equation 1, where R is the reflectance, S is the sample, D is the dark reference, and W is the white reference.

\[ R = \frac{S-D}{W-D} \times 100\% \]  

(1)

After the correction process was done, the object was detected using the background removal method to separate the sample from the Teflon background. Next, three region of interest (ROI) with 10x10 pixels in size, was placed in the middle of each sides of the apple which has the highest reflection from the light sources. The averaging stage was done by the window averaging method where the pixel size that has been determined from the ROI will be averaged so that it becomes one pixel with 224 spectral bands. From one pixel, there are hundreds of features that have been extracted, for that we need a reduction algorithm in order to choose some parts of the data from several features to be processed. In this case, Principal Component Analysis was used as a dimensional reduction for the Support Vector Machine classification model, while for the Decision Tree and Random Forest classification models has their own process called as the feature selection. For data validation, cross-validation (CV) was done by dividing the data into training data and testing data by 80% and 20%. The last step is to evaluate the performance of the classification model that has been made using confusion matrix. Figure 2 shows the hyperspectral image processing design.

3. Result and Discussion

3.1. Reflectance Profile

From a total of 240 region of interests, Figure 3 shows the reflectance profiles of 20 Malang apples that was acquired. Having a lot of data makes it difficult to see the difference in the graph, thus the data was averaged from both classes in order to see the difference between the two. As seen in Figure 4 which shows the averaged reflectance profiles, the non-waxed line was slightly lower than the waxed line in the wavelength range of 800 to 900 nm. This information could mean that the waxed apples has a higher reflection because of the beeswax coating. To process this information further, Machine Learning was applied to classify the samples into two class, “waxed” and “non-waxed”.

![Figure 2. VNIR Image Processing.](image)
3.2. Cross-Validation Results of the Classification Models

Table 1 shows the average cross-validation accuracy of the three classification models. In the Support Vector Machine model, the average accuracy value of the training data and testing data obtained is not good. These results are said to be underfitting, where the model cannot be made well through training data or generalize new data. Generalization in this case refers to how well the concepts learned by the current model are applied to specific examples that are not visible to the model.

In the Decision Tree model, the average value of accuracy from training data and testing data has a huge difference. This result is said to be overfitting, where the model studies detail and noise in training data so that it has a negative impact on the model when studying new data. However, the details and noise on the training data do not necessarily apply to the testing data so that the model fails to generalize it on data testing and has decreased accuracy compared to the training data.

In the Random Forest model, the average accuracy value of the training data and testing data obtained, has a better result than the Decision Tree model. Technically, Random Forest is a collection of Decision Trees where the results of its classification use the majority vote method and eliminate details and noise in the training data, thus making the classification result better than a single decision tree.

| Model                  | Average Cross-Validation Accuracy |
|------------------------|----------------------------------|
|                        | Training Data (%) | Testing Data (%) |
| Support Vector Machine | 71.46              | 67.08            |
| Decision Tree          | 98.33              | 75.83            |
| Random Forest          | 99.90              | 85.83            |

3.3. Confusion Matrix Results of the Classification Models

Table 2 shows the confusion matrix result of the PCA-SVM model. Poor accuracy results occur because there are other features in apples besides the wax coating and SVM cannot choose the features used in the classification. In this case, SVM needs to select the kernel function and determine the correct parameter C. Although both can improve the performance of SVM, it needs more time to compute because for each different kernel function, the SVM algorithm becomes a different algorithm as well. Whereas to determine the proper kernel function, we must provide the actual structure of the data as input.
Table 2. Confusion Matrix for Testing Data of Support Vector Machine Model.

| Prediction Class | Waxed | Non-waxed |
|------------------|-------|-----------|
| Real Class       |       |           |
| Waxed            | 13    | 11        |
| Non-waxed        | 2     | 22        |

Table 2 shows the confusion matrix result of the Decision Tree model. The Decision Tree model provides better classification results than the Support Vector Machine model because it uses the node splitting method. This method classifies the 'waxed' and 'non-waxed' classes based on their reflectance. Each branch represents results and each split class label is called a leaf or node. This method can provide better classification results compared to the scatter plot method and determine the hyperplane of the Support Vector Machine model.

Table 3. Confusion Matrix for Testing Data of Decision Tree Model.

| Prediction Class | Waxed | Non-waxed |
|------------------|-------|-----------|
| Real Class       |       |           |
| Waxed            | 20    | 4         |
| Non-waxed        | 5     | 19        |

Table 3 shows the confusion matrix result of the Decision Tree model. The Decision Tree model provides better classification results than the Support Vector Machine model because it uses the node splitting method. This method classifies the 'waxed' and 'non-waxed' classes based on their reflectance. Each branch represents results and each split class label is called a leaf or node. This method can provide better classification results compared to the scatter plot method and determine the hyperplane of the Support Vector Machine model.

Table 4. Confusion Matrix for Testing Data of Random Forest.

| Prediction Class | Waxed | Non-waxed |
|------------------|-------|-----------|
| Real Class       |       |           |
| Waxed            | 22    | 2         |
| Non-waxed        | 2     | 22        |

Table 4 shows the confusion matrix result of the Random Forest model. The Random Forest model provides the best results from the two previous models. In theory, Random Forest is a collection of Decision Trees so that it can provide better results than a single Decision Tree. In addition to using the node splitting method in each Decision Tree, this model uses the majority vote method to determine the most classes from each tree so that it can provide maximum results.

4. Conclusion
Based on the results obtained in this study, it can be concluded that the identification system of beeswax coating in Malang apples using VNIR imaging can be made using a classification prediction modelling algorithm. Usually for most cases, Support Vector Machine was the best classification model, but for this case, it has failed to generalize new data for some reason. This problem could lead to another study to identify which cases would affect different classification models. The Random Forest algorithm was the best performing classification model out of the other two models with 91.67% accuracy of the
training data and a slight difference between the training data and the testing data from the cross-validation process.

Acknowledgement
This study was carried out with the partial support of Universitas Indonesia (PITTA Research Grant 2019) and Ministry of Research, Technology and Higher Education of the Republic of Indonesia (PUPT Research Grant 2019).

References
[1] A. H Bahnasawy, “Effect of Wax Coating on the Quality of Cucumber Fruits during Storage,” J. Food Process. Technol., vol. 05, no. 06, 2014.
[2] H. Wang, H. Zhu, Z. Zhao, Y. Zhao, and J. Wang, “The study on increasing the identification accuracy of waxed apples by hyperspectral imaging technology,” Multimed. Tools Appl., vol. 77, no. 20, pp. 27505–27516, 2018.
[3] N. T. Vetrekar et al., “Non-invasive hyperspectral imaging approach for fruit quality control application and classification: case study of apple, chikoo, guava fruits,” J. Food Sci. Technol., vol. 52, no. 11, pp. 6978–6989, 2015.
[4] V. T. Kore, S. S. Tawade, and J. Kabir, “Application of Edible Coatings on Fruits and Vegetables Application of Edible Coatings on Fruits and Vegetables Department of Post Harvest Technology of Horticultural Crops ,” no. December, 2016.
[5] Q. Liu, K. Sun, J. Peng, M. Xing, L. Pan, and K. Tu, “Identification of Bruise and Fungi Contamination in Strawberries Using Hyperspectral Imaging Technology and Multivariate Analysis,” Food Anal. Methods, vol. 11, no. 5, pp. 1518–1527, 2018.
[6] K. Nagasubramanian, S. Jones, S. Sarkar, A. K. Singh, A. Singh, and B. Ganapathysubramanian, “Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems,” Plant Methods, vol. 14, no. 1, pp. 1–14, 2018.
[7] M. A. Teena, A. Manickavasagan, L. Ravikanth, and D. S. Jayas, “Near infrared ( NIR ) hyperspectral imaging to classify fungal infected date fruits,” J. Stored Prod. Res., vol. 59, pp. 306–313, 2014.