Classifying Physical Morphology of Cocoa Beans Digital Images using Multiclass Ensemble Least-Squares Support Vector Machine

Armin Lawi¹ and Yudhi Adhitya²
¹Department of Computer Science, Hasanuddin University, Makassar
²Department of Informatics, Al-Asy’ariah Mandar University, Polewali Mandar

E-mail: yudhiadhitya@gmail.com

Abstract. The objective of this research is to determine the quality of cocoa beans through morphology of their digital images. Samples of cocoa beans were scattered on a bright white paper under a controlled lighting condition. A compact digital camera was used to capture the images. The images were then processed to extract their morphological parameters. Classification process begins with an analysis of cocoa beans image based on morphological feature extraction. Parameters for extraction of morphological or physical feature parameters, i.e., Area, Perimeter, Major Axis Length, Minor Axis Length, Aspect Ratio, Circularity, Roundness, Ferret Diameter. The cocoa beans are classified into 4 groups, i.e.: Normal Beans, Broken Beans, Fractured Beans, and Skin Damaged Beans. The model of classification used in this paper is the Multiclass Ensemble Least-Squares Support Vector Machine (MELS-SVM), a proposed improvement model of SVM using ensemble method in which the separate hyperplanes are obtained by least square approach and the multiclass procedure uses One-Against-All method. The result of our proposed model showed that the classification with morphological feature input parameters were accurately as 99.705% for the four classes, respectively.

1. Introduction
Cocoa plant (Theobroma cacao L.) is one of the important plantation commodities as a source of industrial raw materials trade commodities that can increase the country's foreign exchange and income of cocoa farmers. Indonesia is one of the producers of cocoa [1], most of Indonesia's cocoa production is exported to America, Singapore, Malaysia, Brazil, and China. Indonesia's cocoa beans production significantly increases, but the quality produced is very low and various of them are less fermented, not dry enough, beans size is evenly high skin content, high acidity, taste is very diverse and inconsistent. This is reflected in the relatively low price of Indonesian cocoa beans and discounted prices compared to similar products from other producer countries [2].

Indonesian cocoa farmers generally apply a variety of fermentation ways in terms of beans quantity, fermentation means and time. Fermentation is done in baskets, simple wooden crates or plastic bags. What farmers do is not real fermentation because most farmers keep the harvested beans in plastic bags for 1-2 days then dried by drying in direct sunlight on cement floors, mats or woven bamboo. The requirements or conditions used to determine the quality of cocoa beans in Indonesia are contained in the Indonesian national standard of cocoa beans SNI 2323-2008. The Indonesian national standard regulates the classification of the quality of dry cocoa beans as well as general requirements.
and in particular to maintain the consistency of the quality of the cocoa beans produced. At the exporter level, the separation is carried out by using machinery primarily for the classification of cocoa beans. The results of this cocoa bean sorting will be determined by taking samples of cocoa beans to be analyzed in the laboratory in accordance with the standards of cocoa beans quality classification [3].

Quality examination of cocoa beans is done using traditional and manual procedures; i.e., using the visual method on cocoa beans by choosing one-by-one. The human vision must accurately see the object on the surface of the cocoa beans. In plain view, a human without special knowledge can differentiate and classify cocoa beans. Usually, they only armed with experience and knowledge gained earlier. However, manual checks have limitations such as tired eyes and different analytical results of each examiner.

The development of automation technology and intelligent system is one form of technology created by humans to facilitate human activities and has made many changes in the production and processing of plantation commodities. Image processing technology provides an alternative to manual inspection. Digital technology has a variety of input devices including the camera. The output produced by the camera is an image (digital image). Imagery can be analyzed and processed to get useful information for the user.

This system has two important stages, i.e., image analysis and pattern recognition. Image analysis has standard techniques for identifying, measuring, and acquiring large quantities of quantitative data. Image processing techniques include image capture, pre-processing, interpretation, quantization and image classification. But unfortunately, the resulting image is still not in accordance with the results expected by the user. Therefore, the existence of a process that can process an image is needed by the user. The discipline that gave birth to the techniques to process the image is called Digital Image Processing [4, 5].

2. Related Works
Image processing techniques have been widely used in the field of agriculture such as determining the type of defects of coffee beans, edamame quality determination, quality inspection of RSS rubber, mango quality determination, and identification of maturity level of lemon and mangos teen, identification of defective cocoa beans, quality determination of cocoa beans.

I Wayan Astika, et al. use ANN structures to develop the relationship between input parameters, quality components of cocoa beans and outputs. ANN classified cocoa beans into 4 parameters namely: Normal Beans, Broken Beans, Fractured Beans, and Skin Damaged Beans. The classification of the size of the damage beans reaches an accuracy of 79.25%, which consists of accuracy: 84.38% for Normal Beans, 51.72% for Broken Beans, 98.29% for Fractured Beans, and 20% for Skin Damaged Beans [6].

S. Nurmuslimah, created a software system that starts with taking pictures of files to display on the system interface, image is processed using edge detection sobel to get numeric data value. Furthermore, these data are used as data input training Neural Network Back propagation. After training data is obtained, then the data is used for the testing process. From the testing process, the output of the created system is able to provide information about the quality of cocoa beans. Using Back propagation method with alpha = 0.6, hidden layer = 3, fault tolerance = 0.0001, target = 0.9, resulting in a system that has a level of accuracy of (76%) has an error rate (24%) in determining the quality of cocoa beans [7].

Data mining is a process uses statistical techniques, calculations, artificial intelligence and machine learning to extract and identify useful information and related knowledge from large databases [8]. SVM initially can only classify data in two classes [9, 10, 11]. However, further research SVM was developed so that it can classify data over two classes (multiclass) [9, 12, 13, 14]. Classifying M-classes means predicting the class labels $C_m$, $m = 1, ..., M$ one way to solve the M-class problem by formulating it into binary L classification problems [11, 12].
SVM concept is simply described as trying to find the best hyper plane that provides as a separator of two classes in the input space. Pattern which is a member of two classes: +1 and -1 and share alternate field separators. The best dividing fields can not only separate the data but also have the largest margins. Margin is the distance between the fields of separator (hyper plane) with the closest pattern of each class.

Let \( \{x_1, ..., x_n\} \) be the dataset and \( y_i \in \{+1, -1\} \) is the class label of the \( x_i \) data. The two classes are separated by a pair of parallel bounding plane. The first delimiter field limits the first class while the second delimiter field limits the second class, so it is obtained [11, 12]:

\[
\begin{align*}
    x_i, w + b &\geq +1 \text{ for } y_i = +1 \\
    x_i, w + b &\leq -1 \text{ for } y_i = -1
\end{align*}
\]

(1)

The best dividing fields with the largest margin values can be formulated into quadratic programming problems:

\[
\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i
\]

(2)

with constraints: \( y_i(x_i, w + b) \geq 1 - \xi_i \) and \( \xi_i \geq 0 \), where \( l = 1, ..., n \) is a slack slack variable that determines the level of misclassification of the data samples, whereas \( C > 0 \) is a parameter.

Method for classifying data that cannot be separated linearly is kernel method. The kernel method transforms the data into the feature space dimension so that it can be linearly separated on the feature space. The kernel method can be formulated:

\[
K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)
\]

(3)

Commonly used kernel functions are as follows:

- The linear kernel: \( K(x_i, x_j) = x_i^T x_j \)
- Kernel polynomial: \( K(x_i, x_j) = (\gamma x_i^T x_j + r)^p, \gamma \geq 2 \)
- RBF Kernel (Radial Basis Function): \( K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \)

LS-SVM was first introduced by Suykens and Vandewalle in 1999. LS-SVM is one of the SVM modifications that solve linear equations [12, 13, 14, 15, 16, 17]. If the SVM separator field is given as in (3), then for LS-SVM is given as the following objective function.

\[
\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \xi^T \xi
\]

(4)

with constraint: \( y_i(x_i, w + b) \geq 1 - \xi_i \)

The above equation can be solved after forming Lagrangian:

\[
L = \frac{1}{2} \|w\|^2 + \frac{C}{2} \xi^T \xi - \sum_{i=1}^{l} a_i (y_i (\varphi(x_i), w + b) - 1 + \xi_i)
\]

(5)

where \( a_i \) is a Lagrangian multiplier whose value can be either positive or negative.

To optimize the conditions in (5), a decrease of \( w, b, \xi, \text{ and } a \) is equal to zero. The results of the process are as follows.

\[
\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{l} a_i y_i \varphi(x_i)
\]

(6)

\[
\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{l} a_i y_i = 0
\]

(7)

\[
\frac{\partial L}{\partial \xi} = 0 \rightarrow a = \gamma \xi, \quad i = 1, ..., N
\]

(8)

\[
\frac{\partial L}{\partial a} = 0 \rightarrow y_i (\varphi(x_i), w + b) - 1 + \xi_i = 0, i = 1, ..., N
\]

(9)
Using the One-Against-All (OAA) method, a binary model is constructed \( k \) (\( k \) is the number of classes). Each \( i \)-class model is trained by using the entire data. For example, there is a classification problem with 3 classes. For training use 3 pieces of binary classification and thus its objective function is given as follows.

\[
\min_{w_i, b_i, \xi_{t,i}} \frac{1}{2} \sum_{i=1}^{m} (w_i)^T w_i + \frac{C}{2} \sum_{i=1}^{m} \xi_{t,i}^2
\]

with constraint: \( y_{t,i}(\varphi_i(x_t)(w_i)^T + b_i) \geq 1 - \xi_{t,i} \).

Confusion matrix is a table that states the amount of test data that is correctly classified or not. The table calculates the following parameters in order to evaluate accuracy, sensitivity and false discovery rate of the learner LS-SVM model from the tested data.

- **True Positive (TP)** is the number of documents from class 1 is correctly classified as class 1.
- **True Negative (TN)** is the number of documents from class 0 are correctly classified as class 0.
- **False Positive (FP)** is the number of documents from class 0 incorrectly classified as class 1.
- **False Negative (FN)** is the number of documents from class 1 that are misclassified as class 0.

The calculation of accuracy is expressed by the equation:

\[
Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100\%
\]  (11)

The calculation of sensitivity is expressed by the equation:

\[
Sensitivity = \frac{TP}{TP + FN} \times 100\%
\]  (12)

The calculation of false discovery rate is expressed by the equation:

\[
False \ Discovery \ Rate = \frac{FP}{FP + TP} \times 100\%
\]  (13)

3. **Experimental Details**

The process of classifying cocoa beans is explained by the following activities:

1. Cocoa bean image capture
2. Extraction of morphological features using image processing.
3. Conducting the training process to obtain the classification model
4. Conducting the classification process using test data
5. From the classification process in obtaining the results of the classification of cocoa beans based on morphological features

3.1. **Physical Morphological Features**

Morphology is the geometric property of an image. In this case is the size and shape of the characteristics of cocoa beans, obtained from binary image analysis. The morphological features are extracted from binary imagery, and then displaying the following morphological features: Area, Perimeter, Major Axis Length, Minor Axis Length, Aspect Ratio, Circularity, Roundness, Ferret Diameter.

Cocoa bean image is processed using grayscale and thresholding and edge detection is done by using kernel sobel to detect the edge of the image. After that, the process of segmentation and then performed the process of morphological feature extraction.
3.2. Dataset

Based on the morphological features there are four classes namely, first class is Normal Beans, second class is Broken Beans, third class is Fractured Beans and fourth class is Skin Damaged Beans. Attributes for dataset consist of 8 pieces; i.e., Area, Perimeter, Major Axis Length, Minor Axis Length, Aspect Ratio, Circularity, Roundness, Feret Diameter. The data sample for dataset 2 is 2,400 items.

Variables used in this research are:

- Normal Beans, Broken Beans, Fractured Beans, and Skin Damaged Beans
- $y_i =$ class of dataset that is class 1, class 2, class 3 and class 4.
- $x_i =$ data features consist of 8 variables, i.e., Area, Perimeter, Major Axis Length, Minor Axis Length, Aspect Ratio, Circularity, Roundness, Feret Diameter.
- $\alpha =$ Lagrange multiplier.
- $w =$ normal field to support vector.
- $b =$ distance of the bounding plane to the center point.

Values of $\alpha$ and $b$ will be obtained after the training is completed. Value of $\alpha$ is then used to find the value of $w$.

3.3. Classification Process

The classification process is divided into two, training and testing process. The data used is the data of color feature extraction and morphological features as much as 100% data so it is expected to get the accuracy of multiclass classification on LS-SVM. Training process aims to build the MELS-SVM model by finding the parameters, i.e values $\alpha$, $w$, and $b$. After forming the MELS-SVM model, proceed with the testing process on the data to see the accuracy of the LS-SVM technique using the OAA method.

Based on the workflow, the classification process begins with the training process on the train data to obtain $\alpha$ and $b$ values using the Matlab r2016a software and the additional toolbox LSSVMlabv which can be downloaded on the site http://www.esat.kuleuven.be/sista/lssvmlab/toolbox.html. In the training process that generates $\alpha$ and $b$ values, the RBF kernel is used. Furthermore, after the obtained values of $\alpha$ and $b$, the next step is to find the value of $w$ then the value of $w$ and $b$ are used to arrange the separator function.

After getting the separation function model from the training post, proceed to the testing process. The testing process begins by inserting the item values ($x_i$) of the test data, then obtaining the prediction class. From this prediction class will be calculated the accuracy level of the method by finding the total class that is correctly predicted by the class of the test data.

![Figure 1. Results of the image analysis process for morphological features.](image-url)
3.4. Multiclass Algorithm Method
The analysis used in this multiclass classification is the OAA classification method. The multilingas method algorithm as follows:

1. Dataset input.
2. Identify the input dataset
   a. The values of the training data feature \((x_i)\)
   b. Class of training data \((y_i)\)
   c. Values feature test data \((x_t)\)
   d. Class of test data \((y_t)\)
3. Initiate objects on LS-SVM before performing the training process with the initlssvm function
   a. Specify data of training data feature \((x_i)\)
   b. Specifies the training data class \((y_i)\)
   c. Choose a classifier to classify data
   d. Selects the kernel and its parameters to use
4. Selecting the multilingual method code used (code_OneVsAll for OAA)
5. Conduct training process with trainlssvm function
6. Calculating values \(w\)
7. Make predictions based on the model obtained and determine data feature test data \((x_t)\) with simlssvm function
8. Create a confusion matrix
9. Calculate the level of accuracy with the formula:
   \[
   \lambda = \frac{C}{N} \times 100\%
   \]
   where \(C\) is the correct total of predictions and \(N\) is the total of all data tested.

3.5. Implementation Results on dataset
The separator function for the one against all method with the RBF (Radial Basis Function) kernel using the parameter \(\sigma = 0.5\) for dataset is as follows.

\[
f_1(x) = x_i \cdot w_1 + b = x_i \begin{pmatrix} 474.060 \\ 620.798 \\ 676.654 \\ 1,036.838 \\ 57.949 \\ -1,039.173 \\ -270.937 \\ 1,086.816 \end{pmatrix} + (-0.1041),
\]

\[
f_2(x) = x_i \cdot w_2 + b = x_i \begin{pmatrix} -2,866.019 \\ -1,880.457 \\ -2,295.395 \\ -1,234.875 \\ -2,077.650 \\ -598.255 \\ 443.043 \\ -2,034.123 \end{pmatrix} + (-0.7476),
\]
\[ f_3(x) = x_i \cdot w_3 + b = x_i. \]
\[
\begin{pmatrix}
-342.200 \\
-393.094 \\
-379.771 \\
-476.487 \\
-160.096 \\
-152.709 \\
925.336 \\
-483.535
\end{pmatrix}
+ (-0.4019). \quad (16)
\]

\[ f_4(x) = x_i \cdot w_4 + b = x_i. \]
\[
\begin{pmatrix}
-2,830.185 \\
-1,829.274 \\
-2,250.925 \\
-1,171.307 \\
-2,050.838 \\
-517.928 \\
130.739 \\
-1,976.306
\end{pmatrix}
+ (-0.7463). \quad (17)
\]

4. Results and Discussion

Values of \( w \) and \( b \) are based on RBF kernel and its parameters for the use of the One Against All method on dataset can be seen on table 1.

| \( w \) value | \( b \) value |
|--------------|--------------|
| \( w_1 \)    | \( b_1 \)    |
| \( w_2 \)    | \( b_2 \)    |
| \( w_3 \)    | \( b_3 \)    |
| \( w_4 \)    | \( b_4 \)    |
| 474.060      | -2,866.019   |
| 620.798      | -1,880.457   |
| 676.654      | -2,295.395   |
| 1,036.838    | -1,234.875   |
| 57.949       | -2,077.650   |
| -1,039.173   | -598.255     |
| -270.937     | 443.043      |
| 1,086.816    | -483.535     |

Based on table 1 and figure 2, the use of RBF kernel types and using the parameter \( \sigma = 0.5 \) has the highest accuracy on dataset has the highest accuracy (99.705). Accuracy for classification with the number of classes: 4 classes, i.e.: Normal Beans, Broken Beans, Fractured Beans, and Skin Damaged Beans.
Figure 2. Accuracy level of OAA method for each kernel in dataset.

5. Conclusion
In this research, the cocoa beans image is processed using grayscale and thresholding, and then edge detection is done by using a sobel kernel to detect the edges of the image. After segmentation process which extracts feature that produces data based on the extraction of physical morphological features parameters i.e. Area, Perimeter, Major Axis Length, Minor Axis Length, Aspect Ratio, Circularity, Roundness, Ferret Diameter.

The accuracy level is derived from the number of data items categorized into the correct class by the MELS-SVM model. The level of accuracy using one against all method on datasets with RBF type (Radial Basis Function) using parameter $\sigma = 0.5$ resulting 99.705% with four class classification.

Acknowledgments
The authors would like to express heartfelt thanks to Department of Computer Science, Hasanuddin University and Department of Informatics, Al-Asy’ariah Mandar University, Polewali Mandar for the research support and assistances.

References
[1] ICCO Annual Report 2012/2013 International Cocoa Organization
[2] Haryadi M and Supriyanto 2001 Pengolahan Kakao Menjadi Bahan Pangan, Pusat Antar Universitas Pangan dan Gizi (Yogyakarta: Universitas Gajah Mada)
[3] Biji kakao [SNI] Standard Nasional Indonesia 2008 SNI; (SNI 2323:2008, ICS 67.140.30 Badan Standardisasi Nasional)
[4] Sutoyo T, Mulyanto E, Suhartono V, Nurhayati D, Oky W 2009 Teori Pengolahan Citra Digital (Yogyakarta: ANDI)
[5] Pratt, William K 2007 Digital image processing, PIKS Scientific inside (John Wiley) 4th Edition
[6] Astika I W, Solahudin M, Kurniawan A, Wulandari Y 2010 Determination of Cocoa Bean Quality with Image Processing and Artificial Neural Network (Indonesia: University Bogor)
[7] Nurmuslimah S 2016 Implementasi Metode Backpropagation Untuk Mengidentifikasi Jenis Biji Kakao Yang Cacat Berdasarkan Bentuk Biji (Indonesia: ITATS)
[8] Turban E, et al. 2005 Decision Support Systems and Intelligent Systems (Yogyakarta: Andi Offset)
[9] Sembiring K 2007 Tutorial SVM Bahasa Indonesia (Bandung: Institut Teknologi Bandung)
[10] Kurniawan D and Supriyanto C 2013 Journal Teknologi Informasi
[11] Vapnik V and Cortes C 1995 Support-Vector Networks. Machine Learning 273-297
[12] Gestel T, Suykens J, Lanckriet G, Lambrechts A, Moor B and Vandewalle J 2002 Multiclass LS-SVMs: Moderated Outputs and Coding-Decoding Schemes. Neural Processing Letters 45-58
[13] Jafar N, Thamrin S A, Lawi A 2016 Proceeding of International Conference on Computational Intelligence and Cybernetic
[14] Nurkamila J, Lawi A and Thamrin S A 2016. Least Squares Support Vector Machine Untuk Klasifikasi Data Multikelas (Makassar : Hasanuddin University)
[15] Marliani R R, Lawi A and Thamrin S A 2016 Algoritma Adaboost untuk Optimasi Ensemble Least Squares Support Vector Machine (Makassar : Hasanuddin University)
[16] Xu Y, Lv X, Wang Z and Wang L 2014 Journal of Information Science and Engineering 1773-1787
[17] Zhou L, Lai K K and Yu L 2010 J. Expert Systems with Applications 127-133