eISSN: 1984-3909
Coffee Science, e192206, 2024

Doi: https://doi.org/10.25186/.v19i.2206

Analysis of shape features by applying gain ratio and machine learning for coffee bean classification

Anindita Septiarini[1], Hamdani Hamdani[1], Enny Itje Sela[2], Nurul Hidayat[3], Lasmedi Afuan[3]

[1]Department of Informatics, Mulawarman University, Samarinda, East Kalimantan, Indonesia
[2]Department of Informatics, University of Technology Yogyakarta, Yogyakarta, Special Region of Yogyakarta, Indonesia
[3]Department of Informatics, Jenderal Soedirman University, Purwokerto, Cental Java, Indonesia
Contact authors: anindita@unmul.ac.id; hamdani@unmul.ac.id; ennysela@uty.ac.id; nurul@unsoed.ac.id; lasmedi.afuan@unsoed.ac.id

Received in December 23, 2023 and approved April 24, 2024

ABSTRACT

Coffee is one of the daily consumed beverages in many countries. It is yielded from coffee beans, which have proceeded through several processes. Several common coffee beans have been produced in Indonesia, such as Arabica, Robusta, Liberica, and Excelsa. Nevertheless, many coffee fanatics are unable to distinguish the various coffee bean types visually based on those shapes. Accordingly, it is necessary to classify the types of coffee beans. This work proposes a classification method of coffee bean type to differentiate four coffee bean classes: Arabica, Robusta, Liberica, and Excelsa. The work applied training and testing steps. Both involved ROI detection, pre-processing, segmentation, feature extraction, selection, and classification. Image processing was used in ROI detection, pre-processing, and segmentation to simplify the procedure and separate the coffee bean from the background. The feature extraction produced 14 shape features to distinguish the coffee bean's class, but the proposed method's performance has yet to reach the optimal result. The gain ratio was used to reduce the features; hence, only 4 features were selected, including aspect ratio, eccentricity, equivalent diameter, and area. These features were utilized as input data for classification using Naive Bayes, Artificial Neural Network (ANN), Support Vector Machine (SVM), C4.5, and decision tree. The proposed method used 4 features and a decision tree classifier. The local dataset has 400 coffee bean photos in four classes of 100 images each. The images were divided for training and testing using k-fold 10 cross-validation. The accuracy evaluation parameter reached 0.995.

Key words: Coffee beans; otsu method; features reduction; cross-validation; decision tree.

1 INTRODUCTION

Coffee, brewed from roasted and ground coffee beans, is a common beverage. Coffee beans have risen in prestige in recent years. They are both works of artistic and technical skill and scientific artifacts. Coffee with a novel sensory perception is becoming desirable as consumers desire a more satisfying coffee-drinking experience. Coffee also possesses anti-cellulite, anti-cancer, anti-aging, and UV protection properties, in addition to its antioxidant and anti-inflammatory effects. That's why there's been such a surge in coffee's popularity in recent years. As a result, the cost of green coffee beans is at a 10-year high (Febrianto; Zhu, 2023; Santos et al., 2021).

Brazil, Vietnam, Colombia, and Indonesia are among the majority of the countries that produce and export coffee beans (Santos et al., 2020; Santos et al., 2021). Indonesia is one of the world's leading coffee growers, and its exports are increasing (Agus, 2014). Exports of coffee beans in Indonesia are one of the country's main sources of foreign funds. The country's economy relies heavily on the commodity due to the important function that coffee beans play in the nation's economy. Additionally, the coffee industry offers work opportunities for tens of thousands of Indonesians, including farmers and workers located throughout the country.

Various coffee beans such as Arabica, Robusta, Liberica, and Excelsa coffee beans are cultivated in Indonesia. They have unique flavor profiles and distinct qualities. Arabica coffee is renowned for its gentle flavor and acidity, whereas Robusta coffee is more pungent and contains more caffeine. Liberica and Excelsa are uncommon varieties prized for their distinctive and unique flavors (Agus, 2014). Each type of coffee bean can be distinguished visually based on the shapes; however, only a select few possess this information, and they are unable to recognize it. Therefore, a method for coffee bean classification was required.

Machine vision with an image processing approach was appropriate for classifying coffee bean varieties. It has been utilized in agricultural and food research for several works, such as identifying crop diseases (Darwish et al., 2020; Hamdani et al., 2021; Septiarini et al., 2021), monitoring the plant's development (Fahmi et al., 2018), grading the fruit maturity (Behera et al., 2021; Septiarini et al., 2020; Bazame et al., 2021; Lü et al., 2018) estimating the mass of fruits and vegetables (Jana; Parekh; Sarkar, 2020), also recognition of various kinds of fruits and vegetables (Munera et al., 2019; Piedad et al., 2018; Tan et al., 2018; Ji et al., 2019; Palacios-Morillo et al., 2016). In addition, coffee bean classification was implemented based on color, texture, or shape (Oliveira et al., 2016; Septiarini et al., 2022a; Febriana et al., 2022; Zhang; Liu; He, 2018).

Object classification was done using machine vision, typically involving three primary processes involved in this case, including image processing, feature extraction, and classification. Common techniques for image processing included pre-processing and segmentation (Septiarini, et al., 2022a; Septiarini, et al., 2022b; Sharif et al., 2018). Pre-processing comprised the employing of image scaling (Febriana et al., 2022; Septiarini et al., 2022a; Zou et al., 2021) and color space converting techniques (Septiarini et al., 2021; Zou et al., 2021). In contrast, Otsu (Dutta; Talukdar; Bora, 2022) and edge detection (Iqbal et al., 2018; Septiarini et al., 2022a; Sharif et al., 2018) were usually used for segmentation.

Furthermore, feature extraction based on color, shape, and texture was implemented (Alcayde et al., 2019; Arsalane et al., 2020). The color features were derived from color moments (Patil; Kumar, 2017; Septiarini et al., 2020), and histogram based (Behera et al., 2021; Hamdani et al., 2021; Tan et al., 2018), while the shape features were derived from the invariant moment (Rangkuti; Harjoko; Putra, 2021). While texture often employed gray level co-occurrence matrix (GLCM) (Behera et al., 2021; Utaminingrum et al., 2022), and Gabor Filter (Patil; Kumar, 2017). Several classifiers, such as ANN (Darwish et al., 2020; Oliveira et al., 2016; Hamdani et al., 2021; Piedad et al., 2018; Utaminingrum et al., 2022), Naive Bayes (Oliveira et al., 2016; Yang et al., 2019), Decision Tree (Zheng et al., 2023), SVM (Ji et al., 2019; Piedad et al., 2018; Septiarini et al., 2022a; Yang et al., 2019), KNN (Arboleda; Fajardo; Medina, 2018; Tan et al., 2018) were generally applied to various kinds of objects during the classification as the final main process.

Many previous studies have attempted to use image processing methods to classify coffee beans. The roasting grade classification of Arabica coffee beans was differentiated into three classes (light, medium, and dark). Six type features were derived, including kurtosis, entropy, mean, skewness, variance, standard deviation, and average intensity. The features of standard deviation and average intensity were extracted in each channel of RGB color space; hence ten features were obtained to distinguish each class. Those features fed into the SVM classifier. Four SVM kernels –Linear, polynomial, radial basis, and sigmoid functions– were utilized to evaluate the classifier performance. The polynomial kernel resulted in a maximum accuracy value of 100% using cross-validation with k-fold values of 5 and 10. The dataset used to evaluate the method consist of 150 images (50 images for each class) (Septiarini et al., 2022a). In another work, the image of roasted Excelsa coffee beans was acquired using a smartphone with white background. The color features based on all channels of RGB color space were extracted. The utilization of an artificial neural network was employed to categorize the various degrees of roast for coffee beans, specifically classifying them into the categories of light, medium, and extremely dark roast. The dataset comprises a total of 180 photos, with each class containing 60 images. The findings indicated that the approach developed successfully determined the roast level in coffee beans with a accuracy of 97.22% (Sarino et al., 2019).

Several advancements were made by developing a computer vision system for green coffee bean color classification using L*a*b* color space measurement. Whitish, cane green, green, and bluish-green coffee beans were separated using the Bayes classifier with ANN as the transformation model. The generalization error of the neural network models was 1.15%, while the Bayesian classifier achieved an accuracy value of 100% that indicated all data as to their expected classes were correctly categorized (Oliveira et al., 2016). Furthermore, a new dataset was produced with 8,000 images. It was split into 4 categories: Peaberry, Longberry, Premium, and Defect. This work proposed a lightweight and intelligible intelligent system that uses deep learning to aid farmers in sorting green bean arabica by variety. The benchmark deep learning models ResNet-18 and MobileNetV2 were used to demonstrate the baseline classification performance of the dataset. On average, these models could classify data with an accuracy of 81.12% and 81.31% (Febriana et al., 2022). Imaging techniques automated classifying coffee beans into three classes (Robusta, Excelsa, and Liberica). The shape features were determined, including bean area, perimeter, equivalent diameter, and roundness percentage. The dataset was divided into 195 images for training and 60 images for testing. The coffee granules were identified using two classifiers, ANN and KNN. The accuracy of ANN and KNN is 96.66% and 84.12%, respectively. It indicated that ANN could be utilized with the given dataset (Arboleda et al., 2018).

Prior study limitations resulted in the erroneous, necessitating development of the technique. The work related to the classification of coffee beans is challenging due to the similarity in color, shape, and texture of the bean types, which makes differentiation difficult. Regarding our investigation, no studies classified the coffee bean into four classes: Arabica, Excelsa, Liberica, and Robusta. In Arboleda et al. (2018), coffee beans were classified into three types (Robusta, Excelsa, and Liberica). The other previous studies classified the coffee bean according to roast grade (Sarino et al., 2019; Septiarini et al., 2022a), quality (Arboleda et al., 2018), and Arabica coffee variety (Oliveira et al., 2016; Febriana et al., 2022; Zhang et al., 2018). Therefore, this study was conducted to classify coffee beans into four distinct types: Arabica, Excelsa, Liberica, and Robusta. The shape features were extracted and reduced to create the feature sets, which served as the input data for the subsequent procedure. The feature selection was implemented using the grain ratio to reduce unimportant features. Following this, the KNN classifier was applied in the classification process. The proposed method successfully classified all coffee bean varieties with a maximum accuracy value.

The subsequent sections of this work are organized in the following manner. In Section 2, the dataset description and the recommended method employed are presented. The experimental results of the suggested method are outlined in Section 3.

Section 4 culminates by providing a conclusion of the research's discoveries and prospective directions for future work.

2 MATERIAL AND METHODS

Analyzing shape features for coffee bean classification requires a dataset of images. It was used to generate knowledge-based to recognize the type of coffee bean. This work defined a dataset to propose an approach for classifying coffee beans.

2.1 Dataset Description

The dataset utilized in this work was a set of images of coffee beans captured using a built-in digital camera on a smartphone (Sony a6000) with a resolution of 12 Megapixels. These images were saved in JPEG format with a size of 3376 × 6000 pixels. The acquisition process was carried out by placing the coffee bean in a studio minibox with 1 LED lamp with a power of 220 V using a simple white background. The studio minibox has dimensions of 30 cm in height, 20 cm in length, and 10 cm in width. The smartphone was mounted on a tripod, facing the object at a slope of 30 degrees, and is 30 cm away from the studio minibox. The dataset consists of four classes of coffee beans (arabica, excelsa, liberia, and robusta), with a total of 400 coffee bean images collected, with 100 images for each class. The images were then randomly divided into two sets: training and testing sets. Examples of coffee bean images in each class are presented in Figure 1.

2.2 Proposed Methodology

The objective of the suggested methodology was to accurately classify the different classes of coffee bean photos within the testing dataset. The process was bifurcated into two distinct stages: training and testing. The data for each phase was obtained from the training and testing sets. Both of the stages had four main processes: (1) ROI detection, (2) pre-processing, (3) segmentation, and (4) feature extraction. In the training phase, the features selection was applied to obtain the appropriate with a smaller number of features. Meanwhile, in the testing phase, only extracted the selected features. ROI detection consisted of five sequential steps: resizing, gray-scaling, median filtering, thresholding, and bounding box. Subsequently, preprocessing was done by implementing gray-scaling and histogram equalization steps followed by segmentation. The segmentation employed three steps: thresholding, morphology operations, and filling holes. Furthermore, the feature extraction generated the feature set as the input data in the learning process. Several classifiers performed the learning process to yield a predicted class of coffee beans (arabica/excelsa/liberica/robusta), followed by the method evaluation to justify the proposed method. An overview of all processes in the coffee bean classification method is depicted in Figure 2.

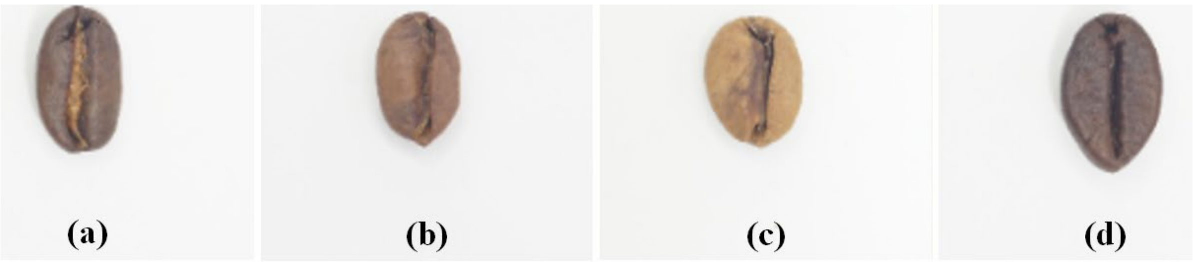


Figure 1: Examples of Coffee bean images with various varieties: (a) arabica, (b) excelsa, (c) liberica, and (d) robusta.

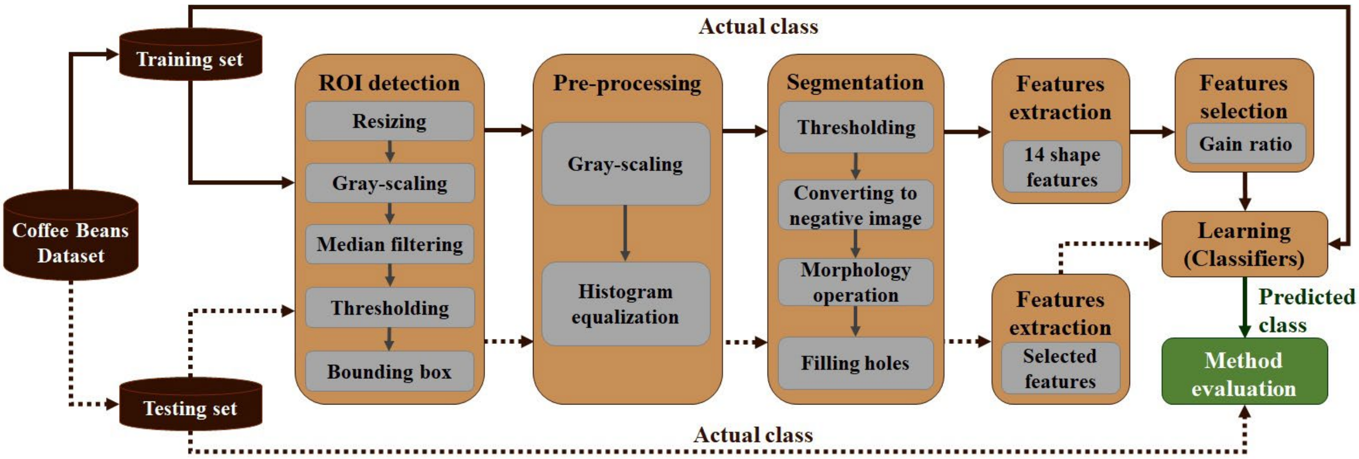


Figure 2: The overview of the process stages in the proposed Method.

2.3 Region of Interest (ROI) Detection

This process attempts to generate a sub-image that emphasizes the coffee bean region. This image subset is known as the ROI image. The proportions of the ROI images varied based on the size of the coffee beans. These dimensions were smaller than the original image because a significant portion of the background had been eliminated. Creating the ROI image decreased the computational processing time and enhanced the outcomes of the subsequent process. For computational efficiency, the ROI detection procedure was first used to resize the original images from 6000 × 3376 to 640 × 480 pixels (Figure 3(a)) (Lü et al., 2018; Zhuang et al., 2018). Subsequently, gray-scaling was performed to convert the RGB color to gray level color (Figure 3(b)), followed by median filtering (Figure 3(c)) to minimize the noise that occurs as a result of uneven illumination causing the occurrence of shadows, thereby altering the original shape of the coffee beans. The Otsu method was then used to implement thresholding because it has been extensively employed to estimate the location of an object in an image (Dutta et al., 2022; Nyalala et al., 2019). On the basis of the resulting binary image (Figure 3(d)) from the thresholding, the bounding box (Figure 3(e)) was set as a reference to construct the ROI image (Figure 3(f)). Figure 3 depicts the final image of each stage of ROI detection.

2.4 Pre-processing

In this process, gray scaling was initially applied by converting the ROI image from RGB color space to gray level color to simplify the subsequent process. Afterward, the histogram was equalized to make the characteristics of the coffee beans from each class more pronounced, allowing them to be distinguished more readily to achieve optimal classification results. The resulting image of each step in this process is depicted in Figure 4.

2.5 Segmentation

This study used segmentation to separate the coffee bean area from the background. There were three required steps to complete this process. Firstly, thresholding was performed using the Otsu method (Dutta et al., 2022; Nyalala et al., 2019) on the image produced by the previous process (Figure 4(c)).

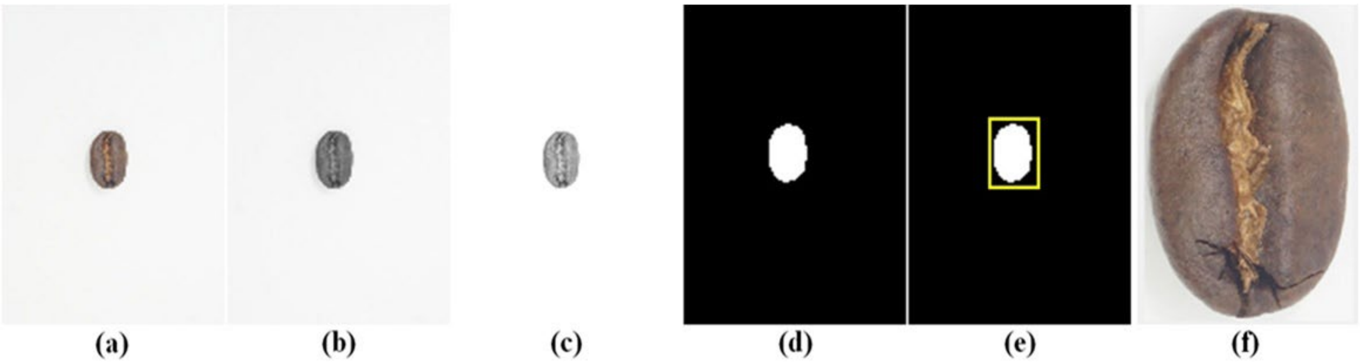


Figure 3: The resulting images of each step in the detection ROI: (a) resizing, (b) gray-scaling, (c) median filtering, (d) thresholding, (e) bounding box, and (f) ROI image.

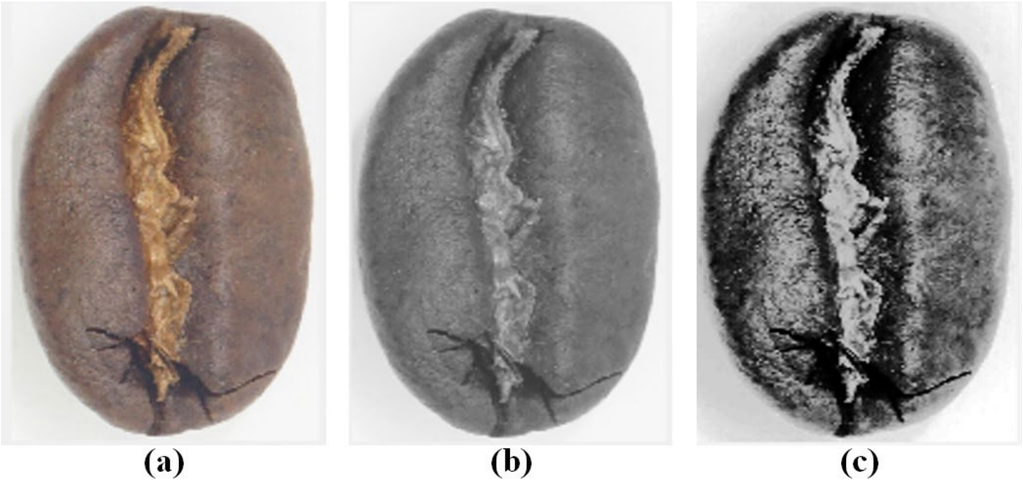


Figure 4: The resulting images of each step in the pre-processing: (a) ROI image, (b) gray-scaling, and (c) histogram equalization.

Subsequently, the resulting image of thresholding was converted to a negative image. Hence the presence of numerous holes, the last step of filling holes was carried out. Lastly, due to improper illumination during image capture resulting in the formation of shadow around the coffee bean in the image, the shadow was interpreted as the coffee bean area, thereby increasing its size. Therefore, the morphology operation required an erosion operation to resize the coffee bean tend to its actual size. Figure 5 depicts the resulting image of each segmentation process.

2.6 Feature Extraction

The characteristic of each coffee bean type can be differentiated visually based on the shape. Therefore, the shape feature is an appropriate descriptor to distinguish the type of coffee bean regarded from the image because there is a difference in structure between Arabica, Robusta, Liberica, and Excelsa coffee bean. The shape feature has been employed successfully to classify objects in previous studies (Patil; Kumar, 2017; Zhang; Wang; Duan, 2020; Zheng et al., 2023; Zou et al., 2021). This study extracted 14 features to investigate the most discriminant shape feature as a descriptor. The shape features were defined as follows (Zhang; Wang; Duan, 2020):

a. Area (*A*) is the number of pixels detected as coffee beans by the method.

b. Perimeter (*P*) is the number of pixels detected as coffee beans boundary.

c. Length (*L*) is the length of the coffee bean region.

d. Width (*W*) is the width of the coffee bean region.

e. Equivalent diameter (*D*) is the diameter of a circle equal to the coffee bean region.

f. Aspect ratio (*Ar*) is the ratio of the length to the width of the coffee bean region. *Ar* is defined by Equation 1.

$$Ar = \frac{L}{W} \tag{1}$$

g. The ratio of the equivalent ellipse axis (Re) is the ratio of the coffee bean equivalent ellipse's long axis to its short axis. Re is resulted using Equation 2.

$$\mathrm{Re} = \frac{a_{eq}}{b_{eq}} \tag{2}$$

h. Eccentricity (*Ec*) is the ratio of the distance between the coffee bean equivalent ellipse's focal points to its long axis. *Ec* is generated using Equation 3.

$$Ec = \frac{c_{eq}}{a_{eq}} \tag{3}$$

i. Compactness (*Ca*) is the ratio of the diameter of the equivalent circle to its length. *Ca* is computed using Equation 4.

$$Ca = \frac{D}{L} \tag{4}$$

j. The Heywood circularity factor (*Hcf*) quantifies the coffee bean's proximity to a circle. *Hcf* is applied using Equation 5.

$$Hcf = \frac{P}{\left(2 \times \sqrt{\pi \times 2}\right)} \tag{5}$$

k. Extent (*Ex*) is the ratio of the coffee beans to the external rectangular area. Ex is derived using Equation 6.

$$Ex = \frac{A}{A_{br}} \tag{6}$$

l. Solidity (So) is the ratio of the coffee beans region to the convex hull area. So is obtained using Equation 7.

$$So = \frac{A}{A_{ch}} \tag{7}$$

m. Arc is the ratio of the coffee bean to the circumcircle area. Arc is defined using Equation 8.

$$Arc = \frac{A}{A_{ec}} \tag{8}$$

n. Complexity (Cl) is the ratio of the square of the coffee bean's perimeter to its area. Cl is calculated using Equation 9.

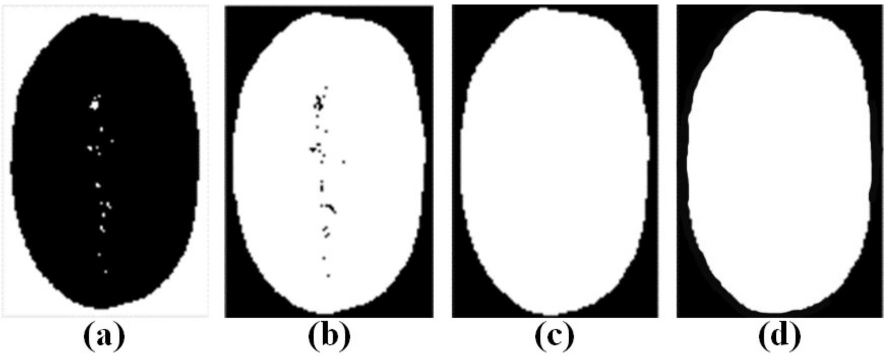


Figure 5: The resulting images of each step in the segmentation: (a) thresholding, (b) negative image, (c) filling hole, and (d) morphology operation.

$$Cl = \frac{P * P}{A} \quad (9)$$

The illustration of the coffee bean features parameter is depicted in Figure 6. The red lines in Figure 6 represent the external rectangle of the coffee bean contour, the green lines indicate the equivalent ellipse of the coffee bean contour, the yellow lines denote the convex hull of the fish contour, and the purple lines present the circumcircle of the coffee bean contour. A is the coffee bean's surface area, while perimeter P is the bean's boundary. Meanwhile, L and W represent the coffee bean's length and width. The corresponding ellipse of the coffee bean contour has a long axis ($2a_{eq}$), a short axis ($2b_{eq}$), and a focal length ($2c_{eq}$). Areas of a circle, a convex hull, and an exterior rectangle are denoted by A_{ec}, A_{ch}, and A_{br}, respectively.

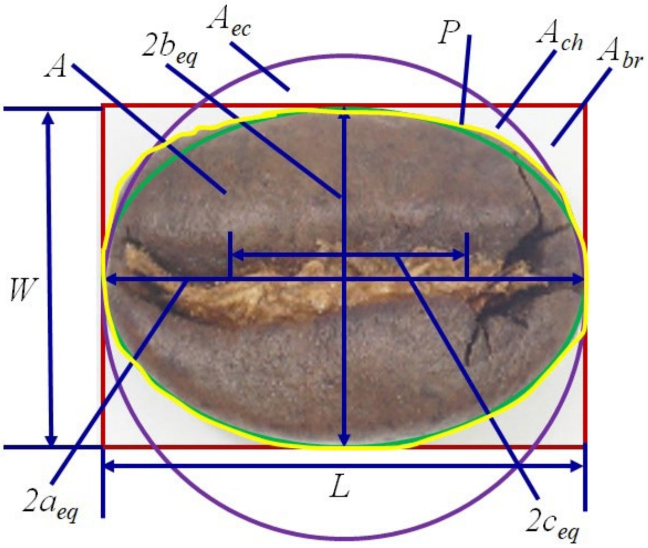


Figure 6: The illustration of the shape features parameter based on the coffee bean contour.

2.7 Feature Selection

A total of 14 features were extracted for this work. Nevertheless, not all of them may play an essential role. Therefore, a feature selection process was required to generate the necessary features and eliminate the remainder. The decreased number of features simplifies the ensuing process and reduces the computation time. In previous works, feature selection was undertaken using the gain ratio to reduce the number of features while achieving optimal classification results. In addition, this method has been applied to various objects associated with this work (Jamalludin et al., 2021; Pasha; Mohamed, 2022; Prasetiyo et al., 2021; Trabelsi et al., 2017).

This process is intended to decrease the number of features. The process yields shape information that serves as an essential indicator for distinguishing the four varieties of coffee beans: arabica, excelsa, liberica, and robusta. In this work, an attribute (*att*) was used for measuring information theory to show the difference between prior and expected posterior uncertainty. An *att* was chosen if its informational gain surpassed all other attributes. Within Information Theory Measures, Information Gain (*IG*) is a well-known attribute evaluation metric. The feature selection method computes the informative gain of each attribute *att* based on its contribution to class *Y*. This assignment is determined by the Equation 10 (Trabelsi et al., 2017):

$$IG(att) = H(Y) - H_{att}(Y) \quad (10)$$

H(*Y*) employs Equation 11 to determine the entropy of the class *Y*. Entropy, a mathematical function, approximates the quantity of information contained or delivered by an information source. The entropy is utilized for relevance measurements using Equation 11:

$$H(Y) = \sum_{i} -P(\upsilon_i)\log_2 P(\upsilon_i) \quad (11)$$

P(*vi*) indicates the probability of having the value *vi* for class *i* by adding to total values. By contributing to class *Y*, the Equation 12 calculates the entropy of the *att* attribute. *K* is the total number of partitions per class, *Pj* is the probability of finding a given number of instances of an attribute split, and *Yj* is the number of the *jth* division by class. Since this method works with categorical data, numerical attributes should be classified prior to the evaluation process.

$$H_{att}(Y) = \sum_{j=1}^{K} P_j H_{att}(Y_j) \quad (12)$$

The Gain Ratio (GR) attribute is assessed to address the limitation of Information Gain. The metric of Information Gain is commonly employed to evaluate the significance of a characteristic, although its effectiveness diminishes when applied to attributes that exhibit variability in their values. The purpose of the gain ratio is to discourage the excessive growth of nodes. It is designed to have a high value when data is evenly distributed among nodes and a low value when all data belongs to a single branch. Each attribute of the gain ratio is determined using Equation 13:

$$GR_{att} = \frac{IG(att)}{H(att)} \quad (13)$$

where $H(att)$ = Σj – P (v_j) log$_2$ $P(v_i)$ and $P(v_j)$ denotes the probability of having the value v_j by adding to the attribute j's overall value.

Based on the gain ratio (*GR*) values calculated for all features, eight features have *GR* values greater than 0. These consist of *Ar, Ec, D, A, L, P, W*, and *Cl*, with *GR* values of 0.622, 0.487, 0.479, 0.477, 0.474, 0.473, 0.462, and 0.279, respectively. A feature with a high *GR* value typically has an essential role; conversely, as the value decreases, the role of the feature becomes unimportant and can be discarded. Table I summarizes the outcomes of the GR values for each feature. Meanwhile, the examples of feature extraction results from features with a GR value of more than 0 are shown in Figure 7.

Based on the GR value in Table 1, this work generated 7 feature sets that contain features with a GR value of more than 0. The feature sets were formed to determine the most discriminatory combination of shape features capable of achieving optimal classification results. Those feature sets consist of the following:

a. Set-1: consists of 8 features, including *Ar, Ec, D, A, L, P, W*, and *Cl*. It was formed based on features with a minimum *GR* value of 0.279.

b. Set-2: consists of 7 features, including *Ar, Ec, D, A, L, P*, and *W*. It was formed based on features with a minimum *GR* value of 0. 462.

c. Set-3: consists of 6 features, including *Ar, Ec, D, A, L*, and *P* It was formed based on features with a minimum *GR* value of 0. 473.

d. Set-4: consists of 5 features including *Ar, Ec, D, A,* and *L*. it was formed based on features with a minimum *GR* value of 0.474.

e. Set-5: consists of 4 features, including *Ar, Ec, D*, and *A*. It was formed based on features with a minimum *GR* value of 0. 479.

f. Set-6: consists of 2 features, including *Ar* and *Ec*. It was formed based on features with a minimum *GR* value of 0. 487.

g. Set-7: consists of 1 feature, including *Ar*, it was formed based on feature with minimum *GR* value of 0.622.

2.8 Classification

The categorization of data was the motivation for implementing the classification process. Classification is one of the tasks where machine learning has been employed (Behera et al., 2021; Ji et al., 2019; Munera et al., 2019; Rajasekhar; Babu, 2020; Septiarini et al., 2022a; Velásquez et al., 2021). The field of machine learning focuses on the development and analysis of algorithms for learning from data and making predictions. It can be used to analyze visual information for recurring patterns. The success of a machine learning model is contingent on several factors, including the quality and quantity of training data, the methodology selected, and the accuracy with which the model's parameters are tuned. In this work, coffee bean classification was accomplished through machine learning. Several machine learning methods were employed, including Naive Bayes, ANN, SVM, C.45, and Decision Tree. Moreover, these methods have been successfully utilized to classify diverse objects (Behera et al., 2021; Piedad et al., 2018; Rajasekhar; Babu, 2020; Septiarini et al., 2022a; Zou et al., 2021).

2.9 Evaluation Method

Cross-validation with the k-folds value of 10 was conducted to separate training and testing data to evaluate the method performance (Septiarini et al., 2021). The performance is indicated by three evaluation parameters: precision, recall, and accuracy based on a confusion matrix multi classes that present information on the predicted class of coffee bean types resulting from the proposed method against the actual class (Deng et al., 2016). The types of coffee beans are divided into four classes, namely arabica (C1), excelsa (C2), liberica (C3), and robusta (C4). The confusion matrix for coffee bean classification with four classes is depicted in Table 2.

N_{ii} denotes the number of coffee bean images classified correctly as *Ci* using the proposed method. Meanwhile, the N_{ij} parameter represents the number of class *Ci* images the proposed method misclassifies into class *Cj*. Otherwise, N_{ji} calculates the number of class *Cj* images misclassified into class *Ci* by the method. The evaluation parameters had values ranging from 0 to 1. The parameter value close to 1 denotes that the proposed method was considered robust and reliable. The evaluation parameters of precision, recall, and accuracy are defined using Equations 14 - 16 (Deng et al., 2016):

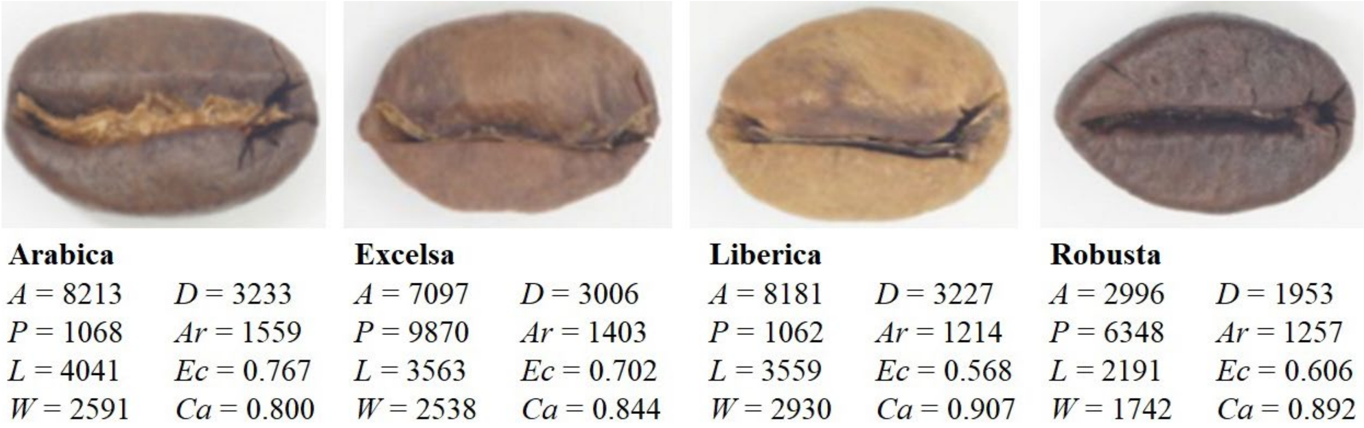


Figure 7: The examples of the features extraction result which has a GR value not equal to 0.

Table 1: The results of candidate selected features based on Gain Ratio rank.

Feature	IG Value	GR Value
Ar	1.794	0.622
Ec	0.221	0.487
D	1.681	0.479
A	1.681	0.479
L	1.666	0.474
P	1.686	0.473
W	1.844	0.462
Cl	0.571	0.279
So	0	0
Ex	0	0
Ca	0	0
Arc	0	0
Hfc	0	0
Re	0	0

Table 2: The confusion matrix for coffee bean classification with four classes.

		Predicted class				
		C1	*C2*	*C3*	*C4*	
Actual Class	*C1*	N_{11}	N_{12}	N_{13}	N_{14}	→ N_{ij}
	C2	N_{21}	N_{22}	N_{23}	N_{24}	→ N_{ii}
	C3	N_{31}	N_{32}	N_{33}	N_{34}	→ N_{ji}
	C4	N_{41}	N_{42}	N_{43}	N_{44}	

$$Precision = \frac{N_{ii}}{\sum_{j=1}^{n} N_{ji}}, \tag{14}$$

$$Recall = \frac{N_{ii}}{\sum_{j=1}^{n} N_{ji}}, \tag{15}$$

$$Accuracy = \frac{\sum_{i=1}^{n} N_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}}, \tag{16}$$

2 RESULT

Several investigations were conducted to validate the most discriminatory features and the most suitable classifier for the dataset employed in this work. Several feature sets, namely Set-1, Set-2, Set-3, Set-4, Set-5, Set-6, and Set-7, which feed various classifiers, were used to evaluate the method. Each classifier incorporated these collections. Hence, employing a 10-fold cross-validation technique, a dataset of 400 images (with 100 images for each class) was separated into four distinct classes, namely Arabica, Excelsa, Liberica, and Robusta, resulting in a total of 35 experiment scenarios. The proposed method's performance was evaluated by comparing the actual class of coffee bean based on observation result and the classification results of the proposed method.

3 DISCUSSION

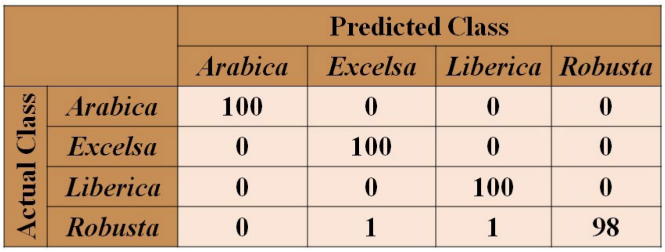
It was assessed by employing three assessment parameters: precision, recall, and accuracy. Table 3 presents a comprehensive overview of the performance evaluation of the coffee bean classification method across different feature sets, employing a range of classifiers.

Table 3 displays the varying performance of the proposed method. The extracted shape features exhibit high discriminatory properties. Consequently, the four coffee bean classes were distinguished by coupling these features with an appropriate classifier. The erroneous classification method for coffee beans achieved the lowest performance, as indicated by an accuracy value of less than 0.70 by implementing Set-6 features and the SVM classifier. In contrast, the optimal method performance was achieved by employing the features in Sets-4 and Set-5 to a decision tree classifier with an accuracy of 0.995. Nevertheless, Set-5 was deemed superior to Set-4 due to it containing fewer features. The method's performance indicates that selecting appropriate features and classifiers enhanced its performance. SVM, Naïve Bayes, ANN, C.45, and the decision tree generate the performance ranking from lowest to highest based on the applicable classifier. The performance of C.45 and the decision tree was competitive, in order applying all feature sets yielded an accuracy value greater than 0.950. In addition, the optimal performance of the method results in two misclassifications within the Robusta class, with one Robusta class image being misclassified as Excelsa and the other as Liberica. The details of the optimal classification results are presented in a confusion matrix, as shown in Figure 8. It depicts the inaccuracy in coffee bean images of the Robusta class. The misclassifications were found in two images of the Robusta class, which were classified as Excelsa and Liberica.

An automatic coffee bean classification method generated using a shape features approach with the implementation of the decision tree classifier. The proposed method's reliability was assessed on many coffee bean images obtained from four classes. The proposed method only uses four shape features, including aspect ratio, eccentricity, equivalent diameter, and area, obtained through feature selection by the gain ratio method. Each feature has a unique value, as shown in Table 1; therefore, those applied in the proposed method.

Table 3: The comparison of performance based on the combination of several feature sets and classifiers

Classifiers	Evaluation Parameters	Set-1	Set-2	Set-3	Set-4	Set-5	Set-6	Set-7
Naïve Bayes	Precision	0.710	0.716	0.709	0.666	0.669	0.700	0.714
	Recall	0.680	0.695	0.693	0.650	0.655	0.693	0.738
	Accuracy	0.680	0.695	0.693	0.650	0.655	0.693	0.738
ANN	Precision	0.990	0.993	0.980	0.872	0.855	0.714	0.701
	Recall	0.990	0.993	0.980	0.868	0.873	0.738	0.735
	Accuracy	0.990	0.993	0.980	0.868	0.873	0.738	0.735
SVM	Precision	0.723	0.724	0.723	0.694	0.694	0.695	0.606
	Recall	0.655	0.643	0.640	0.605	0.605	0.605	0.610
	Accuracy	0.655	0.643	0.640	0.605	0.605	0.605	0.610
C.45	Precision	0.987	0.987	0.987	0.988	0.988	0.953	0.953
	Recall	0.988	0.988	0.988	0.988	0.988	0.953	0.953
	Accuracy	0.988	0.988	0.988	0.988	0.988	0.953	0.953
Decision Tree	Precision	0.980	0.980	0.980	0.995	0.995	0.974	0.974
	Recall	0.980	0.980	0.980	0.995	0.995	0.973	0.973
	Accuracy	0.980	0.980	0.980	0.995	0.995	0.973	0.973



		Predicted Class			
		Arabica	*Excelsa*	*Liberica*	*Robusta*
Actual Class	*Arabica*	100	0	0	0
	Excelsa	0	100	0	0
	Liberica	0	0	100	0
	Robusta	0	1	1	98

Figure 8: Confusion matrix of classification result using Set-5 and decision tree.

In this paper, we compared the results obtained by the proposed method with the performance results associated with the coffee bean classification method for various purposes from several prior works, as summarized in Table 4. It provides an overview of the development studies related to the classification methods of coffee bean types. These previous studies utilized various feature extraction and classification techniques. No related works classify coffee beans with a similar class type.

The proposed method achieved an accuracy of 99.5%; moreover, it successfully surpassed the achievement result of the previous work (Sarino et al., 2019; Febriana et al., 2022; Arboleda et al., 2018) using smaller set of features. The proposed method's results demonstrate that the four shape features can be applied to classify the coffee bean. However, it is lower than (Septiarini et al., 2022a; Oliveira et al., 2016) which managed to achieve an accuracy value of 100% with a larger number of features and a smaller amount of data (150 images and 120 images). The proposed method has not been able to overcome the slight differences in shape between the Liberica and Excelsa. To improve the performance of the proposed method, it is necessary to combine the features extracted from coffee bean images with color or texture features.

Table 4: The summarized of previous works coffee classification

No.	Methods	Dataset	Classes	Accuracy
1.	Employing color features derived from the RGB color space in combination with a Support Vector Machine (SVM) using a polynomial kernel (Septiarini et al., 2022a).	150 images of arabica coffee bean	Roasting grades: light medium, and dark	100%
2.	The features were derived from the RGB color space and utilized with an ANN classifier (Sarino et al., 2019).	180 images of excelsa coffee bean	Roasting grades: light, medium, and very dark	97.22%
3.	Performing L*a*b color space and Bayesian classifier (Oliveira et al., 2016).	120 images of arabica coffee bean	Coffee bean color: whitish, cane green, green, and bluish green	100%
4.	Deep learning implemented using MobileNetV2 (Febriana et al., 2022).	8000 images of arabica coffee bean	Sorting green bean: peaberry, longberry, premium, and defect	81.31%
5.	Applying shape features (area, perimeter, equivalent diameter, roundness percentage) and ANN classifier (Arboleda et al., 2018).	255 images of coffee bean	Coffee bean species: robusta, excelsa, and liberica	96.66%
6.	Proposed method usin the gain ratio to select the shape features and decision tree classifier	400 images of coffee bean	Coffee bean species: arabica, excelsa, liberica, and robusta	99.5%

4 CONCLUSIONS

The experimental results in this work reveal that the shape features combined with machine learning can be used to classify the coffee beans, as indicated by the accuracy value of 0.995. In addition, the gain ratio produced the most discriminative shape features, including aspect ratio, eccentricity, equivalent diameter, and area, against the dataset incorporated with the decision tree classifier. The machine learning algorithm has been essential in producing optimal performance of the coffee bean classification method. This work has the potential for improvement and development to achieve optimum performance by investigating more discriminatory features.

5 AUTHORS' CONTRIBUTION

Conceptual idea: Septiarini, A.; Hamdani, H., Methodology design: Septiarini, A.; Hamdani, H., Data collection: Hidayat, N.; Afuan, L., Data analysis and interpretation: Septiarini, A.; Sela, E.I.; Hidayat, N., and Writing and editing: Septiarini, A.; Sela, E.I.; Afuan, L.

6 REFERENCES

AGUS, N. The impact of food safety standard on Indonesia's coffee exports. **Procedia Environmental Sciences**, 20:425-433, 2014.

ALCAYDE, M.; ELIJORDE, F. I.; BYUN, Y. Quality monitoring system for pork meat using computer vision. **2019 IEEE Transportation Electrification Conference and Expo, Asia-Pacific - ITEC Asia-Pacific**, p. 1-5, 2019.

ARBOLEDA, E. R.; FAJARDO, A. C.; MEDINA, R. P. Classification of coffee bean species using image processing, artificial neural network and K nearest neighbors. **2018 IEEE International Conference on Innovative Research and Development (ICIRD**), p. 1-5, 2018.

ARSALANE, A. et al. Artificial vision and embedded systems as alternative tools for evaluating beef meat freshness. **2020 IEEE 6th International Conference on Optimization and Applications - ICOA**, p. 1-6, 2020.

BAZAME, H. C. et al. Detection, classification, and mapping of coffee fruits during harvest with computer vision. **Computers and Electronics in Agriculture**, 183:106066, 2021.

BEHERA, S. K. et al. Maturity status classification of papaya fruits based on machine learning and transfer learning approach. **Information Processing in Agriculture**, 8(2):244-250, 2021.

DARWISH, A.; EZZAT, D.; HASSANIEN, A. E. An optimized model based on convolutional neural networks and orthogonal learning particle swarm optimization algorithm for plant diseases diagnosis. **Swarm and Evolutionary Computation**, 52:100616, 2020.

DENG, X. et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. **Information Sciences**, 340-341:250-261, 2016.

DUTTA, K.; TALUKDAR, D.; BORA, S. Segmentation of unhealthy leaves in cruciferous crops for early disease detection using vegetative indices and Otsu thresholding of aerial images. **Measurement**, 189:110478, 2022.

FAHMI, F. et al. Image processing analysis of geospatial uav orthophotos for palm oil plantation monitoring. **Journal of Physics**, 978:012064, 2018.

FEBRIANA, A. et al. USK-COFFEE Dataset: A Multi-class green arabica coffee bean dataset for deep learning. **2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)**, p. 469-473, 2022

FEBRIANTO, N. A.; ZHU, F. Coffee bean processing: Emerging methods and their effects on chemical, biological and sensory properties. **Food Chemistry**, 412:135489, 2023.

HAMDANI, H. et al. Detection of oil palm leaf disease based on color histogram and supervised classifier. **Optik**, 245:167753, 2021.

IQBAL, Z. et al. An automated detection and classification of citrus plant diseases using image processing techniques: A review. **Computers and Electronics in Agriculture**, 153:12-32, 2018.

JAMALLUDIN, M. D. et al. Implementation of feature selection using gain ratio towards improved accuracy of support vector machine (SVM) on youtube comment classification. **2021 International Seminar on Application for Technology of Information and Communication (iSemantic)**. p.22-31, 2021.

JANA, S.; PAREKH, R.; SARKAR, B. A De novo approach for automatic volume and mass estimation of fruits and vegetables. **Optik**, 200:163443, 2020.

JI, Y. et al. Non-destructive classification of defective potatoes based on hyperspectral imaging and support vector machine. **Infrared Physics & Technology**, 99:71-79, 2019.

LÜ, J. et al. Immature citrus fruit detection based on local binary pattern feature and hierarchical contour analysis. **Biosystems Engineering**, 171:78-90, 2018.

MUNERA, S. et al. Maturity monitoring of intact fruit and arils of pomegranate cv. ‘Mollar de Elche’ using machine vision and chemometrics. **Postharvest Biology and Technology**, 156:110936, 2019.

NYALALA, I. et al. Tomato volume and mass estimation using computer vision and machine learning algorithms: Cherry tomato model. **Journal of Food Engineering**, 263:288-298, 2019.

OLIVEIRA, E. M. de. et al. A computer vision system for coffee beans classification based on computational intelligence techniques. **Journal of Food Engineering**, 171:22-27, 2016.

PALACIOS-MORILLO, A. et al. Differentiation of Spanish paprika from protected designation of origin based on color measurements and pattern recognition. **Food Control**, 62:243-249, 2016.

PASHA, S. J.; MOHAMED, E. Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction. **Informatics in Medicine Unlocked**, 32:101064, 2022.

PATIL, J. K.; KUMAR, R. Analysis of content based image retrieval for plant leaf diseases using color, shape and texture features. **Engineering in Agriculture, Environment and Food**, 10(2):69-78, 2017.

PIEDAD, E. J. et al. Postharvest classification of banana (*Musa acuminata*) using tier-based machine learning. **Postharvest Biology and Technology**, 145:93-100, 2018.

PRASETIYO, B. et al. Evaluation of feature selection using information gain and gain ratio on bank marketing classification using naïve bayes. **Journal of Physics**, 1918(4):042153, 2021.

RAJASEKHAR, K.; BABU, T. R. G. Analysis and classification of dermoscopic images using spectral graph wavelet transform. **Periodica Polytechnica Electrical Engineering and Computer Science**, 64(3):313-323, 2020.

RANGKUTI, A. H.; HARJOKO, A.; PUTRA, A. E. A novel reliable approach for image batik classification that invariant with scale and rotation using MU2ECS-LBP algorithm. **Procedia Computer Science**, 179:863-870, 2021.

SANTOS, E. M. D. et al. Coffee by-products in topical formulations: A review. **Trends in Food Science and Technology**, 111:280-291, 2021.

SANTOS, F. F. L. D. et al. Quality assessment of coffee beans through computer vision and machine learning algorithms. **Coffee Science**, 15:e151752, 2020.

SARINO, J. N. C. et al. Classification of coffee bean degree of roast using image processing and neural network. **ResearchGate**, 8(10):3231-33, 2019.

SEPTIARINI, A. et al. Pixel quantification and color feature extraction on leaf images for oil palm disease identification. **2021 7th International Conference on Electrical, Electronics and Information Engineering - ICEEIE**. p. 1-5. 2021.

SEPTIARINI, A. et al. Multi-Class support vector machine for arabica coffee bean roasting grade classification. **2022 5th International Conference on Information and Communications Technology - ICOIACT**, p. 407-411, 2022a.

SEPTIARINI, A. et al. Image processing techniques for tomato segmentation applying k-means clustering and edge detection approach. **2021 International Seminar on Machine Learning, Optimization, and Data Science – ISMODE**, p. 92-96, 2022b.

SEPTIARINI, A. et al. Maturity grading of oil palm fresh fruit bunches based on a machine learning approach. **2020 Fifth International Conference on Informatics and Computing - ICIC**, p. 1-5, 2020.

SHARIF, M. et al. Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection. **Computers and Electronics in Agriculture**, 150:220-234, 2018.

TAN, K. et al. Recognising blueberry fruit of different maturity using histogram oriented gradients and colour features in outdoor scenes. **Biosystems Engineering**, 176:59-72, 2018.

TRABELSI, M.; MEDDOURI, N.; MADDOURI, M. A New feature selection method for nominal classifier based on formal concept analysis. **Procedia Computer Science**, 112:186-194, 2017.

UTAMININGRUM, F. et al. Descending stairs and floors classification as control reference in autonomous smart wheelchair. **Journal of King Saud University - Computer and Information Sciences**, 34(8):6040-6047, 2022.

VELÁSQUEZ, S. et al. Classification of the maturity stage of coffee cherries using comparative feature and machine learning. **Coffee Science**, 16:e161710, 2021.

YANG, X. et al. Machine learning for cultivar classification of apricots (*Prunus armeniaca* L.) based on shape features. **Scientia Horticulturae**, 256:108524, 2019.

ZHANG, C.; LIU, F.; HE, Y. Identification of coffee bean varieties using hyperspectral imaging: influence of preprocessing methods and pixel-wise spectra analysis. **Scientific Reports**, 8: 2166, 2018.

ZHANG, L.; WANG, J.; DUAN, Q. Estimation for fish mass using image analysis and neural network. **Computers and Electronics in Agriculture**, 173:105439, 2020.

ZHENG, Y. et al. Adaptive neural decision tree for EEG based emotion recognition. **Information Sciences**, 643:119160, 2023.

ZHUANG, J. et al. Detection of orchard citrus fruits using a monocular machine vision-based method for automatic fruit picking applications. **Computers and Electronics in Agriculture**, 152:64-73, 2018.

ZOU, K. et al. Broccoli seedling pest damage degree evaluation based on machine learning combined with color and shape features. **Information Processing in Agriculture**, 8(4):505-514, 2021.