Vol. 15, No. 3 (2026): 895 - 906 http://dx.doi.org/10.23960/jtepl.v15i3.895-906

JURNAL TEKNIK PERTANIAN LAMPUNG

ISSN 2302-559X (print) / 2549-0818 (online)
Journal homepage : https://jurnal.fp.unila.ac.id/index.php/JTP

Grading Coffee Beans using Extraction of Shape-Based Features Coupled with Support Vector Machine

Agus Dharmawan[1,✉], Rudiati Evi Masithoh[2], Siswoyo Soekarno[1], Hanim Zuhrotul Amanah[2]

[1] Department of Agricultural Engineering, Faculty of Agricultural Technology, University of Jember, INDONESIA.

[2] Department of Agricultural and Biosystems Engineering, Faculty of Agricultural Technology, Gadjah Mada University, INDONESIA.

Article History:

Received : 07 July 2025
Revised : 29 July 2025
Accepted : 28 August 2025

Keywords:

Coffee bean,
Grading,
Shape-based feature extraction,
SVM.

Corresponding Author:
✉ agusd@unej.ac.id
(Agus Dharmawan)

ABSTRACT

Evaluating coffee beans through a computer vision system (CVs) requires a large number of visual attributes to be extracted, but may affect prediction accuracy. Therefore, it is essential to reduce the large features to gain better prediction accuracy by generating new data that represents the most informative dimensions of the original data. Previous studies are limited to comparing different methods of feature extraction. The objective of this research was to explore the comparison of six feature extraction methods (PCA, EFA, LDA, SVD, ICA, and PLS) combined with support vector machine (SVM) as a supervised approach to predict three groups of coffee beans, namely long-berry, normal, and peaberry, for grading issues. SVM with three kernel functions (linear, RBF, and sigmoid) was used to construct a superior classification model. Data were acquired from coffee images processed to generate shape-based features. The results show that LDA provides a better visualization in separating sample classes according to the score plot with 2 variables obtained. The combination of SVM and LDA has a better recognition of coffee beans for grading, which is higher than that of other combinations. A combination of SVM-sigmoid with EFA gave mostly the worst recognition. Our findings proved that the investigation of feature extraction methods and SVM successfully achieve accurate results on grading coffee beans.

1. INTRODUCTION

Coffee grading categorizes coffee beans on the basis of various criteria before being introduced into the market (Wallelign, 2020). Each coffee producer has developed its grading criteria to facilitate a fair system of market pricing and to set minimum standards for export (International Coffee Organization, 2018). Physical attributes of coffee beans, such as appearance, size, shape, thickness, weight, and color, play an important role in determining their quality (Koklu & Ozkan, 2020; Tran *et al*., 2017). The manual assessment of coffee is performed by trained experts who are qualified to describe its visual characteristics. It uses human vision, which may lead to several disadvantages, including inconsistent results and subjectivity due to the effects of the expert's mood, perception, and fatigue (Meenu *et al*., 2021). Nowadays, at the industrial level, coffee producers have utilized a grading machine to confirm the size uniformity of coffee beans, as documented by Widyotomo & Mulato (2005). This machine has table graders with different hole sizes and is installed at a particular angle, approximately 10°. The vibration allows the beans to flow down and pass through a hole of a certain size.

Recently, computer vision (CV) and machine learning (ML)-based techniques have been frequently employed in a broad range of applications, including the recognition of agro-food products (Meenu *et al*., 2021). The CV makes it possible to produce standard images, process them using image processing algorithms, extract their features, and identify them according to the descriptors generated (Lin *et al*., 2021; Bedaso *et al*., 2022; Meenu *et al*., 2021). The

identification of food items based on their physical attributes can be done by employing either statistical methods or ML techniques. ML techniques include Neuro-Fuzzy (Pazoki *et al*., 2014), Support Vector Machine (Koklu & Ozkan, 2020; Lopes *et al*., 2019), Neural Networks (Koklu & Ozkan, 2020; Turi *et al*., 2013), Decision Trees (Koklu & Ozkan, 2020; Lopes *et al*., 2019), k-Nearest Neighbors (Koklu & Ozkan, 2020; Lopes *et al*., 2019), and Random Forest Classifier (Oliveira *et al*., 2021).

Support vector machine (SVM), which was originally proposed by Vapnik and Chervonenkis in 1995, has been successfully developed to solve classification and calibration problems (Brereton & Lloyd, 2010). The method is categorized as supervised non-parametric statistical learning, which is presented with a set of predictive data and labeled targets. SVM aims to find a hyperplane that separates the dataset into a discrete, predefined number of classes (Mountrakis *et al*., 2011). Three critical hyperparameters in SVM include Kernel function, gamma, and *C*. Many kernel functions can be used, such as linear, normalized polynomial, RBF, sigmoid, GaussianRBF, etc. (Bhavsar & Panchal, 2012). To find the best algorithm, the optimal values of parameters gamma and *C* are selected by performing a grid search from several possible values using training data (Wallelign, 2020). These values are then used to train and test the dataset. However, before the original data is introduced to SVM, a hybrid variable reduction step is required (Brereton & Lloyd, 2010).

The number and type of food items (or features) affect an appropriate model to be selected, as well as its prediction accuracy and processing time (Meenu *et al*., 2021). Feature data may contain high dimensionality, irrelevant, redundant, noisy, and highly correlated information, which leads to the risk of overfitting or degradation of prediction accuracy (Howley *et al*., 2006). Feature reduction is crucial for data with high dimensionality or that has many variables (or multivariable). The two most common approaches used include feature selection and feature extraction. Feature selection tries to choose features that contain the information required to distinguish between classes and enhance prediction performance (Kumar & Bhatia, 2014). Feature extraction aimed to seek new data with lower dimensionality by transforming the original data with high dimensionality. Unlike feature extraction, feature selection, which uses a small number of variables, may eliminate some useful information in the data (Mehmood *et al*., 2012).

Feature extraction approach can be grouped into two categories: linear and non-linear methods. PCA, one of the most common linear methods, uses a classical statistical approach to transform correlated variables of a dataset into a new set of uncorrelated variables (Howley *et al*., 2006). Other linear methods have been recorded in several studies, including exploratory factor analysis (EFA), linear discriminant analysis (LDA) (Soleimanipour *et al*., 2018; Jayaprakash *et al*., 2020), singular value decomposition (SVD) (Adebayo & Olumide, 2016; Husin *et al*., 2012), independent component analysis (ICA) (Jayaprakash *et al*., 2020), and partial least-squares regression (PLSR) (Fordellone *et al*., 2018).

We only found limited studies comparing the utilization of various linear feature extraction methods as an alternative to reduce the dimension of feature data to improve the effectiveness and efficiency of a classification model. This study used SVM, a tool to solve the grading of coffee beans, with the experiment of different kernel functions. Images of coffee beans, which were collected from an RGB camera, were processed to extract their morphological properties. The samples were from Arabica coffee and were divided into three groups: long-berry, premium, and peaberry. Therefore, the objective of this study was to develop an SVM-based classifier in combination with feature extraction methods in order to recognize three classes of coffee sizes and to draw a comparison of which combinations achieve better predictive accuracy among the sample groups.

The main contributions of this paper are as follows: firstly, this research provides a clear comparison of six linear feature extraction methods, namely PCA, EFA, LDA, SVD, ICA, and PLSR, using shape-based features of Arabica beans to decompose its original data from high-dimension to lower-dimension and to visualize samples' clustering. Secondly, the dataset with reduced variables is introduced to the SVM model with three kernel functions (linear, RBF, sigmoid) to predict three classes of coffee samples. Thirdly, the experiments using six linear feature extraction methods and three SVM kernels are analyzed to draw a comparison to select which best combination that gives excellent accuracy for grading coffee beans. We hope the approach obtained in this study could be a promising alternative for routine industrial applications in grading and quality control of coffee samples.

2. MATERIALS AND METHODS

2.1. Data Collection

Coffee samples were Arabica beans purchased from a trusted local trader in Temanggung, Central Java, Indonesia, and were from full-washed coffee processing. Before image acquisition, fine beans were separated from damaged beans and were cleaned manually to remove dirt and the remaining dried-attached silverskin. The image acquisition system consisted of four main components: (1) a digital camera (Fujifilm X-A3, 24MP - APS-C CMOS Sensor) to take light information from samples and convert them into digital images, (2) two LED light sources (length 250 mm) to illuminate samples, (3) a black box (220×170×130 mm) with bad-light reflection matter, and (4) a computer-software system to store, display, and process digitized images. The lens of the camera was positioned perpendicularly 130 mm above a sample base. The lights were at an angle of approximately 450 from the center of the sample base. The camera was set in a close-up mode, autofocus-on, and zoom-off. The images were taken on a black background and stored in JPEG format. A desktop computer was an Intel(R) Core (TM) i5, CPU 8.00 GB (RAM), 512 GB SSD capacity with Microsoft Windows 11 Home Single Language, and 64 Operating system, equipped with Visual Studio Code as a Python IDE.

All acquired images (Figure 1, left) were resized from a 6000×4000-pixel size to 2250×1500 pixels. The images were then pre-processed to obtain region-of-interests (ROIs). The ROIs, the largest square portion of a single bean, were cropped from the raw images, see Figure 1 (right). A filter method was applied to eliminate noise appearing on the images. This filter used a 2D Gaussian smoothing 5×5 kernel. Using an Otsu thresholding method, the segmentation stage was applied to remove the background and objects not part of the coffee image. A total of 1010 beans were collected and distributed into longberry (251), premium (415), and peaberry (344).

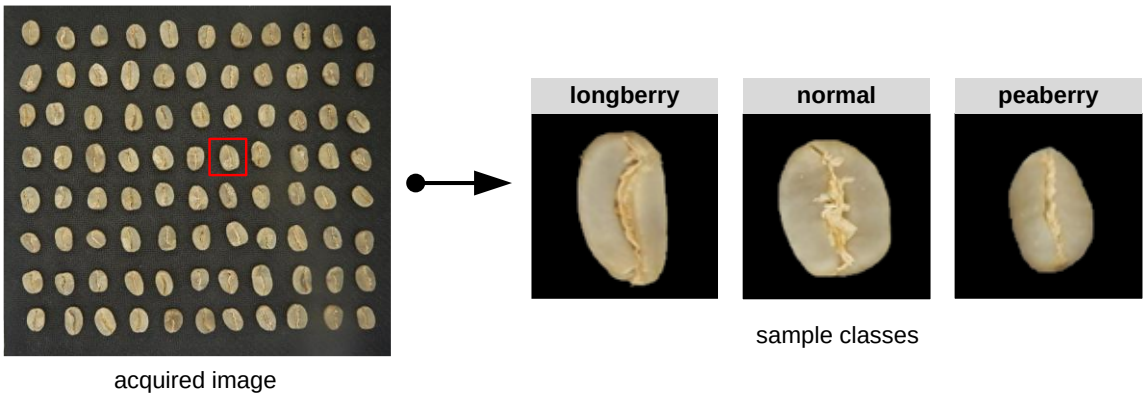

Figure 1. Original and cropped images

A further analysis is to extract quantitatively feature information from images, specifically physical dimensions that characterize their appearance, so that they may be unambiguously classified. To do so, all images were converted into binary images and transformed into two colors: the object should be white, and the background should be black. Image contouring then identifies an object's structural lines by combining all continuous points along the boundary to find its contour. A contour typically refers to edge pixels with a similar color intensity. Edge detection is a required step to determine which pixels should be considered edge pixels using one of the popular edge detection techniques, the Canny operation. To differentiate the three groups of coffee samples, namely long-berry, premium, and peaberry, features regarding the shape of coffee are extracted into 19 relative values, listed in Table 1.

2.2. Feature Extraction

The shape features contain data that have different ranges. It is necessary to transform the values of features in the data to a similar and consistent scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values. *Z*-score normalization (or standardization) is one of the most common techniques for feature scaling. All values in feature data are centered around their mean per unit of standard deviation.

Table 1. Shape descriptors of coffee bean

Group	Feature name	Description	Equation
Contour-based	Perimeter	The number of pixels that perform the distance/ arch length (or) around the edge of an object.	p
	Area	The number of pixels taken up by the surface area of an object.	A
	Convex area	The number of pixels in the smallest convex polygon that can contain the area of an object.	A_C
	Major axis length	The pixel distance between the endpoints of the longest line that can be drawn through the object.	L
	Minor axis length	The pixel distance between the endpoints of the longest line that can be drawn through the object while standing perpendicular to the major axis.	l
	Bounding box area (A_B)	The number of pixels in a bounding box or rectangle drawn on the outer edge of an object.	$L \times l$
	Equivalent diameter	The diameter of the circle where the area is the same as the contour of an object.	$\sqrt{\frac{4A}{\pi}}$
Region-based	Compactness	The ratio of the surface area of an object to the area of a circle whose diameter is equal to the maximum diameter of the object. A circle is used as it is the most compact shape of an object. The measure takes a maximum value of 1 for a circle.	$\frac{4\pi A}{p^2}$
	Roundness (or circularity)	The ratio of the area of an object to the area of a circle with the same convex perimeter. A circular object equals 1 and less than 1 for an object that departs from circularity.	$\frac{4\pi A}{p_C{}^2}$
	Complexity	The ratio between the square perimeter to the unit area of an object.	p^2 / A
	Aspect ratio (eccentricity or ellipticity)	The ratio of width to height of the bounding rectangle of the object. It is particularly useful when talking about ellipses. The result is given as a value from 0 to 1.	$\frac{l}{L}$
	Elongation	The ratio between the length and width of an object in the bounding box. A roughly square or circularly shaped object is equal to 1. The more elongated-shaped object, the more decreased from 1.	$\frac{l_B}{L_B}$
	Extent (or rectangularity)	The ratio of the contour area to the bounding rectangle area. It is also defined as the proportion of pixels in the bounding box that was also in the object region. A perfectly rectangular object has a value of 1.	$\frac{A}{A_B}$
	Solidity	The ratio of the contour area to its convex hull area, or is mentioned as the proportion of pixels in the object region that was also in the convex hull. A value of 1 signifies a solid object. A value less than 1 represents an irregular boundary object.	$\frac{A}{A_C}$
	Convexity	Measures the relative amount of an object that differs from a convex object. It computes the ratio of the perimeter of an object's convex hull to the perimeter of the object itself. A convex object has a value of 1. An object with an irregular boundary will be less than 1.	$\frac{p_C}{p}$
Shape factors		$SF_1 = \frac{L}{A}$; $SF_2 = \frac{l}{A}$; $SF_3 = \frac{A}{\frac{L}{2}\cdot\frac{L}{2}\cdot\pi}$; and $SF_4 = \frac{A}{\frac{L}{2}\cdot ½ \cdot\pi}$	

Source: (Dawson-Howe, 2014; Wirth, 2004; Koklu *et al*., 2021; Zhang *et al*., 2017).

In this step, six techniques, namely PCA, EFA, LDA, SVD, ICA, and PLS, were used for feature reduction and data visualization. They transformed the original data with a large feature into a smaller feature. PCA formed new variables, known as principal components (PCs), which represent the most common variations of the original data. Practically, the PCs that explained more than 85% of the cumulative variance were used to replace the original variables. EFA also transformed original variables into a smaller set of linear combinations called factors. Before applying EFA, the data suitability was assessed using the magnitude of communalities to check the acceptability of the sample size. Bartlett's test of sphericity and KMO measure of sampling adequacy were also used to evaluate the strength of the relationship among the original variables. The number of factors was determined based on the decomposed eigenvalues above 1.00. A low-dimensional data was projected from high-dimensional data using LDA.

The new data formed by LDA had two variables since it could not be larger than the number of original variables and target classes. ICA obtained the projected data based on the top independent components. To transform the dataset, PLS used the feature data and target classes in the algorithm. Among all these techniques, the sorted

eigenvalues that represented the proportion of variance were also provided in the result analyses. For better comprehension, in this study, we also use PCA to map the relationships between samples and between variables. By plotting the first two features of the transformed data, the patterns of how samples relate to each other could be visualized. The PCA biplot (or loading plot) was also used to depict the relationship between variables that significantly contribute to the projected features.

2.3. SVM Classifier

This study used SVM to solve classification problems regarding the inspection of the three types of Arabica coffee samples. The model inputs were the projected data obtained after the analyses of feature extraction. The three kernel functions -linear, RBF, and sigmoid- were applied in the SVM algorithm. Each kernel function has its tunable parameters, such as C in a 'linear' function, and γ and C in 'RBF' and 'sigmoid' functions. Hyperparameter optimization is applied to find the best combination of these hyperparameters using *Scikit-Learn's* GridSearchCV. The model classified the images of coffee beans into three classes: class 0 for a group of longberry beans, class 1 for premium beans, and class 2 for peaberry beans.

2.4. Model Evaluation

K-fold cross-validation is used to assess the ability of the SVM model to predict new data. It uses a training dataset partitioned by 70% of the total dataset, and the remaining 30% of the total dataset is stored to test the model. The term K refers to the number of folds in the training data that will be split into. One-fold is used for validation, and the remaining K–1 folds are used to train the model. This process is carried out in K iterations, where K is determined as 10. The effectiveness of the model was evaluated in terms of the average classification accuracy and error estimation generated during training.

A confusion matrix is a specific table layout used to visualize the performance of the classifier by comparing the predicted targets with the actual targets of both training and testing datasets (Géron, 2019; Ahad *et al*., 2023). The predicted target represented the labels (or class value) generated by the model, while the actual targets represented the initial labels of the original data (Saputra & Kristiyanti, 2022). This confusion matrix is given in Figure 2, containing the number of true and false classified samples, which are then computed to determine the values of true positive (*TP*), true negative (*TN*), false negative (*FN*), and false positive (*FP*), as shown in Table 2. Although the confusion matrix gives a lot of information, it is more convenient to express it in concise metrics, such as accuracy, precision, recall, and F1 score. The formulas are given in Equations (1-4). The results are given as a value from 0 to 1. The more excellent the prediction model, the closer the result to 1.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

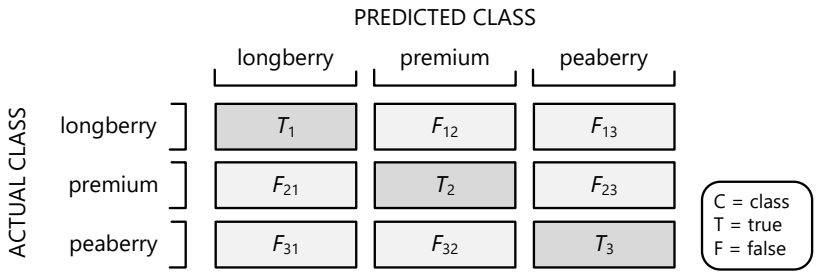

Figure 2. Scheme of three-class confusion matrix

Table 2. Determination of *TP*, *TN*, *FP*, and *FN*

Classes	*TP*	*TN*	*FP*	*FN*
Longberry	T_1	$T_2 + T_3 + F_{23} + F_{32}$	$F_{21} + F_{31}$	$F_{12} + F_{13}$
Premium	T_2	$T_1 + T_3 + F_{13} + F_{31}$	$F_{12} + F_{32}$	$F_{21} + F_{23}$
Peaberry	T_3	$T_1 + T_2 + F_{12} + F_{21}$	$F_{13} + F_{23}$	$F_{31} + F_{32}$

$$\text{precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 \text{ score} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}} \tag{4}$$

3. RESULTS AND DISCUSSION

3.1. Physical Characteristics of Arabica Coffee

The physical attributes of coffee beans play an important role in identifying their profile in the grading system. Table 3 proved that the three types of Arabica bean shapes were considerably different according to their size and geometry. The long-berry beans are larger in shape than premium and peaberry beans. Hence, they have higher values in terms of area, perimeter, axis lengths, bounding area, convex area, and equivalent diameter. Based on their physical features, the compactness values of the premium beans are greater than the long-berry and peaberry beans, showing that they are more oval or not perfectly circular. The higher values of roundness for the premium beans indicate that they are rounder than the other two beans.

Table 3. Morphological characteristics of longberry, premium, and peaberry

Name of Features	Longberry		Premium		Peaberry	
	mean	SD	mean	SD	mean	SD
Area	8059.50	837.18	7027.51	789.02	5606.86	782.84
Perimeter	352.48	19.04	321.30	18.20	290.89	22.03
Minor axis length	129.63	8.33	109.68	7.27	103.90	9.53
Major axis length	79.67	5.52	82.08	5.89	69.03	5.23
Bounding box area	10332.67	1120.19	9018.09	1033.13	7229.75	1045.81
Convex area	8181.61	848.98	7128.28	799.31	5699.34	799.626
Equivalent diameter	101.16	5.25	94.45	5.27	84.29	5.803
Compactness	0.81	0.027	0.85	0.018	0.83	0.028
Roundness	0.89	0.026	0.94	0.016	0.91	0.025
Complexity	15.46	0.538	14.47	0.326	15.16	0.532
Aspect ratio	0.62	0.051	0.75	0.062	0.67	0.062
Elongation	0.64	0.054	0.77	0.064	0.70	0.068
Extent	0.78	0.028	0.78	0.021	0.78	0.028
Convexity	0.95	0.007	0.95	0.006	0.95	0.008
Solidity	0.99	0.003	0.99	0.003	0.98	0.004
SF_1	0.01	0.001	0.01	0.001	0.01	0.001
SF_2	0.02	0.001	0.02	0.001	0.02	0.001
SF_3	1.62	0.135	1.33	0.109	1.50	0.141
SF_4	0.99	0.005	0.99	0.004	0.99	0.005

According to the aspect ratio, the premium beans were characterized as more eccentric or more ellipse in shape. Elongation and extent measure the suitability of an object, respectively, to the square and rectangular shapes. The higher value obtained from premium beans indicates that the shape is relatively close to a square or circle. The same values obtained from extent, 0.78, indicate that the beans have a similar geometry, not perfectly rectangular in shape. The solid shape was also given by all three beans. The beans do not represent an irregular boundary object because they are relatively close to 1.00. The complexity feature of long-berry and peaberry beans tends to have the same values. The more complex the shape of an object, the higher it is. Finally, the less significant values are also given from SF1, SF2, SF3, and SF4.

Table 4. The number of variables vs extracted variance

Variable	PCA		EFA		LDA		SVD		ICA		PLS	
no.	EV	CV	EV	CV	EV	CV	EV	CV	EV	CV	EV	CV
1	8.23	43.30	8.22	43.27	0.62	62.00	91.16	43.30	-	-	8.01	48.51
2	6.34	76.66	6.33	76.58	0.38	100.0	80.00	76.70	-	-	4.25	74.27
3	2.39	90.23	2.38	89.11			49.11	89.30	-	-	2.97	92.25

Note: EV = explained variance, CV = cumulative variance (in %).

3.2. Determining the number of extracted variables

The number of reduced features for those techniques is first discussed according to their extracted variance, as given in Table 4. The number of PCs was obtained from the first features that achieved >80% of the explained variance. Therefore, we selected three PCs because they contributed to the cumulative variance of 90.23% (PC1: 43.30%, PC2: 33.36%, and PC3: 13.57%). The new features obtained from LDA only have a dimension of 2 (two) since it could not be larger than the minimum number of input variables and classes. SVD and PLS also generate three variables that account for 89.30% and 92.25% of the total variance explained, respectively.

In EFA, we also determined three factors that account for 89.11% of the total variance explained. However, EFA requires pre-analysis to ensure whether the relationships within the data exist using Bartlett's test of sphericity and the Kaiser–Meyer–Olkin (KMO) Measure of Sampling Adequacy. In this study, the KMO value obtained was 0.695 or larger than 0.5, indicating that our data is fair enough for factor analysis to commence. The result of Bartlett's test of Sphericity had a significant statistical test of less than 0.05, showing that the input variables relate to one another to perform EFA. Different from other feature extraction methods, which can determine the number of new variables from extracted variance, ICA only searches for new features that are non-Gaussian and statistically independent. We directly select three features from the top independent components, so the new variables obtained from ICA have the same dimension as the other techniques.

The visualization of sample clustering based on PCA is also presented. From Figure 3, we can see that the two-dimensional PCA score plot clearly displays grouping based on their morphological data. The long-berry beans are spatially grouped along the positive score of PC1. The normal beans are positioned in the positive score of PC2. The peaberry beans are separated from the other two groups because they are placed in the positive score of PC1 and the negative score of PC2. If we see data points overlapping with other samples, this indicates that the samples have similar shape characteristics.

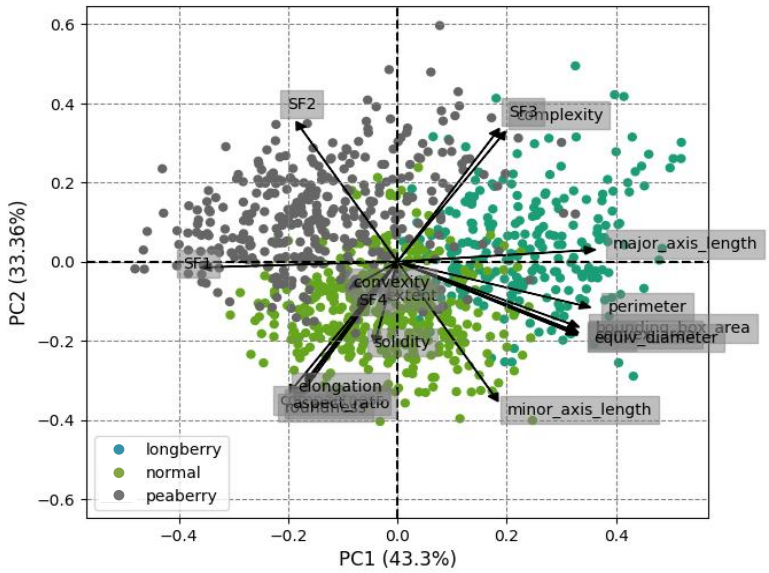

Figure 3. The 2D-PCA biplot

The PCA loading is also obtained to evaluate which shape features contribute the most to each variance (or principle component) (de Almeida *et al*., 2021). According to Figure 3, certain features indicated a significant contribution to PC1 with their positive values, including area, perimeter, axis lengths, convex area, bounding box area, equivalent diameter, complexity, and SF3. The negative values are provided by elongation, compactness, aspect ratio, roundness, F1, and SF2. A weak contribution was given by solidity, convexity, extent, and SF4. Hence, the loadings of these features resulted in a good separation between samples (long-berry, normal, peaberry). The clear separation between normal and peaberry beans can be visualized according to the PC2 axis. The variables with strong influences were provided by complexity, axis lengths, SF2, SF3, elongation, compactness, aspect ratio, and roundness.

3.3. Validation and Prediction using SVM Classifier

In this section, the performance of two SVM algorithms is presented. Before running the algorithms, we performed hyperparameter tuning to figure out the best combinations of SVM parameters from each Kernel function. There are two major SVC parameters required to be tuned, including regularisation (C) and Gamma (γ). The C parameter was tuned from (0.1, 1, 10, 100, 1000) while γ was tuned from (0.001, 0.01, 0.1, 1). Table 5 displays the optimal parameters found using Grid Search. We then take these results to train our SVM model, make predictions using the testing set, and perform classification.

Table 5. Results of tuning the hyperparameters of the SVM classifier

Kernel functions	Parameters	Feature extraction methods					
		PCA	EFA	LDA	SVD	ICA	PLS
Linear	C	0.1	1.0	10	100	100	0.1
RBF	C	1	100	1000	1000	1000	10
	gamma	0.1	0.001	0.001	0.1	0.1	0.01
Sigmoid	C	100	10	1000	100	100	100
	gamma	0.001	0.001	0.01	1	1	0.01

Table 6. Results of the evaluation metrics

SVM Kernels	Feature Extraction Methods	Validation	Training				Testing			
		avg. accuracy ± std. dev.	Ac (%)	Pr (%)	Re (%)	Fs	Ac (%)	Pr (%)	Re (%)	Fs
Linear	PCA	69.0 ± 5.0	89.6	84.7	84.3	0.845	92.3	89.0	88.7	0.888
	EFA	71.0 ± 6.0	88.6	83.0	83.6	0.830	89.0	85.5	84.2	0.842
	LDA	81.0 ± 5.0	91.2	86.9	86.7	0.868	92.5	89.9	88.1	0.889
	SVD	71.0 ± 8.0	90.8	86.6	85.8	0.862	90.8	85.9	86.2	0.859
	ICA	71.0 ± 3.0	90.8	86.8	86.0	0.863	91.0	86.9	85.9	0.863
	PLS	71.0 ± 6.0	90.5	86.0	85.4	0.857	89.7	85.7	84.3	0.848
RBF	PCA	83.0 ± 4.0	91.1	87.3	86.7	0.869	90.5	86.0	85.1	0.853
	EFA	72.0 ± 5.0	90.2	85.3	84.9	0.851	90.3	86.3	85.1	0.856
	LDA	81.0 ± 2.0	90.8	86.4	86.1	0.862	94.3	91.0	91.7	0.913
	SVD	73.0 ± 5.0	91.4	87.6	86.9	0.872	89.2	84.0	83.6	0.838
	ICA	72.0 ± 4.0	90.7	86.7	85.9	0.862	90.5	86.3	85.5	0.857
	PLS	77.0 ± 6.0	90.2	85.5	85.1	0.853	91.2	87.0	87.2	0.871
Sigmoid	PCA	69.9 ± 6.2	90.7	86.3	85.8	0.860	90.8	86.8	86.0	0.862
	EFA	47.1 ± 5.5	60.4	28.3	34.0	0.206	62.4	14.5	33.3	0.202
	LDA	83.8 ± 3.2	92.3	88.5	88.4	0.885	90.3	86.2	85.0	0.854
	SVD	71.9 ± 5.6	90.9	86.6	86.1	0.863	91.0	86.8	87.0	0.868
	ICA	70.7 ± 5.4	90.2	85.8	85.0	0.853	91.0	87.3	86.1	0.864
	PLS	72.1 ± 4.6	91.1	86.8	86.8	0.868	87.0	80.7	80.8	0.807

Ac = accuracy, Pr = precision, Re = recall, and Fs = F1 score.

Table 6 represents the averaged accuracy and error estimation obtained during the validation processes. These results can be utilized to compare combinations of different feature extraction methods and SVM models and determine the one combination that performs best. Considering these results, combinations of LDA–SVMs (linear,

RBF, sigmoid) and PCA–SVM (RBF) provide the best performance with an accuracy larger than 80%. The table also suggests that combinations of RBF and six feature extraction methods achieved high performance or had higher accuracies.

Table 6 also shows the accuracy, recall, precision, and F1-score results from three SVM Kernels with six feature extraction methods. High values of accuracy, recall, precision, and F1-score represent a better model. The experimental results demonstrate that the combination of EFA and Sigmoid had the lowest performance in discriminating sample classes with 62.4% accuracy, 14.5% precision, 33.3% recall, and 0.202 F1-score generated from testing data. The best predictions were obtained by incorporating LDA with three SVM kernels, reaching larger than 90% of accuracy values. However, using the test data, the combination of LDA and RBF-SVM achieved its best recognition with an accuracy of 94.3%; the resulting confusion matrix is shown in Figure 4.

Even though combining SVM and feature extraction methods using image data has been recorded in a few studies (Zeeshan *et al*., 2020; Zhang & Wu, 2012), most of which only used PCA, the others are not well documented. In this experiment, three kernels were chosen, namely linear, sigmoid, and RBF. The RBF achieved its best classification accuracy because it obtained proper separation between sample classes. According to Figure 5 (upper), unlike linear and sigmoid kernels, the PCA-RBF projected properly non-linear separable data into a higher-dimensional space so its hyperplane can visually separate sample classes.

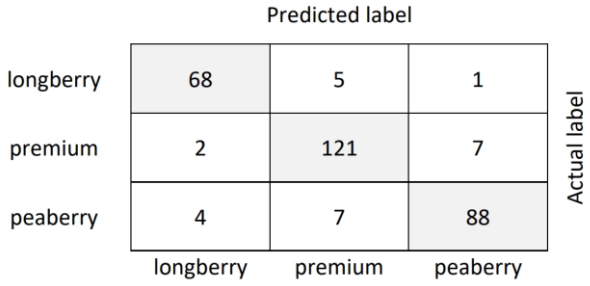

Figure 4. Confusion matrix of LDA and RBF-SVM

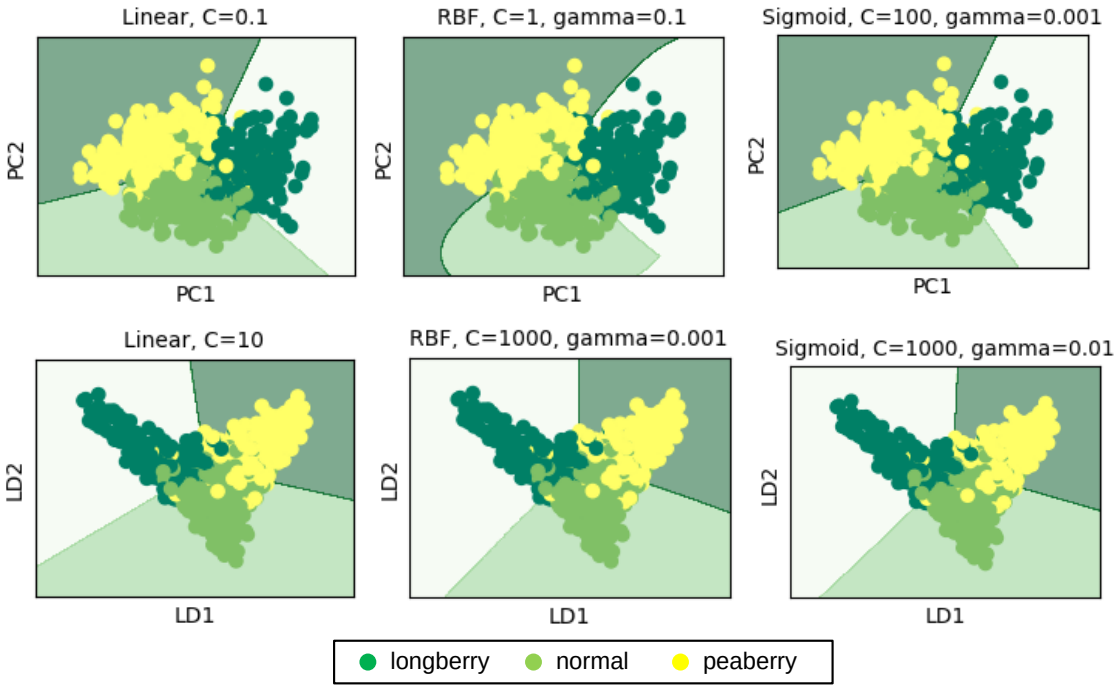

Fig 5. Plots of different SVM Kernels using data extracted by PCA (upper) and LDA (lower)

Few studies by Al-Mejibli *et al*. (2020) and Hastie *et al*. (2004) have demonstrated that the accuracy of Kernels is influenced by the changing of the regularization cost (*C*) and gamma (γ) parameters. The role of these two parameters is to draw a classification margin between the support vectors (high-dimensional data after being processed by SVM) and the separating hyperplane (or decision boundary). Considering Figure 5, the different *C* and gamma values influenced the hyperplanes to be drawn. The parameter *C* estimates the margin size. Its large value gives a smaller margin. Conversely, a smaller value of *C* gives a larger margin. The larger values of *C* give the plotted data points near the decision boundary (Fig. 4, LDA-RBF), while the smaller values involve data further away (Fig. 5, PCA-RBF). In other circumstances, the model with low values of *C* is more regularized than that with higher values (Hastie *et al*., 2004).

The parameter gamma is commonly used for polynomial, RBF, and sigmoid kernels (Al-Mejibli *et al*., 2020). The parameter gamma separates classes by drawing a decision boundary. Higher gamma values lead to a wiggly decision boundary; the decision boundary is closer to the data points. While a low gamma makes a simpler decision boundary, the farther points to the decision boundary. Fig. 5 depicts, at the PCA-RBF plot, that the decision boundary does not bend to the whims of data points. Conversely, at the LDA-RBF plot, the decision boundary bends and twists to classify data points correctly. However, most of our models, feature extraction methods combined with SVM, generated excellent classification because they used combinations of the optimal values *C* and gamma (Table 5) estimated by *k*-fold cross-validation and grid search technique in tuning hyperparameters.

4. CONCLUSION AND FUTURE WORKS

The methodological approach presented in this study could be a powerful tool to maximize the applicability of machine learning models for data analysis applied to agro-food detection. This work proposed a novel classification method based on shape-based image features combined with six linear feature extraction methods - PCA, EFA, LDA, SVD, ICA, and PLS - and a support vector machine with three kernel functions -linear, RBF, sigmoid- for a classification purpose, particularly grading of Arabica beans. The best prediction was obtained while the RBF-SVM was combined with LDA, achieving a classification accuracy of 94.3% using testing data. Bad recognition was only given from a combination of EFA and Sigmoid, gaining an accuracy of 62.4 on testing data. Our proposed procedure has the potential to migrate the mechanical-based grading system to an RGB camera mounted on a conveyor-based coffee grading system. However, future studies should aim at (a) extending our research to assess other agro-food products, and (b) utilizing other machine learning models to compare and increase classification accuracy.

AUTHOR CONTRIBUTION STATEMENT

Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
AD	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓
REM	✓	✓		✓	✓		✓			✓		✓		✓
HZA	✓	✓		✓	✓					✓				
SS										✓				

C: Conceptualization
M: Methodology
So: Software
Va: Validation
Fo: Formal Analysis
I: Investigation
D: Data Curation
R: Resources
O: Writing - Original Draft
E: Writing - Review & Editing
Vi: Visualization
Su: Supervision
Fu: Funding Acquisition
P: Project Administration

REFERENCES

Adebayo, D.S., & Olumide, O. (2016). Fish classification algorithm using single value decomposition. *International Journal of Innovative Research in Science, Engineering and Technology*, ***5***(2), 1621–1628.

Ahad, M.T., Li, Y., Song, B., & Bhuiyan, T. (2023). Comparison of CNN-based deep learning architectures for rice diseases classification. *Artificial Intelligence in Agriculture*, ***9***, 22–35. https://doi.org/10.1016/j.aiia.2023.07.001

Al-Mejibli, I.S., Alwan, J.K., & Abd, D.H. (2020). The effect of gamma value on support vector machine performance with different kernels. *International Journal of Electrical and Computer Engineering*, ***10***(5), 5497–5506. https://doi.org/10.11591/IJECE.V10I5.PP5497-5506

Bedaso, M., Meshesha, M., & Diriba, C. (2022). Grading ethiopian coffee raw quality using image processing techniques. *Research Square,* ***2022***(1). https://doi.org/10.21203/rs.3.rs-1980632/v1

Bhavsar, H.P., & Panchal, M. (2012). A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology*, ***1***(10), 2278–1323.

Brereton, R.G., & Lloyd, G.R. (2010). Support vector machines for classification and regression. *Analyst*, ***135***(2), 230–267. https://doi.org/10.1039/b918972f

Dawson-Howe, K. (2014). *A Practical Introduction to Computer Visio with OpenCV*. John Wiley & Sons, Ltd.

de Almeida, V.E., de Sousa Fernandes, D.D., Diniz, P.H.G.D., de Araújo Gomes, A., Véras, G., Galvão, R.K.H., & Araujo, M.C.U. (2021). Scores selection via Fisher's discriminant power in PCA-LDA to improve the classification of food data. *Food Chemistry,* ***363***, 130296. https://doi.org/10.1016/j.foodchem.2021.130296

Fordellone, M., Bellincontro, A., & Mencarelli, F. (2018). Partial least squares discriminant analysis: A dimensionality reduction method to classify hyperspectral data. *Statistica Applicata*, ***31***(2), 181–200. https://doi.org/10.48550/arXiv.1806.09347

Géron, A. (2019). *Hands-on Machine Learning whit Scikit-Learing, Keras and TensorFlow*. O'Reilly Media, Inc.

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, ***5***(Oct), 1391–1415.

Howley, T., Madden, M.G., O'Connell, M.L., & Ryder, A.G. (2006). The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowledge-Based Systems*, ***19***(5), 363–370. https://doi.org/10.1016/j.knosys.2005.11.014

Husin, Z., Shakaff, A.Y.M., Aziz, A.H.A., Farook, R.S.M., Jaafar, M.N., Hashim, U., & Harun, A. (2012). Embedded portable device for herb leaves recognition using image processing techniques and neural network algorithm. *Computers and Electronics in Agriculture*, ***89***, 18–29. https://doi.org/10.1016/j.compag.2012.07.009

International Coffee Organization. (2018). Grading and classification of green coffee. *The International Coffee Organization*, 1–5. Retrieved from http://www.ico.org/projects/Good-Hygiene-Practices/cnt/cnt_en/sec_3/docs_3.3/Grading & class.pdf

Jayaprakash, C., Damodaran, B.B., Viswanathan, S., & Soman, K.P. (2020). Randomized independent component analysis and linear discriminant analysis dimensionality reduction methods for hyperspectral image classification. *Journal of Applied Remote Sensing*, ***14***(3), 36507. https://doi.org/10.1117/1.JRS.14.036507

Koklu, M., & Ozkan, I.A. (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, ***174***, 105507. https://doi.org/10.1016/j.compag.2020.105507

Koklu, M., Cinar, I., & Taspinar, Y.S. (2021). Classification of rice varieties with deep learning methods. *Computers and Electronics in Agriculture*, ***187***, 106285. https://doi.org/10.1016/j.compag.2021.106285

Kumar, G., & Bhatia, P.K. (2014). A detailed review of feature extraction in image processing systems. *International Conference on Advanced Computing and Communication Technologies (ACCT)*, 5–12. http://dx.doi.org/10.1109/ACCT.2014.74

Lin, H., Sheng, H., Sun, G., Li, Y., Xiao, M., & Wang, X. (2021). Identification of pumpkin powdery mildew based on image processing PCA and machine learning. *Multimedia Tools and Applications*, ***80***(14), 21085–21099. https://doi.org/10.1007/s11042-020-10419-1

Lopes, J.F., Ludwig, L., Barbin, D.F., Grossmann, M.V.E., & Barbon, S. (2019). Computer vision classification of barley flour based on spatial pyramid partition ensemble. *Sensors (Switzerland)*, ***19***(13), 1–17. https://doi.org/10.3390/s19132953

Meenu, M., Kurade, C., Neelapu, B. C., Kalra, S., Ramaswamy, H. S., & Yu, Y. (2021). A concise review on food quality assessment using digital image processing. *Trends in Food Science and Technology*, ***118***(Part A), 106–124. https://doi.org/10.1016/j.tifs.2021.09.014

Mehmood, T., Liland, K.H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, ***118***, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, ***66***(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Oliveira, M.M., Cerqueira, B.V., Barbon, S., & Barbin, D.F. (2021). Classification of fermented cocoa beans (cut test) using computer vision. *Journal of Food Composition and Analysis*, ***97***, 103771. https://doi.org/10.1016/j.jfca.2020.103771

Pazoki, A.R., Farokhi, F., & Pazoki, Z. (2014). Classification of rice grain varieties using two artificial neural networks (mlp and neuro-fuzzy). *Journal of Animal and Plant Sciences*, ***24***(1), 336–343.

Saputra, I., & Kristiyanti, D.A. (2022). *Machine Learning untuk Pemula*. Bandung: Informatika.

Soleimanipour, A., Chegini, G.R., & Massah, J. (2018). Classification of anthurium flowers using combination of PCA, LDA and support vector machine. *Agricultural Engineering International: CIGR Journal*, *20*(1), 219–228.

Tran, H.T.M., Vargas, C.A.C., Slade Lee, L., Furtado, A., Smyth, H., & Henry, R. (2017). Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (*Coffea arabica* L.). *Tree Genetics and Genomes*, ***13***(3), 54. https://doi.org/10.1007/s11295-017-1138-8

Turi, B., Abebe, G., & Goro, G. (2013). Classification of Ethiopian coffee beans using imaging techniques. *East African Journal of Sciences*, ***7***(1), 1–10.

Wallelign, S. (2020). An intelligent system for coffee grading and disease identification [*Master's thesis*]. École Nationale d'Ingénieurs de Brest. Machine Learning.

Widyotomo, S., & Mulato, S. (2005). Performance of a table vibration type coffee grading machine. *Pelita Perkebunan (a Coffee and Cocoa Research Journal)*, ***21***(1), 55–72. https://doi.org/10.22302/iccri.jur.pelitaperkebunan.v21i1.125

Wirth, M.A. (2004). *Shape Analysis & Measurement*. University of Guelph, Computing and Information Science, Image Processing Group. http://www.cyto.purdue.edu/cdroms/micro2/content/education/wirth10.pdf (July 2, 2023),

Zeeshan, M., Prabhu, A., Arun, C., & Rani, N.S. (2020). Fruit Classification System Using Multiclass Support Vector Machine Classifier. *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC* ***2020***, 289–294. http://dx.doi.org/10.1109/ICESC48915.2020.9155817

Zhang, C., Zhang, S., Yang, J., Shi, Y., & Chen, J. (2017). Apple leaf disease identification using genetic algorithm and correlation based feature selection method. *International Journal of Agricultural and Biological Engineering*, ***10***(2), 74–83. https://doi.org/10.3965/j.ijabe.20171002.2166

Zhang, Y., & Wu, L. (2012). Classification of fruits using computer vision and a multiclass support vector machine. *Sensors (Switzerland)*, ***12***(9), 12489–12505. https://doi.org/10.3390/s120912489