https://doi.org/10.48196/016.02.2020.03

Varietal Classification of Selected Green Coffee Beans (*Coffea arabica* L. and *Coffea canephora* Pierre ex A.Froehner) Using Image Processing Software

Katrina Nicole M. Lumagui[1], Luther John R. Manuel[2], Erwin P. Quilloy[3], and Kevin F. Yaptenco[4]

[1]BSABE Graduate, Agricultural, Food & Bioprocess Engineering Division, Institute of Agricultural and Biosystems Engineering, College of Engineering and Agro-Industrial Technology, University of the Philippines Los Baños, 4031 College, Laguna, Philippines (Author for correspondence email: kmlumagui@up.edu.ph)

[2]University Researcher II, Agricultural, Food & Bioprocess Engineering Division, Institute of Agricultural and Biosystems Engineering, College of Engineering and Agro-Industrial Technology, University of the Philippines Los Baños, 4031 College, Laguna, Philippines

[3]Assistant Professor 7, Agribiosystems Machinery and Power Engineering Division, Institute of Agricultural and Biosystems Engineering, College of Engineering and Agro-Industrial Technology, University of the Philippines Los Baños, 4031 College, Laguna, Philippines

[4]Professor 9, Agricultural, Food & Bioprocess Engineering Division, Institute of Agricultural and Biosystems Engineering , College of Engineering and Agro-Industrial Technology, University of the Philippines Los Baños, 4031 College, Laguna, Philippines

ABSTRACT

The study created a classification model that sorts green coffee beans based on its variety, primarily 'Arabica' and 'Robusta' to address coffee adulteration. A total of 1500 green coffee beans were obtained from Cavite, Philippines, allotting 70% and 30% of it for training and testing dataset, respectively. These were then captured through the fabricated image acquisition setup. Image processing techniques were performed to extract the features of the beans namely color, size, shape, and crack using ImageJ software. Fourier analysis was mainly performed for the extraction of the bean shape information. After performing t-test, 3, 4, and 7 parameters of color, size, and shape features were found to be significant, respectively. The 14 features were used in creating 7 classifications setups of different feature combinations through discriminant analysis. The model that yielded the highest accuracy (99%) is the combination of color and size features. This exceeded the subjective varietal classification accuracy (98%) performed by a coffee expert. Thus, the variety of green coffee beans were feasibly classified through image analysis and the created model had surpassed the traditional performance of sorting green coffee beans.

Keywords: classification, green coffee beans, Fourier analysis, image analysis

INTRODUCTION

Coffee has more than 100 species, but the main varieties are 'Arabica' and 'Robusta'. These are the most commercially produced varieties that are frequently compared due to its interchangeable characteristics. However, these varieties have prominent differences in physical and chemical properties. 'Arabica' coffee is considered as a better variety compared to 'Robusta' due to its taste and acidity level. Yet, 'Robusta' is easier to grow and harvest which is why it is twice cheaper than 'Arabica' (Hoffman, 2014). 'Arabica' and 'Robusta' varieties take up the two largest percentage of world's coffee production which is 70% and 27%, respectively (El Sheikha *et al*., 2018). As of July

2020, the production of coffee worldwide is recorded to be 170.94 million bags: each containing 60 kilograms. Most of this production was harvested from Brazil. On the other hand, United States is found to be the country with the highest revenue from the coffee market. 'Arabica' coffee also remained to be expensive twice more than 'Robusta' (Bedford, 2020).

Food frauds and adulterations continuously arise in the world. The recently reported products being adulterated includes milks, wines, and alcoholic beverages by adding water, toxic chemicals, and cheaper alternative ingredients. Hence, the product quality is being degraded as its quantity is increased (Food Fraud Cases, 2020).

As reported by Dynbuncio (2013), coffee is one of the top 10 most adulterated foods. One way of degrading coffee is adding low-value materials after grinding it. Another is adding or replacing an expensive coffee variety by a cheaper one. In fact, news from Federal Food Safety and Veterinary Office (2019) imparts that in Bern, Switzerland, 'Robusta' coffee beans had replaced 'Arabica' in selected packs of coffee with labels stating 100% 'Arabica' coffee beans. After taking samples of the said coffee, a high level of 16-O-methylcafestol contained only in 'Robusta' coffee was determined. Since about half the price of 'Arabica' beans is 'Robusta's' price, this news alerts the affected consumers.

Mostly, farmers use their visual decision making in agricultural works. However, this traditional method is inconsistent and inaccurate. The application of image processing in agriculture must be used since it establishes efficiency and precision in practices such as grading, sorting, and inspecting while lessening uncertainty in the obtained data (Armstrong & Saxena, 2014). There are several agricultural studies that used image processing. One of those is the research conducted by Abebe *et al.* (2013) wherein a computer routine algorithm was developed to classify Ethiopian coffee beans based on their geographic origins. Since manual classification of beans is known to be inaccurate and tedious, the researchers used an imaging technique for enhancing the efficiency in classification which led to a consideration that imaging technique is the most efficient practice in coffee beans classification. Abbaspour-Gilandeh and Azizi (2014) then focused on applying image analysis on potatoes to classify the regularly shaped from the irregularly shaped ones. Fourier transform and geometrical features were also used for better accuracy of classifying the shape of commodity. Using this method, the researchers were able to achieve an accuracy of 98%. Takahashi *et al* (2013) then used image analysis to evaluate the quality of tomato and its color changes at different maturity stages due to the storage duration and temperature.

In the coffee supply chain, the varietal classification model to be developed can help the importer countries and coffee intermediaries to identify if the packages of green coffee beans being delivered are of the correct variety. Aside from preventing adulteration, the color details of a green coffee bean provide information whether it was picked in an immature or an overripe condition. Traditionally, only the coffee experts, people who have enough knowledge on growing and processing coffee beans, can distinguish whether the coffee beans are labelled correctly based on the variety type. However, it is tedious and time consuming if large number of coffee beans are to be inspected by them manually. This traditional sorting is commonly done by manually picking the beans that are not the same variety from the labelled package through the expert's visual decision. On the other hand, in terms of defects in coffee samples, there are already coffee sorter machines wherein the classification is based on the number of defects in a sample. The machine separates coffee beans based on its defects such as being immature, broken, and insect damaged. That way, the quality degradation of coffee to be processed was minimized (Preedy, 2014). Thus, an accurate and efficient varietal classification method is needed as intended by the study. This study can also serve as a fundamental step in developing a green coffee bean varietal sorter machine.

The general objective of this study was to develop a computer-based image analysis system for classifying the green coffee bean varieties. Particularly, the study aimed to establish the physical characteristics of 'Arabica' (*Coffea arabica*

L.) and 'Robusta' (*Coffea canephora Pierre* ex A.Froehner) green coffee beans necessary for image analysis, to develop an image acquisition and analysis system for the classification of coffee bean varieties, and to test the performance of the established varietal classification model.

MATERIALS AND METHODS

The samples used in this study were limited to the 'Arabica' and 'Robusta' green coffee beans bought from three coffee stores in Silang and Amadeo, Cavite, Philippines. The total number of considered green coffee beans is 1500, allotting 70% (525 'Arabica' and 525 'Robusta' beans) for training and 30% (225 'Arabica' and 225 'Robusta' beans) for testing.

Digital Imaging Acquisition

Green coffee beans were manually and singly captured using the image acquisition setup shown in Figures 1 and 2. Each bean was captured in one view, with the cracked side up and endpoints of the crack falling within the white horizontal crosshair of the camera as illustrated in Figure 3. Each snap was saved in Joint Photographic Expert Group (JPEG) format with a horizontal and vertical resolution of

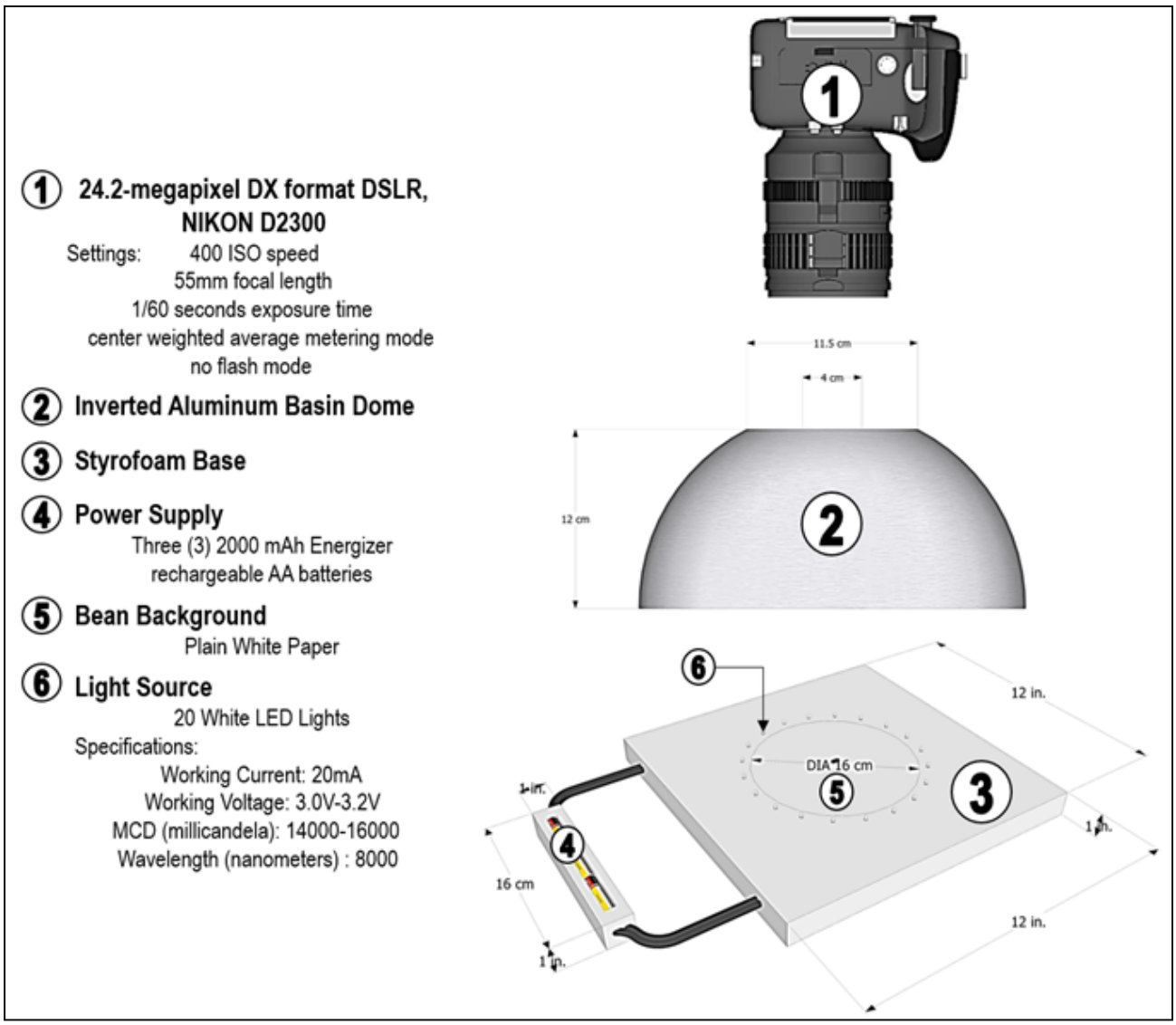

Figure 1. Schematic diagram of the image acquisition setup with specifications.

Figure 2. Actual image acquisition setup used.

300 dpi and an image size of 3008 pixels x 2000 pixels.

Feature Extraction

In this study, four physical properties of green coffee beans were analyzed: (1) color, (2) size, (3) shape, and (4) crack features. Feature extraction was performed using ImageJ 64-bit Java 1.8.0 (ImageJ, 2020), an open-source software for image processing. The flowchart of image processing steps is shown in Figure 4. For each property, different pre-processing steps were performed to have an ease in extracting the object from its background.

Color Feature

For the color feature extraction, each image was pre -processed by subtracting a constant value from it for easier selection of the bean outline. The average RGB color of the bean was then measured. Using Microsoft Excel, these were converted into HSI values using Equations 1, 2, and 3 which were adapted from the study of Gonzalez and Wood (2002) as cited by Abebe *et al.* (2013).

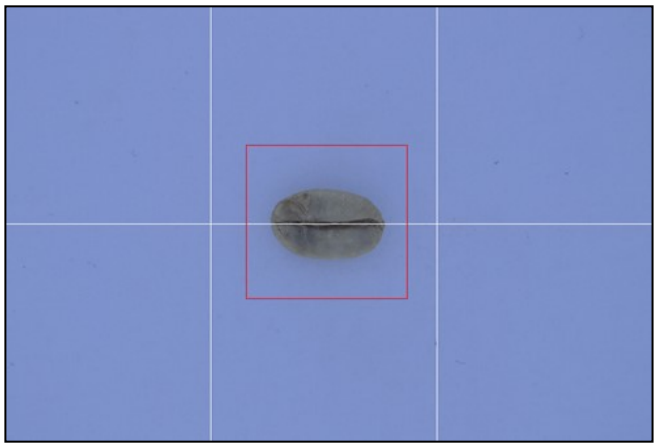

Figure 3. Green coffee bean to be captured.

$$\cos H = \frac{2R-G-B}{2\sqrt{(R-G)^2+(R-B)(G-B)}}$$

Equation 1

where:

H is the hue value
R is the red color value
G is the green color value
B is the blue color value

$$S = 1 - \frac{3}{R+G+B} \min (R, G, B)$$

Equation 2

where:

S is the saturation value
R is the red color value
G is the green color value
B is the blue color value

$$I = \frac{R+G+B}{3}$$

Equation 3

where:

I is the intensity value;
R is the red color value
G is the green color value
B is the blue color value

Size Feature

For the size feature, each image was also pre-processed the same steps as for the color feature. The projected area, major diameter, minor diameter, and perimeter of each bean were measured together. The obtained size data in ImageJ were in square pixels and were processed in Microsoft Excel. For the data calibration, cut-out circles and cut-out ellipses with different diameter and major and minor diameters, respectively were captured and processed in ImageJ like the bean images.

Projected areas, perimeters, and major and minor diameters were also obtained from the cut-outs of circle and ellipse. The relationship of the beans to cut-out circles and the relationship of the same beans to cut-out ellipses were obtained by creating a separate scatter plot for each parameter. The trendline of the plots, one for circle and one for ellipse in each parameter, were then created as well as the respective trendline equations. Moreover, the coefficient of determination (R^2) value in each relationship was obtained. Generally, the relation that had the R^2 value closest to 1 was considered in calibrating the obtained ImageJ data.

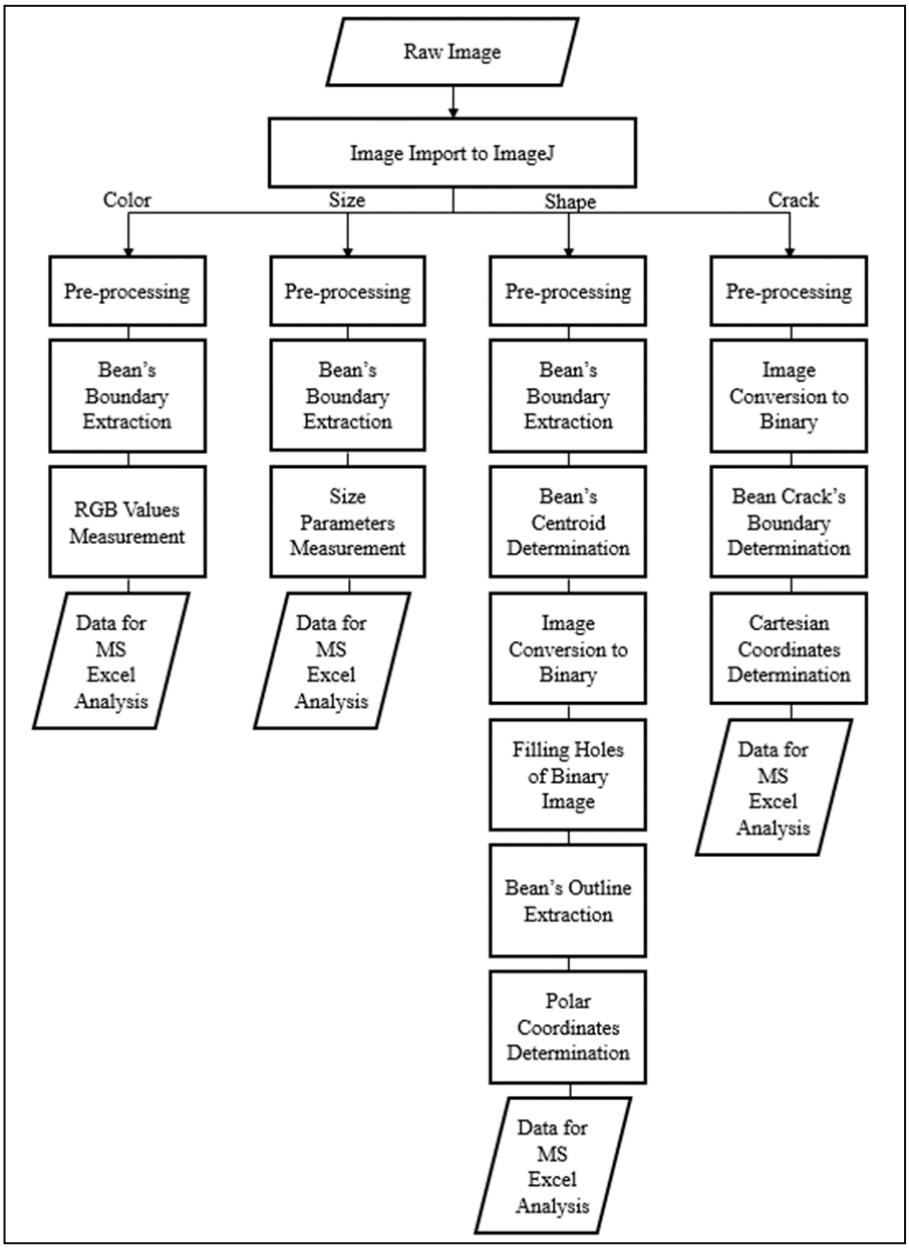

Figure 4. Flowchart of procedures using ImageJ.

Shape Feature

For shape feature, the images were converted to binary images. The bean was selected to obtain the polar coordinates of the bean boundary. The obtained radii were normalized using Microsoft Excel. Using the resulting radius per angle, the geometric signature of the bean was illustrated. The spatial domain was then converted into the Fourier domain through the discrete Fourier transform (DFT) through the Excel add-in called NumXL version 1.65 (NumXL, 2020). The conversion generated Fourier transform magnitudes $\hat{r}(h)$, frequencies (h), and phase angles. Qualitatively, several descriptors should be selected based on the generated plot of Fourier transform magnitudes versus frequencies to approximate the true shape of the bean. According to Sonka *et al.* (1993) as cited by Bowman, Soga, and Drummond (2000), the first ten to fifteen obtained descriptors are sufficient to define a complex object's shape. Kindratenko and

Van Espen (2002) then cited that the first ten to twenty obtained descriptors give a good, reconstructed object's shape. In this study, the first 15 descriptors were utilized since the bean's recreated shape yielded a smooth shape signature that approaches the original one. Illustrated in Equation 4 is the transformation equation to convert space domain into frequency domain using discrete Fourier transform (NumXL, 2008).

$$x_k = \sum_{j=0}^{N-1} x_j e^{-\frac{2\pi i}{N} jk}$$

Equation 4

where:

k is the frequency component
$X_0, \ldots, X_{N-1}$ is the input time series value (radius values for this study)
N is the number of non-missing values in the input time series
i denotes that the exponential function is complex

These selected 15 descriptors were used to perform the inverse discrete Fourier transform (IDFT) using Microsoft Excel. This method constructed a waveform from the selected Fourier descriptors for the shape signatures to be visualized, analyzed, and compared to the original shape signature. The use of the Fourier transform in this study was based on the performed experiment of Bo, Fuguo, Peng, and Zhanwei (2013).

Roundness and circularity were also considered for the shape feature and were calculated as shown in Equation 5 and Equation 6 (NIH, 2020).

$$C = \frac{4\pi x A}{P^2}$$

Equation 5

where:

C is the circularity value
A is the area in cm^2
P is the perimeter in cm

$$R = \frac{4\pi x A}{D^2}$$

Equation 6

where:

R is the roundness value
A is the area in cm^2
D is the major diameter in cm

Crack Feature

According to Pais (2015), the 'Arabica' coffee bean has a central split or crack that is wavier or that resembles an 'S' shape whereas the 'Robusta' coffee bean crack resembles a straight-line shape. To extract this feature, each image was pre-processed by multiplying a constant value from it to extract the bean crack from the coffee bean and its background. The enhanced image was transformed into a binary image then the boundary crack was selected. The cartesian coordinates of the boundary was extracted to serve as the data to be analyzed. The obtained y-coordinates were

Table 1. Color features extracted from 'Arabica' and 'Robusta' training beans.

COLOR FEATURES	'ARABICA'				'ROBUSTA'				T-TEST
	Minimum	Maximum	Mean	Standard Deviation	Minimum	Maximum	Mean	Standard Deviation	
H	0.997	2.972	2.264	0.152	1.772	2.753	2.161	0.081	*
S	0.001	0.177	0.024	0.012	0.016	0.355	0.085	0.066	*
I	0.270	0.537	0.377	0.034	0.208	0.371	0.278	0.025	*

* − significant at $P < 0.05$

normalized, and its variance was computed using Microsoft Excel.

Statistical Analysis

T-test

The processed feature values of the training green coffee beans were analyzed using T-test through XLSTAT add-in of Microsoft Excel to determine whether these values have significant differences between the two varieties for the creation of the classification model.

Discriminant Analysis

Discriminant analysis was performed through XLSTAT wherein 1050 and 450 beans were used as the training and testing dataset, respectively. In all the models created, the covariance matrices of both varieties were assumed to be unequal. Cross-validation was performed in training the classification model to avoid overestimation of data. The testing dataset was utilized to create the confusion matrix which will determine the model's accuracy.

Manual Sorting of Green Coffee Beans

To evaluate the performance of the created varietal classification model, its results was compared to the results of the traditional sorting of green coffee beans based on the variety. A total of 1200 green coffee beans were manually sorted based on its physical appearances by a recommended coffee expert of the College of Agriculture and Food Science, University of the Philippines Los Baños (CAFS-UPLB). The coffee expert is a laboratory technician of the CAFS-UPLB with extensive experience in processing cacao and coffee beans.

RESULTS AND DISCUSSIONS

Features Extraction

Color Feature

Table 1 shows the training color data of beans at $P < 0.05$. As observed, the hue and intensity mean values of 'Arabica' beans were higher compared to 'Robusta'; the opposite was observed for saturation mean values. Considering hue, both varieties have values approaching the mean red color; however, it failed to describe the beans color from the hue chart. These values were affected by the obtained saturation values of both varieties that are closer to 0 than 1 because according to Blotta *et al.* (2011), when the saturation value of an object approaches 0, the hue value will only reflect colors between black and white. Nevertheless, hue values are significantly different between the two varieties based on the t-test. For the saturation values, 'Arabica' beans were whiter than 'Robusta'. For the intensity values, both varieties have values near 0; however, 'Arabica' beans have higher mean value being it lighter than 'Robusta'. The mean values of saturation and intensity were also significantly different between the varieties; thus,

Table 2. Size features extracted from 'Arabica' and 'Robusta' training beans.

SIZE FEATURES	'ARABICA'				'ROBUSTA'				T-TEST
	Minimum	Maximum	Mean	Standard Deviation	Minimum	Maximum	Mean	Standard Deviation	
Area (cm^2)	0.001	1.037	0.665	0.090	0.217	0.639	0.403	0.066	*
Perimeter (cm)	0.294	10.758	3.004	0.550	1.644	4.442	2.225	0.209	*
Major Diameter (cm)	0.031	1.425	1.080	0.098	0.542	0.956	0.773	0.072	*
Minor Diameter (cm)	0.025	1.011	0.780	0.066	0.508	0.854	0.659	0.055	*

* − significant at $P < 0.05$

Table 3. Shape features extracted from 'Arabica' and 'Robusta' training green coffee beans.

SHAPE FEATURES	'ARABICA'				'ROBUSTA'				T-TEST	IMPLIED SHAPE INFORMATION
	Minimum	Maximum	Mean	Standard Deviation	Minimum	Maximum	Mean	Standard Deviation		
Roundness	0.56	1.46	0.73	0.07	0.65	0.99	0.86	0.05	*	
Circularity	0.11	1.04	0.95	0.12	0.28	1.056	1.02	0.05	*	
Coefficient 1	0.05	78.94	1.75	5.02	0.09	29.65	1.28	1.86	ns	
Coefficient 2	8.71	51.85	30.39	7.55	1.72	40.06	14.76	5.56	*	Elongation
Coefficient 3	0.38	53.94	4.35	3.38	0.10	8.79	2.99	1.61	*	Triangularity
Coefficient 4	0.10	49.94	2.61	2.92	0.06	7.79	1.79	1.10	*	Squareness
Coefficient 5	0.02	49.88	2.08	2.75	0.07	8.95	1.23	0.84	*	Asymmetry
Coefficient 6	0.03	48.86	1.50	2.70	0.02	6.59	1.01	0.73	*	Angularity
Coefficient 7	0.07	45.84	1.17	2.56	0.01	6.78	0.77	0.64	ns	
Coefficient 8	0.02	46.59	1.04	2.55	0.02	6.51	0.70	0.62	ns	
Coefficient 9	0.02	44.46	0.89	2.51	0.01	6.72	0.60	0.59	ns	
Coefficient 10	0.02	41.70	0.84	2.41	0.01	6.52	0.56	0.58	ns	
Coefficient 11	0.01	40.94	0.79	2.37	0.02	6.70	0.49	0.54	ns	
Coefficient 12	0.01	37.61	0.71	2.25	0.01	6.62	0.45	0.53	ns	
Coefficient 13	0.01	36.18	0.66	2.21	0.01	6.51	0.41	0.52	ns	
Coefficient 14	0.01	33.369	0.61	2.09	0.02	6.57	0.38	0.53	ns	
Coefficient 15	0.01	31.830	0.60	2.04	0.02	6.54	0.35	0.50	ns	

* − significant at $P < 0.05$; ns – not significant at $P < 0.05$

all the color features were utilized in the model creation.

Size Feature

In size calibration, the linear regression of actual versus estimated size features showed that calibration was marginally better using a circle than ellipse; R^2 values, however, were all greater than 0.999 regardless of shape. Hence, the circle trendline equations for all size parameters: projected area, perimeter, major diameter, and minor diameter were used in calibration as presented in Equations 7, 8, 9, and 10, respectively. Table 2 shows the summarized training size data of the beans at $P < 0.05$.

$$y = 2.2586E\text{-}14\ x^2 + 4.6038E\text{-}06\ x + 5.4940E\text{-}04$$ **Equation 7**

where:

y is the projected area of beans in cm^2
x is the projected area of beans in square pixels

$$y = 3.4356E\text{-}08\ x^2 + 1.8848E\text{-}03\ x + 5.5382E\text{-}02$$ **Equation 8**

where:

y is the perimeter of beans in cm
x is the perimeter of beans in square pixels

$$y = 1.8729E\text{-}08\ x^2 + 2.1391E\text{-}03\ x - 1.5988E\text{-}03$$ **Equation 9**

where:

y is the major diameter of beans in cm^2
x is the major diameter of beans in square pixels

$$y = 2.8225E\text{-}08\ x^2 + 2.1305E\text{-}03\ x + 4.1242E\text{-}03$$ **Equation 10**

where:

y is the minor diameter of beans in cm^2
x is the minor diameter of beans in square pixels

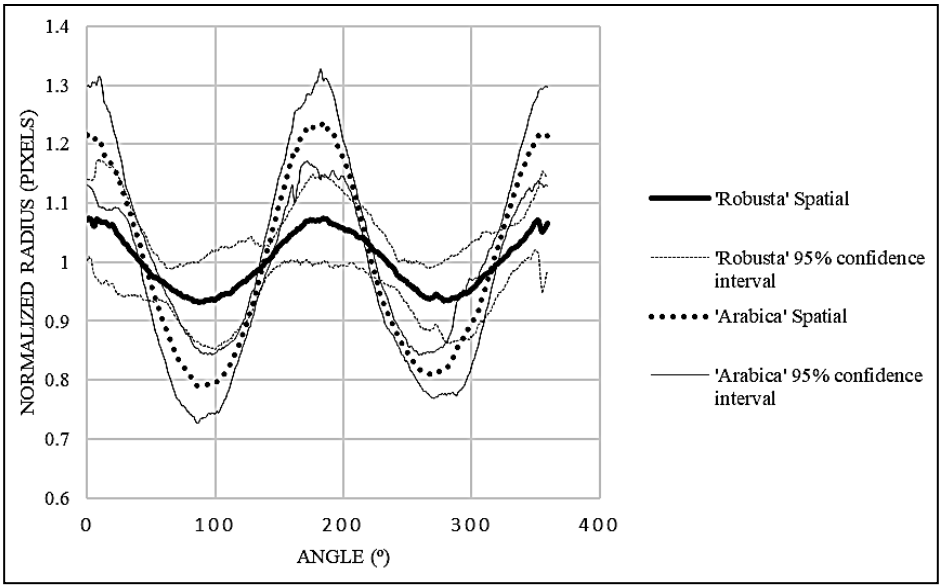

Figure 5. Mean geometric signatures of 'Arabica' and 'Robusta' beans.

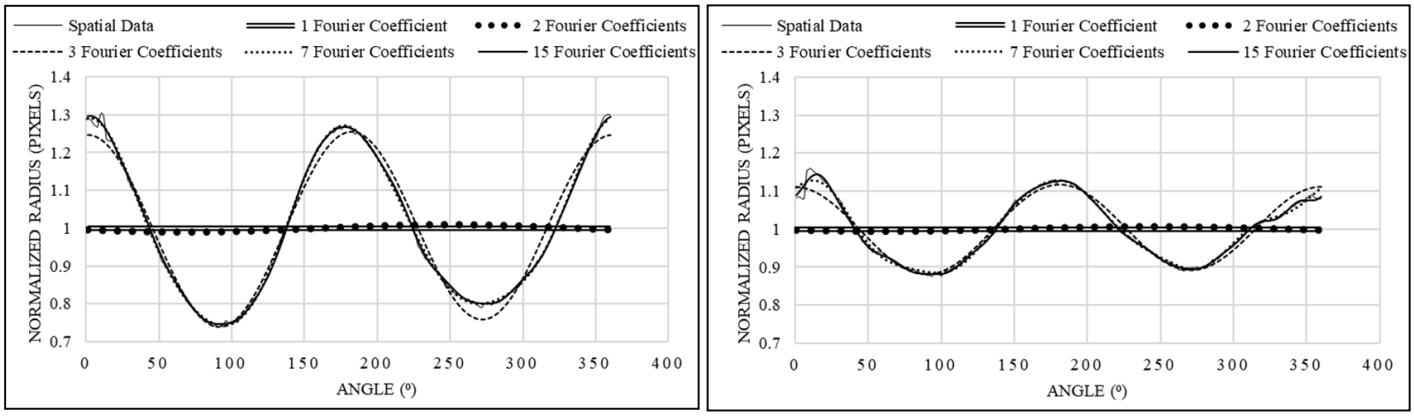

Figure 6. Recreated geometric signatures of 'Arabica' (left) and 'Robusta' (right) bean using 1, 2, 3, 7, and 15 Fourier coefficients.

Table 2 shows that all the size values of 'Arabica' beans were higher than 'Robusta'. There is also a significant difference between the varieties according to t-test; thus, all the size features were utilized in model creation.

Shape Feature

Table 3 shows the training shape data of the beans at $P < 0.05$. For roundness, the mean value of 'Robusta' beans is higher and closer to 1 than 'Arabica'; hence, 'Robusta' is rounder. For circularity, which measures the irregularities of an object, 'Robusta' has higher mean value than 'Arabica'. Aside from these parameters, the selected 15 Fourier coefficients were also analyzed.

Figure 5 shows the geometric signature of 'Arabica' and 'Robusta' bean. The most noticeable difference in the shape signatures is that 'Arabica' beans have higher amplitude than 'Robusta'.

Figure 6 presents the recreated geometric signatures of beans using 1, 2, 3, 7, and 15 Fourier coefficients of 'Arabica' and 'Robusta'. As observed, the generated signature using the selected Fourier data is a simplified version of the spatial data. Also, the lesser the Fourier data used in shape signature recreation, the simpler the geometric signature was recreated.

Fourier coefficients were obtained to analyze the bean shape signatures between the varieties. The objective of Fourier transform is to reduce the dimensionality of data in spatial domain; thus, translating it into frequency domain wherein the resulting Fourier coefficients do not depend on rotation, translation, and starting point. Shown in Figure 7 and 8 are the boundary of one sample 'Arabica' and 'Robusta' beans, respectively based on the spatial and on the inverse Fourier transform using the 15 coefficients. The recreated shapes of 'Arabica' and 'Robusta' green coffee beans are the simplified version of original shapes.

Based on Table 3, 5 of the 15 Fourier coefficients were significantly different between the varieties and were the most useful for the model creation. These coefficients are Coefficient 2 (F[2]), 3 (F[3]), 4 (F[4]), 5 (F[5]), and 6 (F[6]) wherein each has corresponding shape information. This shape information was based on the performed Fourier transformation of simple shapes namely: circle, ellipse, triangle, square, and rectangle by Bo *et al.* (2013) and was supported by Abbaspour-Gilandeh and Azizi (2014) who analyzed the corresponding shape information of potatoes by determining its Fourier coefficients.

All the mean coefficient values of 'Arabica' were higher than those of 'Robusta' as observed in Table

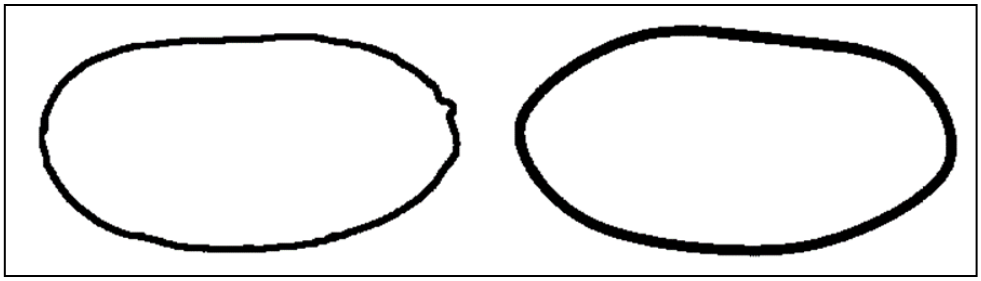

Figure 7. Original boundary (left) and recreated boundary through Fourier Transform (right) of one 'Arabica' bean.

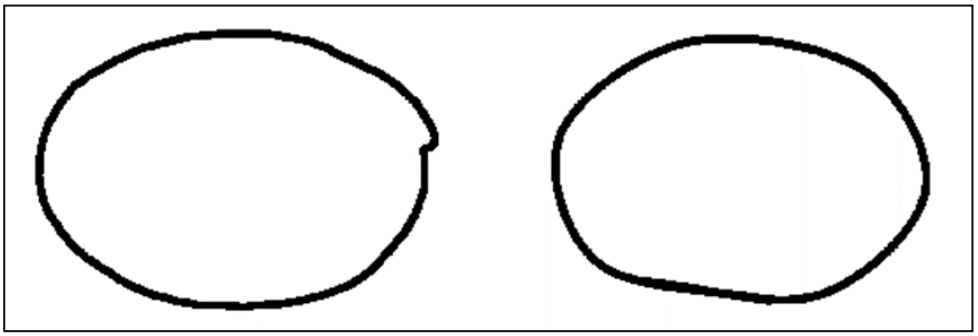

Figure 8. Original boundary (left) and recreated boundary through Fourier Transform (right) of one 'Robusta' bean.

3. For F[2], 'Arabica' bean is more elongated than 'Robusta'. For F[3], 'Arabica' approaches a triangular shape than 'Robusta'. This may be due to the irregularity of 'Arabica' since it has lesser smoothness; thus, having more angles in its shape boundary. This also supports F[4] implying squareness and F[6] denoting angularity of the bean due to the rough edges of 'Arabica'. For F[5], 'Arabica' beans have lack of symmetry or have an unbalanced shape. This is prominent since most of

Table 4. Confusion matrices for the classification models and traditional sorting.

CLASSIFICATION MODEL	ACTUAL VARIETY	PREDICTED VARIETY		TOTAL	% CORRECT
		'Arabica'	'Robusta'		
Significant Color Features	'Arabica'	223	2	225	99%
	'Robusta'	8	217	225	96%
	Total	231	219	450	98%
Significant Size Features	'Arabica'	213	12	225	95%
	'Robusta'	6	219	225	97%
	Total	219	231	450	96%
Significant Shape Features	'Arabica'	164	61	225	73%
	'Robusta'	13	212	225	94%
	Total	177	273	450	84%
Significant Color and Size Features	'Arabica'	223	2	225	99%
	'Robusta'	4	221	225	98%
	Total	227	223	450	99%
Significant Color and Shape Features	'Arabica'	221	4	225	98%
	'Robusta'	7	218	225	97%
	Total	228	222	450	98%
Significant Size and Shape Features	'Arabica'	215	10	225	96%
	'Robusta'	10	215	225	96%
	Total	225	225	450	96%
Significant Color, Size, and Shape Features	'Arabica'	223	2	225	99%
	'Robusta'	6	219	225	97%
	Total	229	221	450	98%
Traditional Sorting	'Arabica'	581	19	600	97%
	'Robusta'	11	589	600	98%
	Total	592	608	1200	98%

Note: The significant color features are hue, saturation, and intensity. The significant size features are projected area, perimeter, major diameter, and minor diameter. The significant shape features are roundness, circularity, F[2], F[3], F[4], F[5], and F[6].

the 'Arabica' beans are difficult to classify whether it approach circles or ellipses.

Crack Feature

For the crack feature, the variance of the bean crack boundary y-coordinates between the varieties was not significant at $P < 0.05$. According to Pais (2015), 'Arabica' green coffee beans has an S-shaped crack while 'Robusta' has a straight; thus, it should have higher mean variance values. However, the expected mean results were not met. Hence, the crack feature is not significantly different between the varieties.

Features Classification

Classification using Color Features

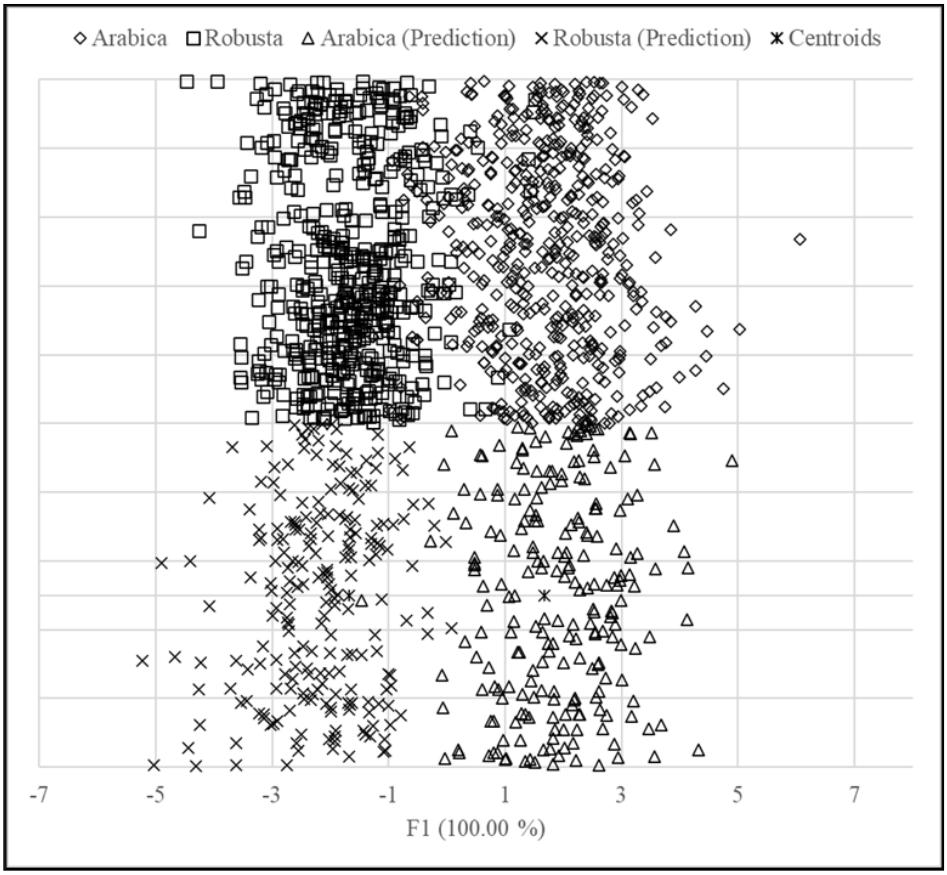

Figure 9. Jitter chart for the significant color features classification model.

In this classification model, the 3 significant color features: hue, saturation, and intensity were utilized to classify bean varieties. A jitter chart was presented in Figure 9 to determine how the varieties are discriminated based on color and to allow the visualization of observations in new space. As observed, the separation of the varieties regarding color feature was distinct but an overlapping was still observed. Thus, an occurrence of misclassification was expected to occur.

Table 4 presents the compiled confusion matrices for all classification models including the results for traditional sorting. It shows the results of the classification and the misclassification of the testing dataset. Based on Table 4, the total accuracy of this classification model is 98%. As observed, there are a greater number of 'Robusta' that exhibit the color of 'Arabica'.

Classification using Size Features

In this classification model, the 4 significant size features: area, perimeter, major diameter, and minor diameter were considered in classifying the bean varieties. Like the previous chart, shown in Figure 10 is the jitter chart for this model. The separation of the varieties is also prominent; yet an overlapping occurred between the varieties. Thus, an occurrence of misclassification is also likely to occur.

Based on Table 4, this model yielded a total accuracy of 96%. Thus, there are 'Robusta' which are bigger and 'Arabica' which are smaller.

Classification using Shape Features

In this classification model, 7 significant shape features were considered: roundness, circularity, and

5 Fourier coefficients to classify the beans based on its variety. Same generalizations in the obtained shape jitter chart shown in Figure 11 as the previous charts were also observed.

As observed in Table 4, this model had a total accuracy of 84%. Hence, more 'Arabica' beans were misclassified compared to 'Robusta'.

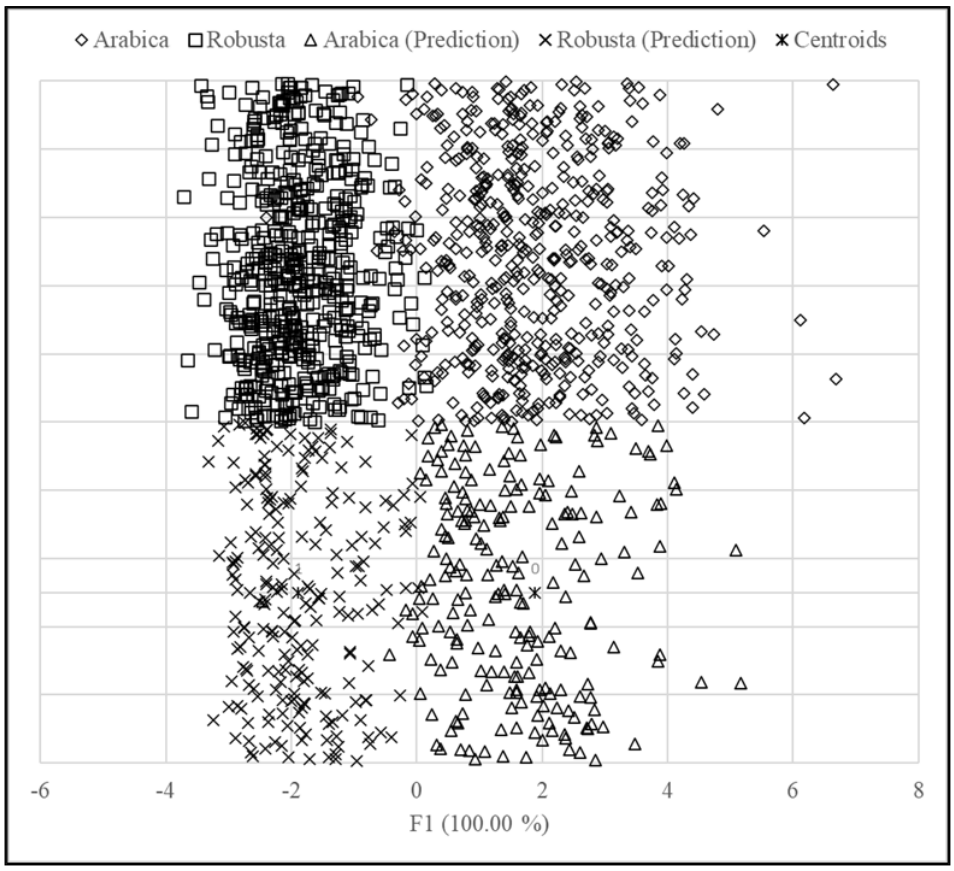

Figure 10. Jitter chart for the significant size features classification model.

Classification using Combined Features of Color and Size

In this classification model, 7 significant green coffee bean features were considered: hue, saturation, intensity, area, perimeter, major diameter, and minor diameter. Since the color and size classification models had high accuracy each, combining both features may yield to higher accuracy. As presented in Table 4, this model yielded a total accuracy of 99% which is the highest among all the created models.

Classification using Combined Features of Color and Shape

Significant features of color and shape were also combined to create a classification model. As shown in Table 4, the combined color and shape model had a total accuracy of 98%.

Classification using Combined Features of Size and Shape (Morphology)

Size and shape of an object refers to morphology. Combining the features of size and shape resulted to the ability of the bean's morphology to classify each into 'Arabica' or 'Robusta' variety. Based on Table 4, this model yielded a 96% total accuracy.

Classification using Combined Features of Color, Size, and Shape

Lastly, all the significant features of the green coffee beans were used to create a varietal classification model. The features include hue, saturation, intensity, area, perimeter, major and minor diameter, roundness, circularity, and 5 Fourier coefficients. This model yielded to a 98% total accuracy as observed in Table 4.

Generally, among all the models, the combined color and size feature yielded the highest total accuracy of 99% whereas the classification using shape feature had the lowest accuracy of 84%.

Performance of the Varietal Classification Model

The results of the manually sorted green coffee beans are summarized also in Table 4. The sorting

of the 1200 green coffee beans by the coffee expert was performed for about 30 minutes. As observed in the data, the total accuracy of manually sorted beans was 98%. Such accuracy was exceeded by the created varietal classification model that tested the combined significant color and size of 450 green coffee beans with 99% accuracy for only about 8 minutes. Therefore, the varietal classification model of the combined color and size features sorts green coffee beans more accurately and efficiently compared to the traditional sorting. However, when the developed classification model was further compared to the classification model of Abebe et al (2013) which classifies green coffee beans based on their botanical origin, the latter is better based on its accuracy having a 100%. The model of Abebe et al (2013) was developed using MATLAB (version R2012b), a proprietary software which is not free to use. On the other hand, the software used in this study are ImageJ which is an open-source and Microsoft Excel which is widely available.

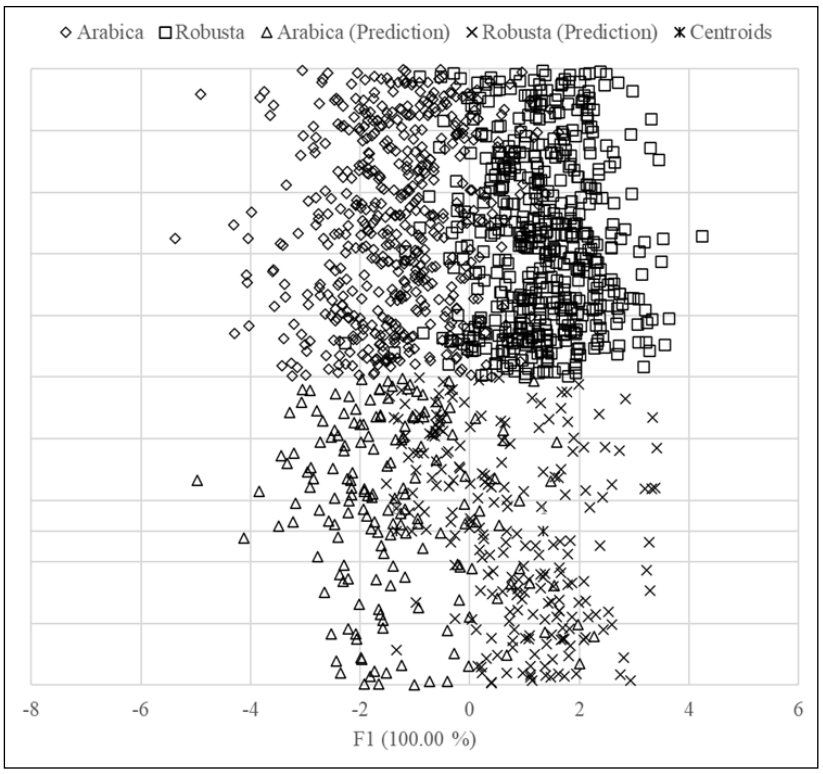

Figure 11. Jitter chart for the significant shape features classification model.

SUMMARY AND CONCLUSION

Green coffee beans were captured using the fabricated dome-shaped image acquisition setup. Each image was then processed for extracting bean features. After the feature extraction procedures, three color, four size, seventeen shape, and one crack feature were extracted in each bean. A total of fourteen green coffee bean features were then found to be significantly different between the two varieties and were considered as the physical characteristics useful for image analysis. These were used in developing and testing seven classification setups wherein the color and size combination model yielded the highest classification accuracy (99%) of classifying 450 beans for about 8 minutes. Furthermore, a subjective varietal classification of 1200 green coffee beans was performed by a coffee expert which resulted to a total accuracy of 98% for 30 minutes. Hence, the combined features of color and size classification model efficiently surpassed the performance of manual bean sorting by a coffee expert based on the classification accuracy and time duration.

RECOMMENDATIONS

The image acquisition setup developed in this study is recommended to have an automatic and faster clicking system to accommodate many samples. Numerous beans in a single image are also recommended using a tray with compartments to address the case of overlapping beans. Moreover, a power source directly connected to the outlet is preferred to have a constant lighting inside the dome. The specifications of the camera and lights used in capturing the samples must be noted to ensure the repeatability of the methods. An improved program that will increase the accuracy and shorten the duration of classifying the coffee beans is also advised. This may include the use of other software that can provide better accuracy. It is

also recommended to use a single programmable software that can pre-process images, extract features, extract Fourier coefficients, and analyze statistical data to minimize importation of data into different software and to further shorten the duration of classification. Large number of samples which came from different provinces in the country is then recommended to consider the diversity of coffee beans.

ACKNOWLEDGEMENT

The authors would like to thank the Agricultural, Food & Bioprocess Engineering Division (AFBED), Institute of Agricultural and Biosystems Engineering (IABE), College of Engineering and Agro-Industrial Technology, University of the Philippines Los Baños, 4031 College, Laguna, Philippines, for the assistance in the conduct of the study and for allowing the use of its facilities. We would also like to thank Alfredo Pascual Sr, of the College of Agriculture & Food Science, UP Los Banos, for lending his expertise and experience in sorting experimental samples for the study. The guidance of Paolo Rommel Sanchez, a faculty member of the IABE, on Fourier analysis is also gratefully acknowledged.

LITERATURE CITED

ABBASPOUR-GILANDEH, Y., & AZIZI, A. (2014). Identifying Irregular Potatoes by Developing an Intelligent Algorithm Based on Image Processing. Journal of Agricultural Sciences, 22 (2016), 32-41. Retrieved July 06, 2019 from https://www.researchgate.net/publication/303739870_Identifying_irregular_potatoes_by_developing_an_intelligent_algorithm_based_on_image_processing

ABEBE, G., GORO, G., & TURI, B. (2013). Classification of Ethiopian Coffee Beans Using Imaging Techniques. East African Journal of Sciences, 7(1), 1-10. doi: 10.1.1.1020.8929

ARMSTRONG, L., & SAXENA, L. (2014). A survey of image processing techniques for agriculture. Proceedings of Asian Federation for Information Technology in Agriculture. Perth, W.A. Australian Society of Information and Communication Technologies in Agriculture, 401-413. Retrieved June 29, 2019 from https://ro.ecu.edu.au/ecuworkspost2013/854

BEDFORD, E. (2020). Coffee Market Worldwide – Statistics & Facts. Retrieved November 08, 2020 from https://www.statista.com/topics/5945/coffee-market-worldwide

BLOTTA, E., BOUCHET, A., BALLARIN, V., & PASTORE, J. (2011). Enhancement of Medical Images in HSI Color Space. *Journal of Physics Conference Series,* doi: 10.1088/1742-6596/332/1/012041

BO, C., FUGUO, L., PENG Z., & ZHANWEI, Y. (2013). Description of Shape Characteristics through Fourier and Wavelet Analysis. Chinese Journal of Aeronautics, (2014),27(1): 160–168. Retrieved August 04, 2019 from https://www.sciencedirect.com/science/article/pii/S1000936113001477

BOWMAN, E.T., SOGA, K., & DRUMMOND, T.W. (2000). Particle Shape Characterization using Fourier Analysis. CUED/D-SoWTR315. Retrieved January 20, 2020 from https://pdfs.semanticscholar.org/4757/eeef6081f487b2e4eae3083bae1b2dae5fe4.pdf

DYNBUNCIO, M. (2013, January 22). Food frauds: 10 most adulterated foods. CBS News.Retrieved November 09, 2020 from https://www.cbsnews.com/media/food-frauds-10 - most-adulterated-foods/

EL SHEIKHA, A.F., LEVIN, R., & XU, J. (2018). Molecular Techniques in Food Biology: Safety, Biotechnology, Authenticity and Traceability. United States: John Wiley & Sons.

FEDERAL FOOD SAFETY AND VETERINARY OFFICE. (2019). Checking of coffee labelling: three false declarations in Switzerland. Retrieved July 28, 2019 from https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-75501.html

FOOD FRAUD CASES. (2020). JRC Food Fraud Monthly Report: European Union, Europe.

HOFFMAN, J. (2014). *The World Atlas of Coffee* (2nd ed.) [Electronic Version]. North America: Firefly Books.

IMAGEJ. (2020). Image Processing and Analysis in Java. Retrieved November 04, 2020 from https://imagej.nih.gov/ij/download

KINDRATENKO, V.V. & VAN ESPEN, P.J.M. (2002). Classification of Irregularly Shaped Micro-Objects Using Complex Fourier Descriptors. Proceedings of the 13th International Conference on Pattern Recognition – ICPR96 2, 285-289. doi: 10.1109/ICPR.1996.546834

NIH. (2020). ImageJ. Retrieved November 08, 2020 from https://imagej.nih.gov

NUMXL. (2020). Excel Time Series Made Easier. Retrieved May 05, 2020 from https://www.numxl.com/

NUMXL. (2008). DFT. Retrieved January 21, 2020from https://www.numxl.com/support/documentation/numxl/reference-manual/spectral-analysis/dft?fbclid=IwAR2E0midOn5qa97UFczm2f3r6LoCkpT6pjuwGOHhXVIMAyCeTdC0UDdmB10

PAIS, A.J. (2015). Making Your Own Coffee Blend with Arabica and Robusta Beans. Retrieved July 02, 2019 from https://ecofriendlycoffee.org/making-coffee-blend-arabica-robusta-beans

PREEDY, V.R. (2014). Coffee in Health and Disease Prevention. United States: Academic Press.

TAKAHASHI, N., MAKI, H., NISHINA, H., & TAKAYAMA, K. (2013). Evaluation of Tomato Fruit Color Change with Different Maturity Stages and Storage Temperatures Using Image Analysis. 5th *IFAC Conference on Bio-Robotics*, *46*(4), 147-149. doi: https://doi.org/10.3182/20130327-3-JP-3017.00034 ■