Using machine learning techniques and different color spaces for the classification of Cape gooseberry (Physalis peruviana L.) fruits according to ripeness level

Carlos Cotrina [1], **Karen Bazán** [1], **Jimy Oblitas** [2], **Himer Avila-George** [Corresp., 3], **Wilson Castro** [Corresp. 1]

[1] Facultad de Ingeniería, Universidad Privada del Norte, Cajamarca, Cajamarca, Peru

[2] Departamento de Tecnología de Alimentos, Universidad de Lleida, Lleida, Cataluña, Spain

[3] Department of Computer Sience and Engineering, Universidad de Guadalajara, Ameca, Jalisco, Mexico

Corresponding Authors: Himer Avila-George, Wilson Castro
Email address: himer.avila@academicos.udg.mx, wilson.castro@upn.edu.pe

The classification of fresh fruits according to their ripeness is typically a subjective and tedious task; consequently, there is growing interest in the use of non-contact techniques such as those based on computer vision and machine learning. In this paper, we propose the use of non-intrusive techniques for the classification of Cape gooseberry fruits. The proposal is based on the use of machine learning techniques combined with different color spaces. Given the success of techniques such as artificial neural networks,support vector machines, decision trees, and K-nearest neighbors in addressing classification problems, we decided to use these approaches in this research work. A sample of 926 Cape gooseberry fruits was obtained, and fruits were classified manually according to their level of ripeness into seven different classes. Images of each fruit were acquired in the RGB format through a system developed for this purpose. These images were preprocessed, filtered and segmented until the fruits were identified. For each piece of fruit, the median color parameter values in the RGB space were obtained, and these results were subsequently transformed into the HSV and L*a*b* color spaces. The values of each piece of fruit in the three color spaces and their corresponding degrees of ripeness were arranged for use in the creation, testing, and comparison of the developed classification models. The classification of gooseberry fruits by ripening level was found to be sensitive to both the color space used and the classification technique, e.g., the models based on decision trees are the most accurate, and the models based on the L*a*b* color space obtain the best mean accuracy. However, the model that best classifies the cape gooseberry fruits based on ripeness level is that resulting from the combination of the SVM technique and the RGB color space.

Using machine learning techniques and different color spaces for the classification of Cape gooseberry *(Physalis peruviana L.)* fruits according to ripeness level

Carlos Cotrina[1], Karen Bazán[1], Jimy Oblitas[2], Himer Avila-George[3,4], and Wilson Castro[1,5]

[1]**Facultad de Ingeniería, Universidad Privada del Norte. Cajamarca, Cajamarca 06002, Peru.**
[2]**Escuela de Doctorado, Departamento de Tecnología de Alimentos, Universidad de Lleida. Lleida 25198, Spain.**
[3]**Centro Universitario de los Valles, Universidad de Guadalajara. Ameca, Jalisco 46600, Mexico.**
[4]**Unidad de Transferencia Tecnológica, CONACYT-CICESE. Tepic, Nayarit 63173, Mexico.**
[5]**Centro de Investigaciones e Innovaciones de la Agroindustria Peruana. Amazonas 1061, Peru.**

ABSTRACT

The classification of fresh fruits according to their ripeness is typically a subjective and tedious task; consequently, there is growing interest in the use of non-contact techniques such as those based on computer vision and machine learning. In this paper, we propose the use of non-intrusive techniques for the classification of Cape gooseberry fruits. The proposal is based on the use of machine learning techniques combined with different color spaces. Given the success of techniques such as *artificial neural networks*, *support vector machines*, *decision trees*, and *K-nearest neighbors* in addressing classification problems, we decided to use these approaches in this research work. A sample of 926 Cape gooseberry fruits was obtained, and fruits were classified manually according to their level of ripeness into seven different classes. Images of each fruit were acquired in the *RGB* format through a system developed for this purpose. These images were preprocessed, filtered and segmented until the fruits were identified. For each piece of fruit, the median color parameter values in the *RGB* space were obtained, and these results were subsequently transformed into the *HSV* and *L*a*b** color spaces. The values of each piece of fruit in the three color spaces and their corresponding degrees of ripeness were arranged for use in the creation, testing, and comparison of the developed classification models. The classification of gooseberry fruits by ripening level was found to be sensitive to both the color space used and the classification technique, e.g., the models based on decision trees are the most accurate, and the models based on the *L*a*b** color space obtain the best mean accuracy. However, the model that best classifies the cape gooseberry fruits based on ripeness level is that resulting from the combination of the SVM technique and the *RGB* color space.

Keywords: Cape gooseberry, artificial neural networks, support vector machines, decision trees, K-nearest neighbors

INTRODUCTION

The Cape gooseberry *(Physalis peruviana L.)*, known as the goldenberry in English-speaking countries and as aguaymanto in Peru, is a plant native to the South American Andes (Salazar et al., 2008; Luchese et al., 2015). This plant has attracted the interest of functional food markets (emerging markets of growing economic importance) due to its medicinal, nutritious, and pharmaceutical properties (Erkaya et al., 2012; Ramírez et al., 2013; Vásquez-Parra et al., 2013). The food industry needs to provide fruits of high and homogeneous quality, being necessary to improving their production methods to ensure high quality during manufacturing and commercialization (Benedito et al., 2006).

For fresh fruits, such as the Cape gooseberry, an important step in ensuring high quality is sorting, which is currently based on the visual inspection of color, size, and shape parameters (Zhang et al., 2014). However, visual inspection has certain disadvantages, as it is subjective, variable, tedious, laborious, inconsistent and easily influenced by the environment (Arakeri and Laksmana, 2016). Consequently, there is growing interest in reducing this subjectivity using innovative and non-contact measurements such as artificial vision systems. These types of systems measure the entire surface of the sample, making them more representative than colorimeters, which are based on point-to-point measurements (Chen et al., 2010; Romano et al., 2012; Sozer, 2016; Brosnan and Sun, 2004).

Computer vision systems (CVSs) are currently employed in the classification of horticultural products and in monitoring such products for defects and bruising (Romano et al., 2012). At present, the development of computer vision systems is focused on defining new methods for the evaluation of color and shape parameters. In this context, color receives special interest because it is an important sensory attribute providing necessary quality information for human perception. Consumers tend to prefer products that have a uniform appearance and vivid colors. Color has been closely associated with quality factors (ripeness, variety, and desirability) and food safety. Therefore, color is an essential element of classification for most food products (Castro et al., 2017; Oliveira et al., 2016; Avila et al., 2015; Wu and Sun, 2013).

Each color that humans can recognize in an image is a combination of the so-called primary colors, *red, green, and blue*, which can be arranged in a color space to facilitate the specification of colors in a standardized and widely accepted form. In essence, a color space is the specification of a three-dimensional coordinate system and a subspace of this scheme in which a single point represents each color. Nevertheless, there is more than one color space, and each color space can be classified into hardware-orientated spaces, human-orientated spaces, and instrumental spaces, as proposed by Wu and Sun (2013).

- *Hardware-orientated spaces*. These color spaces were defined based on the properties of the hardware devices used to reproduce the colors. In this category, the most popular color spaces are *RGB*, *YIQ*, and *CMYK*.

- *Human-orientated spaces*. These color spaces are based on hue-saturation. The most popular color spaces in this category are *HSI*, *HSL*, *HSV* and *HSB*. These spaces correspond to the concepts of tint, shade, and tone, which are specified by an artist based on inherent color characteristics. However, as with human vision, human-orientated spaces are not sensitive to small variations in color and are therefore not suitable for evaluating changes in product color during processing.

- *Instrumental spaces*. Color spaces such as *XYZ*, *L*a*b**, and *L*u*v** are used for color instruments. Unlike hardware-oriented spaces, which have different coordinates for the same color for various output media, the color coordinates of an instrumental space are the same on all output media.

The main features of the color parameters based on the works in Leon et al. (2006) and Zakaluk and Ranjan (2006) are detailed in Table 1. As can be seen, each color space was developed for a particular purpose; each color space has certain advantages when used in classification and identification problems.

Table 1. Color parameters used for classification.

Space	Parameter	Description
RGB	*R*	Red measured in digital image [0, 255]
	G	Green measured in digital image [0, 255]
	B	Blue measured in digital image [0, 255]
HSV	*H*	Hue derived from *RGB* [0, 360]
	S	Saturation derived from *RGB* [0, 100]
	V	Value derived from *RGB* [0, 100]
*L*a*b**	*L**	Luminosity derived from *RGB* [0, 100]
	*a**	Red/green opponent colors [-128, 127]
	*b**	Yellow/blue opponent colors [-128, 127]

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

Thus, although the systems directly provide information in the *RGB* space, some works, such as that realized by Du and Sun (2008), have aimed to determine whether there is any difference in the classification caused by the selected color space or by the utilized segmentation technique.

According to Wu and Sun (2013), "In the color measurement of food, the *L*a*b** color space is the most commonly used due to the uniform distribution of colors and because it is perceptually uniform."

In the image analysis process, pattern recognition is a qualitative analysis method, of which the supervised methods are most commonly used. Supervised learning is an automatic learning task that infers a function given labeled training data. In the fruit inspection industry, the support vector machine (SVM), k-nearest neighbor (KNN), artificial neural network (ANN), and decision tree (DT) pattern classification methods are the most commonly used (Arabasadi et al., 2013; Vithu and Moses, 2016).

The use of computer vision systems to determine ripeness level has been studied for a variety of fruits, including apples, bananas, blueberries, dates, mangoes, and tomatoes. Table 2 summarizes the main studies on fruit ripening using computer vision systems. However, for the Cape gooseberry, there are no reports about the use of image analysis for ripeness stage classification.

Table 2. Fruit/Vegetable ripening evaluation using expert system techniques in different color spaces.

Item	Color space	Processing method	Accuracy	Ref.
Apple	HSI	SVM	95	Xiaobo et al. (2007)
Apple	*L*a*b**	MDA	100	Cárdenas-Pérez et al. (2017)
Avocado	*RGB*	K-Means	82.22	Roa Guerrero and Meneses Benavides (2014)
Banana	*L*a*b**	LDA	98	Mendoza and Aguilera (2004)
Banana	*RGB*	ANN	96	Paulraj et al. (2009)
Blueberry	*RGB*	KNN and SK-Means	85-98	Li et al. (2014)
Date	*RGB*	K-Means	99.6	Pourdarbani et al. (2015)
Lime	*RGB*	ANN	100	Damiri and Slamet (2012)
Mango	*RGB*	SVM	96	Nandi et al. (2014)
Mango	*L*a*b** and HSB	MDA	90	Vélez-Rivera et al. (2014)
Mango	*L*a*b**	LS-SVM	88	Zheng and Lu (2012)
Oil palm	*L*a*b**	ANN	91.67	Fadilah et al. (2012)
Pepper	*HSV*	SVM	93.89	Elhariri et al. (2014)
Persimmon	*RGB* + *L*a*b**	QDA	90.24	Mohammadi et al. (2015)
Tomato	*HSV*	SVM	90.8	El-Bendary et al. (2015)
Tomato	*RGB*	DT	94.29	Goel and Sehgal (2015)
Tomato	*RGB*	LDA	81	Polder et al. (2002)
Tomato	*L*a*b**	ANN	96	Rafiq et al. (2016)
Watermelon	YCbCr	ANN	86.51	Shah Rizam et al. (2009)

Thus, we present a novel study classifying Cape gooseberry fruits using different color spaces and four of the leading supervised learning techniques. The principal objective is to determine which color space and which method of classification are the most appropriate for classifying Cape gooseberry fruits according to their level of ripeness.

MATERIALS AND METHODS

Cape gooseberry fruit samples

A sample of gooseberry fruits from a plantation located in El Faro village, Celendin Province, Cajamarca, Peru [UTM: -6.906469, -78.257071] was collected. This sample consisted of 926 Cape gooseberry fruits at different ripeness levels.

Computer vision system for grading Cape gooseberries

The hardware and software that constitute this system are described below.

- *Conveyor belt*. The conveyor belt is 160 cm long, 25 cm wide, and 80 cm high. The speed is adjustable, and the conveyor is operated by an EPLI motor (MS 632-4 60 Hz, 0.18 KW, 0.25 HP, 220 V, 1570 RPM).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

- *VGA webcam*. The utilized webcam has the following specifications:
  - Trademark: Halion
  - Model: HA-411
  - Resolution: 1280x720 pixels

  This webcam was located 35 cm above the sample.

  The internal walls of the CVS were painted black to avoid light leakage and exterior reflections of the room in a method similar to that realized by Pedreschi et al. (2006).

- *Lightning source*. A directional lighting system composed of two long fluorescent tubes (Philips TL-D Super, cold daylight, 80 cm, 36 W) was distributed symmetrically on both sides of the sample, and a circular fluorescent tube (Philips GX23 PH-T9, cold daylight, 21.6 cm, 22W) was located at the top.

- *Computer*. We used a laptop (Intel(R) Pentium(R) Dual-Core CPU T4200 @ 2.00 GHz and 3.0 GB RAM).

- *Informatics tool for data acquisition*. A computer tool was developed to control the acquisition of the images and their subsequent analysis. This tool was implemented using MATLAB.

Methodology

In subsequent sections, this methodology is described in detail.

Data extraction

In this step, information on the color parameters in three color spaces was collected from each fruit in the sample, and the sample was classified according to its ripeness grade.

Visual classification by color. Fruits in the samples were classified according to the ripening stage in seven states, similar to the method used by Bravo and Osorio (2016), using the surface color as indicated by the NTC 4580 sensor (Colombian Technical Normative) for Cape gooseberry and the visual scale proposed by Fischer et al. (2005); see Figure 1.

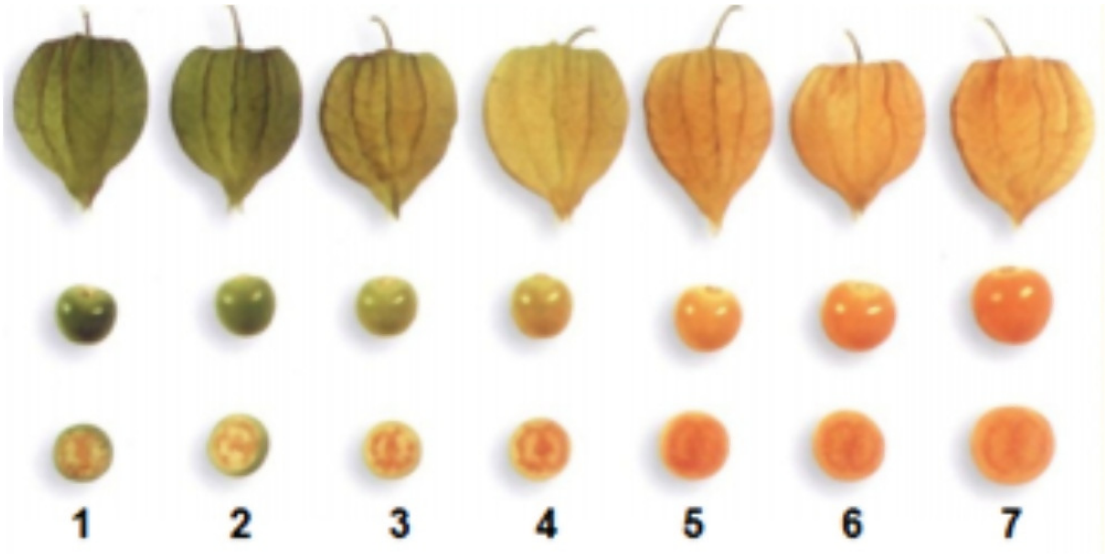

Figure 1. Ripeness states of Cape gooseberry.

Image acquisition and pre-processing. The steps in this stage, based partially on the methodology proposed by Arakeri and Laksmana (2016), are detailed below:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

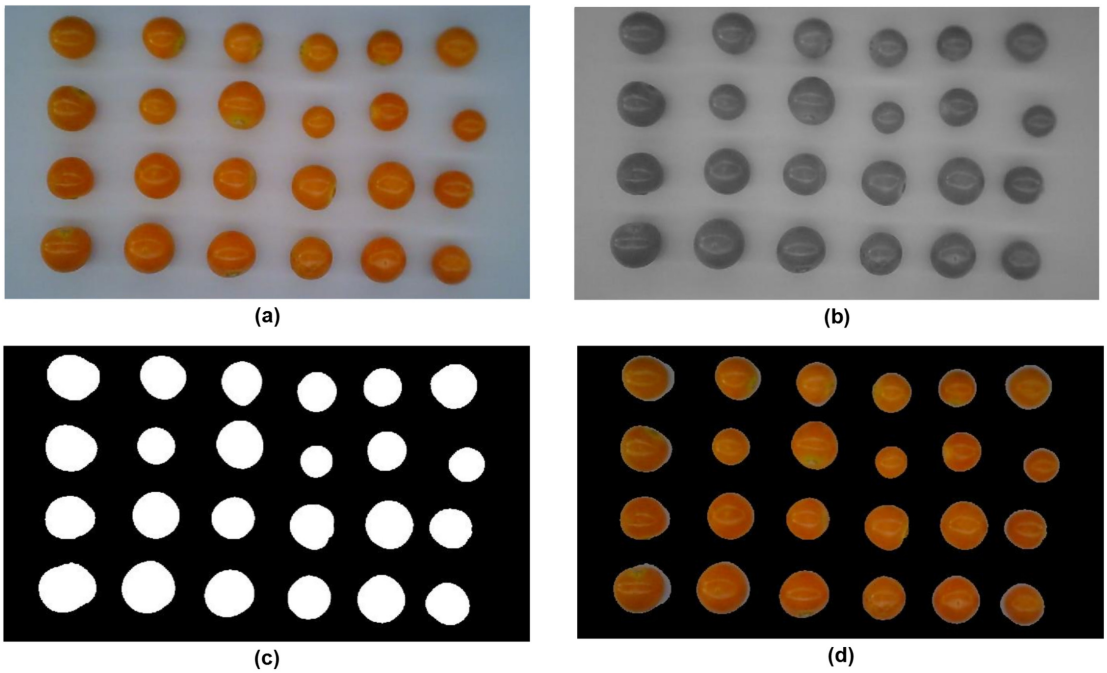

Figure 2. Images of acquisition and preprocessing stages.

- *Location of samples*. Fruits for each class, determined in the previous steep, were placed on the conveyor belt; these fruits were arranged in matrixial dispositions of four rows and between five to seven columns. The fruits were successively displayed on the conveyor belt until the entire class of fruits was completed.
- *Image acquisition*. Fruits were conveyed to the corresponding area for image acquisition. The software component discussed in Section was used to capture the images, and each image was stored (see Figure 2(a)) according to the corresponding class.
- *Image enhancement*. The images were enhanced by a Gaussian filter; the function shown in Eq. 1 was used for this purpose.

$$g(x,y) = \frac{1}{2\pi\sigma} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{1}$$

where
g = Filtered image
x, y = Position of pixel
σ = Standard deviation of Gaussian filter.

- *Segmentation*. The images were converted to grayscale, shown in Figure 2(c), and thresholding segmentation based on Eq. 2 was used. The results were images for which samples are isolated from background and their pixels identified, as shown in Figure 2(d).

$$h(x,y) = \begin{cases} 1 \text{ if } g(x,y) \geq T \\ 0 \text{ if } g(x,y) < T \end{cases} \tag{2}$$

where
h= Segmented image
x, y = Position of pixel
T = Threshold value.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

Obtaining space color parameters. From each fruit in the segmented images, shown in Figure 2(d), the median values of the color parameters in the *RGB* color space were determined similarly to the work of Blasco et al. (2007). Then, these median values were converted to *HSV* and *L*a*b** color spaces using the *rgb2hsv* and *rgb2lab* functions of the MATLAB software. All these values, linked to each fruit in the different classes, were stored in a database for the subsequent modeling stage.

Modeling and validation

The dataset obtained from the 926 fruits evaluated was used for the construction of the models and their subsequent validation. Then, to perform this stage, the dataset was divided into 70 percent (649 fruits) for modeling and 30 percent (277 fruits) for validation.

Modeling. In this stage, four supervised machine learning techniques were used for modeling; these techniques consider the categorical labels when data entries $x_1, x_2, \ldots, x_n$ must be assigned to predefined classes $C_1, C_2, \ldots, C_m$. In multi-class classification, the input is to be classified into only one of n non-overlapping classes. In the following paragraphs, each technique is detailed.

- *ANN*. This non-linear supervised classification method uses mathematical models to simulate biological neural networks. A common type of ANN is the radial basis function ANN (RBF-ANN), which is used for classifying into different classes by finding common features between samples of the known feature class. This type of network has nonlinearity embedded in the transfer functions of its hidden-layer neurons, making the optimization of tunable parameters a linear search (Dash et al., 2000; Kong et al., 2016). Figure 3a shows a schematic representation of this type of ANN, which was proposed by Beale et al. (2012).

  The Neural Network Toolbox in MATLAB was used for the implementation of the sorting models based on the ANN technique. Specifically, we used the *newpnn* function to create and train the RBF-ANN, and the *sim* function was used for the simulation stage.

- *DT*. This technique is a tree-based exemplification of the knowledge used to represent the classification rules. The internal nodes of a tree represent tests of an attribute; each branch represents the outcome of the test, and leaf nodes represent class labels. Traversing the branch from root to leaf node decodes the information enclosed in the form of if-then statements, and each branch leads to a single rule. Figure 3b shows a schematic of this technique, which was proposed by Safavian and Landgrebe (1991). Therefore, DT can be exploited to automatically generate the rules without the need of a human expert (Goel and Sehgal, 2015; Safavian and Landgrebe, 1991).

  For fitting and applying, the multi-class classifier function *fitctree* and *predict* from Matlab's *Machine Learning Toolbox* were used.

- *SVM*. SVM is a supervised non-parametric statistical learning technique that is widely used for classification by constructing a hyperplane or a set of hyperplanes in a high-dimensional space (Xiaobo et al., 2007; Nandi et al., 2014). Figure 3c shows the support vectors and the hyperplane separating the two classes, which are defined by squares and triangles.

  In this case, we use the *fitcecoc* function to fit models and the *predict* function to predict labels; both function are also implemented into Matlab's *Machine Learning* Toolbox

- *KNN*. KNN is a non-parametric classification technique cache of all the training data that predicts the response of the new sample by analyzing a certain number of the nearest neighbors in the feature space of the sample (Unay and Gosselin, 2007; Pourdarbani et al., 2015).

  Figure 3d shows an example of this technique. The element to be classified is the sun symbol. For $k = 3$, this object is classified as the triangle class since there is only one square and two triangles inside the circle that contains them. If $k = 9$, this object is classified as the square class since there are four triangles and five squares inside the outer circle.

  Again, we used MATLAB's Machine Learning Toolbox, where the *fitcknn* function was used to train the model and the *predict* function was used to predict the labels.

Validation. After obtaining the class for each fruit in the validation sample, the performance of each classifier and color space was determined using the confusion matrix. This technique is one of the most

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

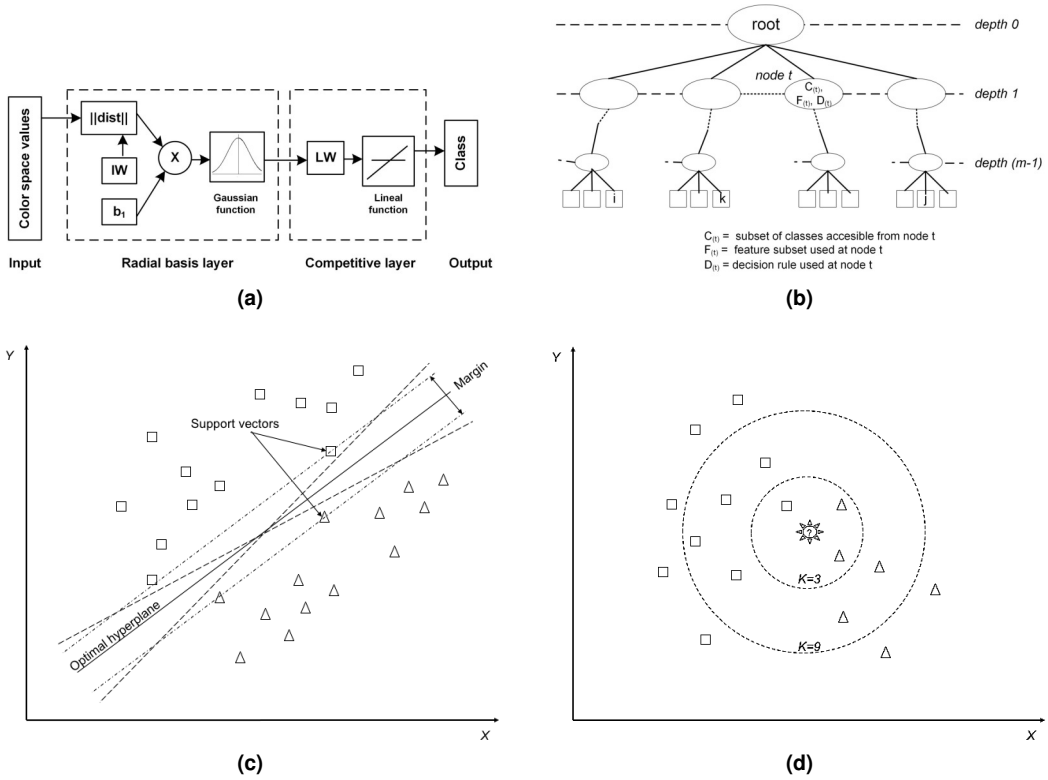

Figure 3. Four supervised machine learning techniques used in this work for modeling. (a) Generalized RBF-ANN structure; (b) Generalized DT structure; (c) SVM example; and (d) KNN example.

commonly used techniques in the machine learning community and contains information about the actual and predicted ratings obtained by a classification system.

A confusion matrix has two dimensions (real and predicted). Each row represents the instances of a real class, while each column represents the cases of a predicted class. Table 3 shows the basic form of the confusion matrix for a multi-class classification problem, where N_{ij} represents the number of samples that belong to class C_i but that are classified as class C_j.

Table 3. Generalized confusion matrix for several classes.

		Predicted class (C_j)				
		C_1	…	C_j	…	C_n
Actual class (C_i)	C_1	$N_{1,1}$	…	$N_{1,j}$	…	$N_{1,n}$
	⋮	⋮	⋮	⋮	⋮	⋮
	C_i	$N_{i,1}$	…	$N_{i,j}$	…	$N_{i,n}$
	⋮	⋮	⋮	⋮	⋮	⋮
	C_n	$N_{n,1}$	…	$N_{n,j}$	…	$N_{n,n}$

From the information contained in a confusion matrix, some performance measures can be defined, among them *accuracy*, *precision*, *recall* and *f-measure*. These measures are determined by classification errors made by the classifier. The terms true positive (TP), true negative (TN), false positive (FP) and false negative (FN) are counters used to contrast the class indicated by the classifier and the actual class, which are defined by Eqs. 3-6. The *positive* and *negative* terms refer to the classification produced by the classifier, while the expressions *true* and *false* refer to whether the classification is consistent with the actual value of the label.

According to Deng et al. (2016), *accuracy* is the proportion of the total number of predictions that

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

were correct and is defined by Eq. 7; *precision* is a measure of the accuracy provided that a specific class has been predicted and is defined by Eq. 8; *recall* is a measure of the ability of a prediction model to select instances of a certain class from a data set and is defined by Eq. 9; and *f-measure* is the harmonic mean of precision and recall and is defined by the Eq. 10.

$$TP_i = N_{ii} \tag{3}$$

$$FN_i = \sum k = 1^n N_{ik} - TP_i \tag{4}$$

$$FP_i = \sum_{k=1}^{n} N_{ki} - TP_i \tag{5}$$

$$TN_i = FN_i + FP_i \tag{6}$$

$$Accuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}} \tag{7}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{8}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{9}$$

$$F\text{-}Measure_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{10}$$

To measure the influence that color space may have on the results of a classification technique, extensive experimentation was carried out, in which the general precision obtained by the degree of ripening for Cape gooseberry fruits was compared for each combination of color parameters and classification techniques. For each of the twelve combinations of the proposed machine learning techniques (ANN, DT, SVM, and KNN) and the selected color spaces (*RGB*, *HSV*, and *L*a*b**), the performance measures were determined using Eqs. 7-10. In this work, we take the *f-measure* as the main measure for analysis.

RESULTS

Cape gooseberry color during ripening

The color parameter values for each color space are shown in Figure 4. As observed in Figure 4a, the parameter *R* presents an upward trend throughout the ripening process, starting at 75 and increasing to a maximum value of 150. In contrast, the parameter *G* begins at 89 and ends at 63, and the parameter *B* shows some variation between 0 and 45.

For the HSV color space, Figure 4b shows that the parameter *H* exhibits a downward trend starting at 0.19 and ending at 0.07. The parameter *S* shows a little variability, with median values fluctuating between 0.98 and 1. Finally, the parameter *V* exhibits an upward trend throughout the maturation process, with a minimum value of 0.35 and maximum of 0.60.

Concerning the *L*a*b** color space, Figure 4c shows that the parameter *L** presents slight variability, oscillating between 36.5 and 45.3. The parameter *a** has an upward trend that starts at -17.37 and reaches a maximum of 32.39. Finally, the parameter *b** fluctuates between 41.48 and 51.53.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

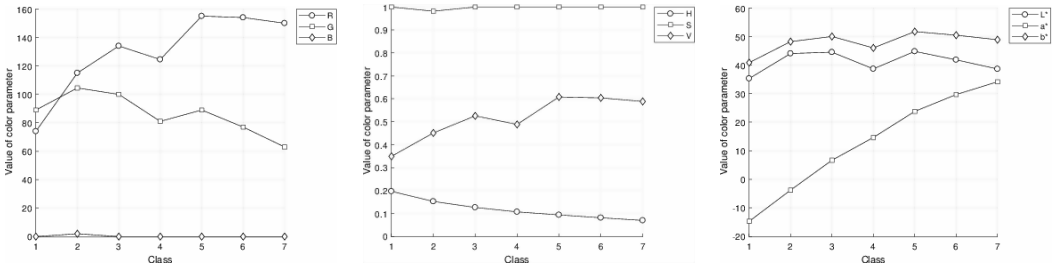

Figure 4. Color spaces for Cape gooseberry at different ripeness levels. (a) *RGB*; (b) *SVM*; and (c) $L*a*b*$.

Table 4. Color parameters in *L*a*b** space in ripeness stage.

Parameter	*Obtained*	*Source*	
		Vásquez-Parra et al. (2013)	Puente et al. (2011)
*L**	38.11 ± 3.89	54.04 ± 0.34	71.37 ± 1.10
*a**	32.40 ± 3.89	23.67 ± 0.30	15.20 ± 0.48
*b**	48.05 ± 3.81	59.85 ± 0.29	61.76 ± 1.34

Mean ± standard deviation.

Model evaluation

The results of the twelve classification models proposed in this research are shown in Tables 5-8, where each table groups the results by machine learning techniques.

Table 5 shows the results of the evaluation of the models based on the ANN technique; obtaining accuracy between 50.18% and 75.75% for *L*a*b** and *HSV* color-based models.

Table 5. Confusion matrix for ANN model using three color spaces.

RGB	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	**Metrics**	
Class 1	30	1	0	0	0	0	0	Precision	69,97 %
Class 2	13	17	5	0	0	0	0	Recall	74,97 %
Class 3	10	3	18	2	1	0	0	Specificity	76,67 %
Class 4	5	0	3	19	5	0	0	F-Measure	76,44 %
Class 5	8	0	0	3	33	6	0		
Class 6	3	0	0	0	11	27	2		
Class 7	6	0	0	0	0	12	34		
HSV									
Class 1	31	0	0	0	0	0	0	**Metrics**	
Class 2	3	19	10	1	2	0	0	Precision	51,21 %
Class 3	0	4	10	2	16	0	2	Recall	44,55 %
Class 4	0	4	0	14	5	0	9	Specificity	49,38 %
Class 5	0	0	0	0	34	0	16	F-Measure	56,38 %
Class 6	0	1	0	0	28	0	14		
Class 7	0	0	0	2	19	0	31		
L*a*b*									
Class 1	31	0	0	0	0	0	0	**Metrics**	
Class 2	0	23	11	1	0	0	0	Precision	76,73 %
Class 3	0	2	28	4	0	0	0	Recall	73,37 %
Class 4	0	0	6	18	7	1	0	Specificity	69,65 %
Class 5	1	0	2	3	37	7	0	F-Measure	72,17 %
Class 6	0	0	0	2	9	28	4		
Class 7	0	0	0	1	0	7	44		

Table 6 shows the results obtained by the DT technique. The DT model based on the *L*a*b** color space achieves better accuracy (75.09%). Although, the results of the models based on the *RGB* and *SVM* color spaces achieved accuracies very close to those obtained by the model based on the *L*a*b** color

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

space ($RGB = 72.92$ and $HSV = 74.37$).

Table 6. Confusion matrix for DT model using three color spaces.

RGB	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7		
Class 1	31	0	0	0	0	0	0	**Metrics**	
Class 2	1	26	8	0	0	0	0	Precision	74,45 %
Class 3	0	4	26	3	1	0	0	Recall	70,71 %
Class 4	0	0	6	21	3	2	0	Specificity	73,40 %
Class 5	0	1	3	7	27	12	0	F-Measure	78,36 %
Class 6	0	0	1	1	6	30	5		
Class 7	0	0	0	1	0	10	41		
HSV									
Class 1	31	0	0	0	0	0	0	**Metrics**	
Class 2	0	24	11	0	0	0	0	Precision	76,00 %
Class 3	0	3	27	3	1	0	0	Recall	72,00 %
Class 4	0	0	5	19	7	1	0	Specificity	73,85 %
Class 5	0	0	1	2	40	7	0	F-Measure	79,56 %
Class 6	0	0	0	1	10	21	11		
Class 7	0	0	0	0	0	8	44		
L*a*b*									
Class 1	31	0	0	0	0	0	0	**Metrics**	
Class 2	0	22	12	0	1	0	0	Precision	77,35 %
Class 3	0	1	24	5	4	0	0	Recall	73,58 %
Class 4	0	0	5	21	6	0	0	Specificity	69,16 %
Class 5	0	0	5	4	34	7	0	F-Measure	73,41 %
Class 6	0	0	0	0	9	28	6		
Class 7	0	0	0	0	0	4	48		

Table 7 presents the results of the SVM models In this the SVM model based on the *RGB* color space obtained the best accuracy (79.42%). The model based on the *L*a*b** color space obtained an accuracy of 77.62%, which is very close to that achieved by the model based on *RGB*. However, the model based on the *HSV* color space yielded poor results ($accuracy = 33.57\%$).

Table 7. Confusion matrix for SVM model using three color spaces.

RGB	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7		
Class 1	31	0	0	0	0	0	0	**Metrics**	
Class 2	0	24	11	0	0	0	0	Precision	80,82 %
Class 3	0	1	28	4	1	0	0	Recall	77,62 %
Class 4	0	0	6	22	3	1	0	Specificity	78,29 %
Class 5	0	0	1	6	39	4	0	F-Measure	76,47 %
Class 6	0	0	0	0	7	31	5		
Class 7	0	0	0	0	0	7	45		
HSV									
Class 1	18	13	0	0	0	0	0	**Metrics**	
Class 2	0	23	0	0	0	0	12	Precision	24,93 %
Class 3	0	4	0	0	1	0	29	Recall	12,42 %
Class 4	0	6	0	0	0	0	26	Specificity	20,75 %
Class 5	0	0	0	0	0	0	50	F-Measure	36,19 %
Class 6	0	0	0	0	0	0	43		
Class 7	0	0	0	0	0	0	52		
L*a*b*									
Class 1	31	0	0	0	0	0	0	**Metrics**	
Class 2	0	25	10	0	0	0	0	Precision	79,06 %
Class 3	0	2	28	4	0	0	0	Recall	75,57 %
Class 4	0	0	6	19	4	3	0	Specificity	72,17 %
Class 5	0	0	1	4	39	6	0	F-Measure	74,66 %
Class 6	0	0	0	0	7	30	6		
Class 7	0	0	0	0	0	9	43		

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

As seen in Table 8, the model based on the *L*a*b** color space obtained the best accuracy (75.81). However, the results of the models based on the *RGB* and *SVM* color spaces reported accuracy results very close to those obtained by the model based on the *L*a*b** color space (*RGB* = 72.56 and *HSV* = 72.92).

Table 8. Confusion matrix for KNN model using three color spaces.

RGB	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7		
Class 1	30	1	0	0	0	0	0	**Metrics**	
Class 2	2	22	10	1	0	0	0	Precision	74,66 %
Class 3	0	4	25	3	2	0	0	Recall	71,47 %
Class 4	0	1	4	21	6	0	0	Specificity	74,52 %
Class 5	0	0	0	3	40	7	0	F-Measure	76,86 %
Class 6	0	0	0	1	12	28	2		
Class 7	0	0	0	1	0	16	35		
HSV									
Class 1	31	0	0	0	0	0	0	**Metrics**	
Class 2	1	25	9	0	0	0	0	Precision	74,08 %
Class 3	0	7	21	2	4	0	0	Recall	70,28 %
Class 4	0	0	4	22	5	1	0	Specificity	73,93 %
Class 5	0	0	0	4	39	7	0	F-Measure	78,97 %
Class 6	0	0	0	1	13	23	6		
Class 7	0	0	0	0	0	11	41		
L*a*b*									
Class 1	31	0	0	0	0	0	0	**Metrics**	
Class 2	0	23	11	1	0	0	0	Precision	77,26 %
Class 3	0	2	28	4	0	0	0	Recall	73,47 %
Class 4	0	0	6	18	7	1	0	Specificity	69,76 %
Class 5	0	0	2	3	38	7	0	F-Measure	72,31 %
Class 6	0	0	0	2	9	28	4		
Class 7	0	0	0	1	0	7	44		

In Figure 5, a summary of the results of this research work is shown, primarily using the *f-measure* for analysis. Each of the twelve proposed classification models obtained different f-measures. The SVM model with the *HSV* color space exhibited the worst results (*f-measure* = 24.24%). The models based on the KNN and DT techniques yielded good results regardless of the color space used (*f-measure* > 73%); between these two techniques, the DT model obtained slightly better results. Additionally, the models based on the *L*a*b** color space give good results (*f-measure* > 75%).

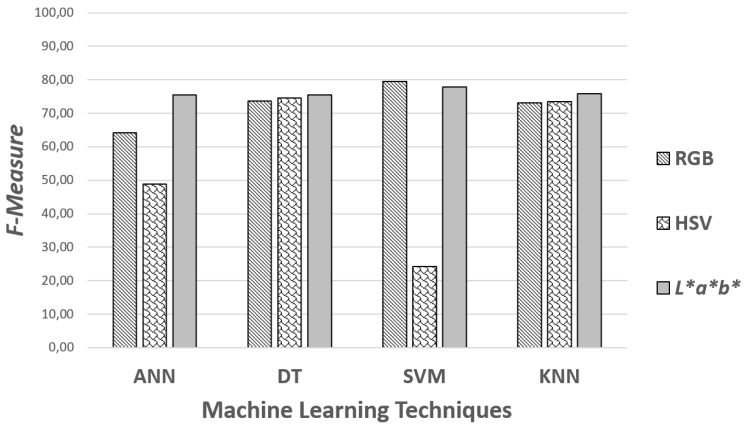

Figure 5. F-measure analysis.

Finally, for the classification of Cape gooseberry fruits by level of ripeness, the best model is that based on the SVM technique and the *RGB* color space. This model obtains an *f-measure* = 79.58%.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

DISCUSSION

As explained in Itle and Kabelka (2009) and Cárdenas-Pérez et al. (2017), changes in the parameters L^*, a^* and b^* are associated with increases in carotenoid levels and a loss of chlorophyll in the pericarp. In this sense Table 4 compares the median values for the $L^*a^*b^*$ space obtained in this work against Puente et al. (2011) and Vásquez-Parra et al. (2013).The differences in the $L^*a^*b^*$ parameters obtained in this research and previous reports are because the cultivar, ripeness stage or cultivation procedure for each sample was different in each study, as suggested by Oliveira et al. (2016).

Understanding that changes in the different color parameters are related to ripening stage is needy to evaluate the accuracy of each technique according to color space used for a classifier, so:

- *ANN.* This technique has already been successfully used to classify fruits according to their level of ripeness. Some examples of the use of this technique can be found in the following works: Paulraj et al. (2009), Damiri and Slamet (2012), Fadilah et al. (2012), and Shah Rizam et al. (2009). As can be seen in Table 5, the accuracy of the ANN models is significantly influenced by the chosen color space. The ANN model based on the $L^*a^*b^*$ color space obtained suitable accuracy (75.45%) in the Cape gooseberry fruit classification, which agrees with the results obtained by Fadilah et al. (2012).

- *DT.* It was successfully used by Goel and Sehgal (2015) to classify tomatoes according to their ripeness level using *RGB* color space and was capable of classifying the fruits with an accuracy of 94.29%. In our case, in addition to the *RGB* color space-based model, we built models for the *HSV* and $L^*a^*b^*$ color spaces, observing that color space slightly influences the quality of the results.

- *SVM.* The SVM technique has been used by Xiaobo et al. (2007) to classify apples using the *HSI* color space. In combination with the *RGB* color space, Nandi et al. (2014) used SVM to classify pieces of mango fruit. In both studies, it was possible to classify the fruits with an accuracy greater than 95%. In our case the best accuracy was 79.42% for *RGB* space color model and the poor ($accuracy = 33.57\%$) for the *HSV* color space.

- *KNN.* Many studies have achieved results with excellent levels of accuracy; e.g., Unay and Gosselin (2007) classified apple stems with an accuracy of up to 99%. Regarding problems related to the classification of fruits according to their maturation degree, Li et al. (2014) presented the results of their study on identifying blueberries in different stages of growth. Among the classification models constructed in their work, the model based on the KNN technique obtained the best accuracy (86%) using the *RGB* color space. Our results using $L^*a^*b^*$ color spaces shown the best accuracy (75.81 %), and similar results for *RGB* and *SVM* ($RGB = 72.56$ and $HSV = 72.92$).

As shown in Tables 5-8, the classification of Cape gooseberry fruits by the degree of ripeness is sensitive to both the color space used and the classification technique used. In this sense, the mean accuracy obtained for *RGB*, *HSV*, and $L^*a^*b^*$ were 72.29%, 57.76%, and 75.99% respectively. These results are similar to those reported by Blasco et al. (2007), who used the LDA classifier and found that *RGB* and $L^*a^*b^*$ present similar and slightly higher accuracy values compared to those obtained in $L^*u^*v^*$ space.

CONCLUSIONS

The purpose of this research was to develop a non-intrusive system for classifying gooseberry fruits according to their degree of maturity. Twelve classification models were developed. These models were the result of combining four machine learning techniques (ANN, KNN, DT, and SVM) and three color spaces (*RGB, HSV, and* $L^*a^*b^*$).

The choice of color space was found to influence the accuracy of the sorting systems, and this dynamic is observed mainly in the models based on the ANN and SVM techniques. Meanwhile, the models based on the KNN and DT techniques yielded good results regardless of the color space used. On the other hand, the models based on the $L^*a^*b^*$ color space produced good results regardless of the machine learning technique employed. However, the classifier developed from the *SVM* technique and *RGB* color space gave the best performance in terms of *accuracy* and *f-measure* ratios.

Future works should evaluate the use of different color spaces associatively to determine whether synergies exist in the ripening classification process.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

REFERENCES

Arabasadi, Z., Khorasani, M., Akhlaghi, S., Fazilat, H., Gedde, U., Hedenqvist, M., and Shiri, M. (2013). Prediction and optimization of fireproofing properties of intumescent flame retardant coatings using artificial intelligence techniques. *Fire Saf. J.*, 61:193–199.

Arakeri, M. and Laksmana (2016). Computer vision based fruit grading system for quality evaluation of tomato in agriculture industry. *Procedia Comput. Sci.*, 79:426–433.

Avila, F., Mora, M., Oyarce, M., Zuñiga, A., and Fredes, C. (2015). A method to construct fruit maturity color scales based on support machines for regression: Application to olives and grape seeds. *J. Food Eng.*, 162:9 – 17.

Beale, M. H., Hagan, M. T., and Demuth, H. B. (2012). Neural network toolbox^{TM}user's guide. In *R2012a*, pages 201–213. The MathWorks, Inc., USA.

Benedito, J., Simal, S., Clemente, G., and Mulet, A. (2006). Manchego cheese texture evaluation by ultrasonics and surface probes. *Int. Dairy J.*, 16(5):431 – 438.

Blasco, J., Aleixos, N., Gómez, J., and Moltó, E. (2007). Citrus sorting by identification of the most common defects using multispectral computer vision. *J. Food Eng.*, 83(3):384–393.

Bravo, K. and Osorio, E. (2016). Characterization of polyphenol oxidase from cape gooseberry (physalis peruviana l.) fruit. *Food Chem.*, 197:185–190.

Brosnan, T. and Sun, D.-W. (2004). Improving quality inspection of food products by computer vision — a review. *J. Food Eng.*, 61(1):3–16.

Cárdenas-Pérez, S., Chanona-Pérez, J., Méndez-Méndez, J. V., Calderón-Domínguez, G., López-Santiago, R., Perea-Flores, M. J., and Arzate-Vázquez, I. (2017). Evaluation of the ripening stages of apple (Golden Delicious) by means of computer vision system. *Biosyst. Eng.*, 159:46 – 58.

Castro, W., Oblitas, J., Chuquizuta, T., and Avila-George, H. (2017). Application of image analysis to optimization of the bread-making process based on the acceptability of the crust color. *J. Cereal Sci.*, 74:194–199.

Chen, K., Sun, X., Qin, C., and Tang, X. (2010). Color grading of beef fat by using computer vision and support vector machine. *Comput. Electron. Agric.*, 70(1):27–32.

Damiri, D. J. and Slamet, C. (2012). Application of image processing and artificial neural networks to identify ripeness and maturity of the lime (citrus medica). *Int. J. Basic Appl. Sci.*, 01(02):171–179.

Dash, P. K., Mishra, S., and Panda, G. (2000). A radial basis function neural network controller for upfc. *IEEE Trans. Power Syst.*, 15(4):1293–1299.

Deng, X., Liu, Q., Deng, Y., and Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Inf. Sci.*, 340:250–261.

Du, C.-J. and Sun, D.-W. (2008). Multi-classification of pizza using computer vision and support vector machine. *J. Food Eng.*, 86(2):234–242.

El-Bendary, N., El Hariri, E., Hassanien, A. E., and Badr, A. (2015). Using machine learning techniques for evaluating tomato ripeness. *Expert Syst. Appl.*, 42(4):1892–1905.

Elhariri, E., El-Bendary, N., Hussein, A. M. M., Hassanien, A. E., and Badr, A. (2014). Bell pepper ripeness classification based on support vector machine. In *IEEE International Conference on Engineering and Technology*, pages 1–6.

Erkaya, T., Dağdemir, E., and Şengül, M. (2012). Influence of cape gooseberry (physalis peruviana l.) addition on the chemical and sensory characteristics and mineral concentrations of ice cream. *Food Res. Int.*, 45(1):331–335.

Fadilah, N., Mohamad-Saleh, J., Abdul Halim, Z., Ibrahim, H., and Syed Ali, S. S. (2012). Intelligent color vision system for ripeness classification of oil palm fresh fruit bunch. *Sensors*, 12(10):14179–14195.

Fischer, G., Miranda, D., Piedrahita, W., and Romero, J. (2005). *Avances en cultivo, poscosecha y exportación de la uchuva (physalis peruviana L.) en Colombia*. Universidad Nacional de Colómbia.

Goel, N. and Sehgal, P. (2015). Fuzzy classification of pre-harvest tomatoes for ripeness estimation – an approach based on automatic rule learning using decision tree. *Appl. Soft Comput.*, 36:45 – 56.

Itle, R. A. and Kabelka, E. A. (2009). Correlation between *l*a*b** color space values and carotenoid content in pumpkins and squash (cucurbita spp.). *HortScience*, 44(3):633–637.

Kong, C., Wang, H., Li, D., Zhang, Y., Pan, J., Zhu, B., and Luo, Y. (2016). Quality changes and predictive models of radial basis function neural networks for brined common carp (Cyprinus carpio) fillets during frozen storage. *Food Chem.*, 201:327–333.

Leon, K., Mery, D., Pedreschi, F., and Leon, J. (2006). Color measurement in l⋆ a⋆ b⋆ units from rgb

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

digital images. *Food Res. Int.*, 39(10):1084–1091.

Li, H., Lee, W., and Wang, K. (2014). Identifying blueberry fruit of different growth stages using natural outdoor color images. *Comput. Electron. Agric.*, 106:91–101.

Luchese, C. L., Gurak, P. D., and Marczak, L. D. F. (2015). Osmotic dehydration of physalis (physalis peruviana l.): Evaluation of water loss and sucrose incorporation and the quantification of carotenoids. *LWT-Food Sci. Technol.*, 63(2):1128–1136.

Mendoza, F. and Aguilera, J. (2004). Application of image analysis for classification of ripening bananas. *J. Food Sci.*, 69(9).

Mohammadi, V., Kheiralipour, K., and Ghasemi-Varnamkhasti, M. (2015). Detecting maturity of persimmon fruit based on image processing technique. *Sci. Hortic.*, 184:123–128.

Nandi, C. S., Tudu, B., and Koley, C. (2014). A machine vision-based maturity prediction system for sorting of harvested mangoes. *IEEE Trans. Instrum. Meas.*, 63(7):1722–1730.

Oliveira, S. F., Gonçalves, F. J. A., Correia, P. M. R., and Guiné, R. P. F. (2016). Physical properties of *physalis peruviana l. Open Agric.*, 1(1):55–59.

Paulraj, M., Hema, C. R., R. Pranesh, K., and Siti Sofiah, M. R. (2009). Color recognition algorithm using a neural network model in determining the ripeness of a banana. In *Proceedings of the International Conference on Man-Machine Systems*, pages 2B71–2B74. Universiti Malaysia Perlis.

Pedreschi, F., León, J., Mery, D., and Moyano, P. (2006). Development of a computer vision system to measure the color of potato chips. *Food Res. Int.*, 39(10):1092–1098.

Polder, G., van der Heijden, G. W. A. M., and Young, I. T. (2002). Spectral image analysis for measuring ripeness of tomatoes. *Trans. ASAE*, 45(4):1155.

Pourdarbani, R., Ghassemzadeh, H., Seyedarabi, H., Nahandi, F., and Vahed, M. (2015). Study on an automatic sorting system for date fruits. *J. Saudi Society Agric. Sci.*, 14(1):83–90.

Puente, L. A., Pinto-Muñoz, C. A., Castro, E. S., and Cortés, M. (2011). Physalis peruviana linnaeus, the multiple properties of a highly functional fruit: A review. *Food Res. Int.*, 44(7):1733–1740.

Rafiq, A., Makroo, H. A., and Hazarika, M. K. (2016). Artificial neural network-based image analysis for evaluation of quality attributes of agricultural produce. *J. Food Process Preserv.*, 40(5):1010–1019.

Ramírez, F., Fischer, G., Davenport, T. L., Pinzón, J. C. A., and Ulrichs, C. (2013). Cape gooseberry (physalis peruviana l.) phenology according to the bbch phenological scale. *Sci. Hortic.*, 162:39–42.

Roa Guerrero, E. and Meneses Benavides, G. (2014). Automated system for classifying hass avocados based on image processing techniques. In *IEEE Colombian Conference on Communications and Computing*, pages 1–6.

Romano, G., Argyropoulos, D., Nagle, M., Khan, M. T., and Müller, J. (2012). Combination of digital images and laser light to predict moisture content and color of bell pepper simultaneously during drying. *J. Food Eng.*, 109(3):438–448.

Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.*, 21(3):660–674.

Salazar, M. R., Jones, J. W., Chaves, B., and Cooman, A. (2008). A model for the potential production and dry matter distribution of cape gooseberry (physalis peruviana l.). *Sci. Hortic.*, 115(2):142–148.

Shah Rizam, M. S. B., Farah Yasmin, A. R., Ahmad Ihsan, M. Y., and Shazana, K. (2009). Non-destructive watermelon ripeness determination using image processing and artificial neural network. *Int. J. Electr. Comput. Eng.*, 3(2):332–336.

Sozer, N. (2016). *Imaging Technologies and Data Processing for Food Engineers*. Springer.

Unay, D. and Gosselin, B. (2007). Stem and calyx recognition on 'jonagold' apples by pattern recognition. *J. Food Eng.*, 78(2):597 – 605.

Vásquez-Parra, J. E., Ochoa-Martínez, C. I., and Bustos-Parra, M. (2013). Effect of chemical and physical pretreatments on the convective drying of cape gooseberry fruits (physalis peruviana). *J. Food Eng.*, 119(3):648–654.

Vélez-Rivera, N., Blasco, J., Chanona-Pérez, J., Calderón-Domínguez, G., Perea-Flores, M. d. J., Arzate-Vázquez, I., Cubero, S., and Farrera-Rebollo, R. (2014). Computer vision system applied to classification of "manila" mangoes during ripening process. *Food Bioprocess Technol.*, 7(4):1183–1194.

Vithu, P. and Moses, J. A. (2016). Machine vision system for food grain quality evaluation: A review. *Trends Food Sci. Technol.*, 56:13–20.

Wu, D. and Sun, D.-W. (2013). Colour measurements by computer vision for food quality control–a review. *Trends Food Sci. Technol.*, 29(1):5–20.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018

Xiaobo, Z., Jiewen, Z., and Yanxiao, L. (2007). Apple color grading based on organization feature parameters. *Pattern Recognit. Lett.*, 28(15):2046 – 2053.

Zakaluk, R. and Ranjan, R. S. (2006). Artificial neural network modelling of leaf water potential for potatoes using rgb digital images: a greenhouse study. *Potato Res.*, 49(4):255–272.

Zhang, B., Huang, W., Li, J., Zhao, C., Fan, S., Wu, J., and Liu, C. (2014). Principles, developments and applications of computer vision for external quality inspection of fruits and vegetables: A review. *Food Res. Int.*, 62:326–343.

Zheng, H. and Lu, H. (2012). A least-squares support vector machine (LS-SVM) based on fractal analysis and CIELab parameters for the detection of browning degree on mango (Mangifera indica L.). *Comput. Electron. Agric.*, 83:47–51.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26691v1 | CC BY 4.0 Open Access | rec: 14 Mar 2018, publ: 14 Mar 2018