Journal of Engineering Science and Technology
Vol. 16, No. 5 (2021) 3923 - 3942

ALMOND KERNEL VARIETY IDENTIFICATION AND CLASSIFICATION USING DECISION TREE

NARENDRA V.G.[1], KRISHANAMOORTHI M.[1,*], SHIVAPRASAD G.[1], AMITKUMAR V.G.[1], PRIYA KAMATH[1]

Department of Computer Science and Engineering,
Manipal Institute of Technology, Manipal Academy of Higher Education,
Manipal,Udupi District, Karnataka State, India-576104
*Corresponding Author: k.moorthi@manipal.edu

Abstract

Nowadays, an identification system is needed in agriculture and food processing industries to boost the efficiency of production to meet international standards. The manual approach is used in product grading and quality control. Unfortunately, it leads to uneven products, higher time expenses, and fatigue by human operators. So, quality assessment is one of the significant factors for products and a great impact on the final prices. In this study, we have proposed an image processing and computational intelligence-based method for identifying and classifying almond varieties as Nonpareil (NP), Mission (MI), Carmel (CR), and California (C). The scanner is used to obtain a kernel image for 2000 samples of almond. The proposed system involves four stages, they are pre-processing (used a median filter to eliminate noise and Otsu threshold algorithm used for segmentation), feature extraction (total 66 features- 4 for geometric; 10 for shape; 37 for colour, and 15 for texture), feature selection and reduction (principal component analysis-PCA and modified sequential floating forward selection-MSFFS), and classification (decision tree). We used three strategies in classification, they are strategy-1 (considered whole features set without feature selection and reduction applied to DT), strategy-2 (considered whole features set with PCA), and strategy-3 (considered whole features set with MSFFS). Overall accuracy is obtained from DT as 80.8% for strategy-1, 90.8% for strategy-2, and 97.13% for strategy-3. Among all, strategy-3 (DT with MSFFS) is outperformed for the classification of almonds kernel variety. The developed method can be easily extended to online sorting machines.

Keywords: Almond kernel variety, Colour, Decision tree, Geometric, Image processing, Intelligent systems, Modified sequential floating forward selection, Principal component analysis, Shape, Texture.

1. Introduction

Almonds are the most nutritious of all nuts. The almonds are incredibly high in protein content and rich in vitamin E, zinc, phosphorus, magnesium, iron, and calcium. Worldwide, almonds are one of the leading nut crops that produce high commercial value. Almond's production in India during 2017–18 was 4000 MT. The almond farming in India was confined to selected hilly areas of Jammu and Kashmir, Uttar Pradesh, and Himachal Pradesh. The types/varieties of almonds found are Kashmiri, Gurbandi, Mamra, Nonpareil, Mission, Carmel, California, etc. Also, almond varieties have been identified through the number of almonds per ounce (18/20, 20/22, 23/25, 25/27, 27/30, 30/32, 32/34, and 34/36) [1]. India is one of the primary producers and exporters of quality almonds and contributes 6 percent of total exports in the world. Also, foreign exchange is earning significantly for the country. So, kernel quality should compete in the international market according to the standards specified by the United Nations Economy Commission for Europe (UNECE) and the United States Department of Agriculture (USDA).

Promoting agricultural products as a standard product variety, there are several factors to play an important role, such as growing, planting, and harvesting. Among these, postharvest is the most essential and directly related to the process of improving the sorting or grading of the product. Improving the quality of new and used almond's reliability strategy is a critical factor in exports and the economic profitability of the final product. Therefore, for implementing such operations, both Computer Vision (CV) and Human Vision (HV) are being used. The HV-based methods are less attractive because of low speed, high costs, and require skilled labour for high accuracy. In the past few years, CV-based applications have used advanced techniques for agricultural and food product sorting and grading with its high efficiency, low cost, and higher speeds [2-9].

In India, most almond-processing industries are practicing manual grading. Except for some mechanical grading tools, still, a labour-intensive manual process is to sort and grade kernel quality. Due to being influenced by physical factors, the manual process is inducing subjective assessment leading to wavering the results. Hence, the production cost is high because of a lack of skilled labour. So, the development of an intelligent system is needed for grading the almond kernel automatically. Hence, we needed a standard image database as a benchmark. Currently, it is not available in India. Hence, initially, this study proposes to create a standard image database for each kernel variety. Next, to maintain international quality standards, we focused on the development of the system for kernel grading using image processing and machine learning technique.

2. Related Work

Image processing and computer vision techniques have been reported in various studies for grain seed analysis. Chen et al. [10] proposed the classification of corn into five varieties. A flatbed scanner was used to acquire non-touching corn kernels. The different types of features (morphological, colour, etc.) extracted from corn kernel image and discriminant analysis were used to get an optimized feature subset. Finally, discriminant analysis and neural network with backpropagation both were combined and used for corn type identification.

The RGB and HSV colour spaces are used to extract colour features from the acquired images of corn seeds. GLCM was used to calculate the texture features and local binary pattern (LBP) and resulted in 5%-6%. More results that are accurate were obtained to compare these features for corn seed classification. Also, in other research, the developed methods have been evaluated two ways to discriminate the wheat into five classes. The first method used individual features such as colour, shape, and texture. The second method used a combination of features such as colour, shape, and texture. The results showed an overall accuracy of 96% by the second method [11-12].

The 11 wheat grain varieties were discriminated against and proposed by Zapotoczny. [13], using surface colour and texture features. The methods of feature reduction were used to constitute an optimized feature set. Finally, the performance was evaluated using different classifiers. The discrimination of chestnut into five classes used the combination of colour (RGB, HSV, CIEL*a*b*), shape, and texture features [14]. Mollazade et al. [15] and Yu et al. [16] extracted the colour features using RGB, Nrgb, and HSI colour spaces, and the calculated texture features from the grey-level co-occurrence matrix (GLCM) for raisins and finally reported the appropriateness of the extracted features. The obtained results show the competence of support vector machines (SVMs) and artificial neural networks (ANNs) for raisins classification into four types.

Colour and geometry are the major sources of food and agricultural commodity inspection, grade, and sort [3]. In several agricultural and foods product, CV-based applications are successfully used to classify or identify quality factors such as colour and size including soybean seeds [17], Coffee [18], dry beans [19], pistachios [20], and Peanuts [21, 22]. Progress in hardware, and image processing-popularized computer vision techniques for automated kernel quality verification of almond parameters such as colour and size. Currently, in the oil processing industry, seed classification, grading, and quality evaluation used CV technology and results in high accuracy compared to those based on HV [23-25]. The agricultural and food product grading, as well as quality evaluation, were done developed CV methods, which use the combined features (shape, colour, and texture) set. For proper performance, there is no guarantee for the categorization of various agricultural products using a feature set including shape, colour, and texture. Therefore, it is suggested to combine different elements to get appropriate classification with high accuracy in various categories of products [26, 27].

Five varieties of rice were identified using morphology and colour features. For classification, classifiers such as neuro-fuzzy and multilayer perceptron were trained and tested with extracted features. Finally, it considered averaged the obtained accuracy. The accuracy of kernel-level rice identification was 99% [28] using a near-infrared spectrophotometer (NIRS). Zou et al. [29] have researched to classify rapeseed varieties. The features were extracted from the kernel-level image and reduce the extracted features using principal component analysis. The back-propagation neural network and distance discriminant analysis classifiers have been used to discriminate against the rapeseed varieties (five), and obtained accuracy was 100%.

Regarding this point that no comprehensive research on the quality grading of almond products has been conducted so far. Therefore, this research developed a robust method based on image processing and computational intelligence for quality grading and classification of this product. The kernels are graded into four classes

including NP, CR, MI and C according to UNECE and USDA standards. This method can be adapted for grading and sorting machines to increase speed and accuracy.

3. Materials and Methods

Figure 1 shows the framework for quality grading and classification of almond variety developed in this research. In the first step, images were taken from almonds then by using appropriate algorithms for segmenting images (i.e., separated from the background). In the next step, after extracting useful features related to size, shape, colour, and texture on almond images, a feature vector was formed. To grade and classify almond variety successfully, it was necessary to find prominent features. Accordingly, features were selected using principal component analysis and modified sequential floating forward selection. In the final step, the almonds were graded and classified into four varieties (NP, MI, CR, and C) [22] by Decision Tree. All of the above-mentioned steps are presented in detail in the following subsections.

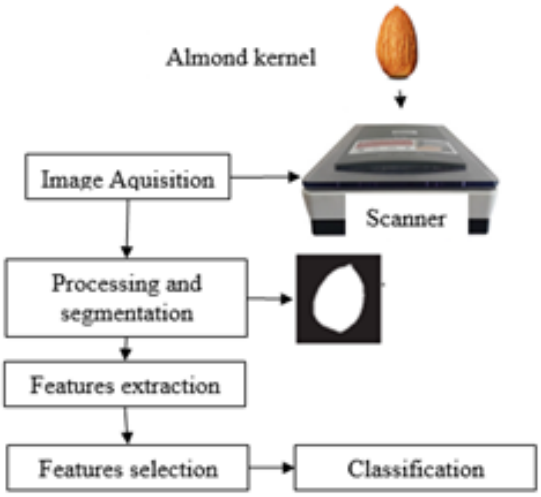

Fig. 1. The framework for quality grading and classification of almond variety.

3.1. Collection of almond kernels

Almonds variety is grown in the season of 2017–18. All samples are collected from the Central Institute of Temperate Horticulture, Srinagar, Jammu and Kashmir, India, and the University of Agricultural Sciences, Dharwad, Karnataka India. Almonds are cleaned (air dust, dirt, and stones) and remove the immature and broken kernels. Then, plastic bags are used to store all the almonds samples and kept at room temperature (25˚−29˚ Centigrade). Figure 2 illustrates the sample images of each kernel variety.

3.2. Image acquisition

Initially, almonds of each variety were manually separated and then a charge-coupled device scanner (ScanJet 3770 with the 24-bit colour of 1200x1200 dots-per-inch) was used for acquiring the images of kernel under laboratory conditions. The images were taken in a way that in each image only almonds of one kernel type were present (as shown in Fig. 1). The images were saved in BMP format with 300 dpi resolution.

Finally, the acquired images were transferred to a personal computer for further analysis. A total of 2000 images were acquired for each variety (500 samples) of almond, including NP, MI, CR, and C (as shown in Fig. 2).

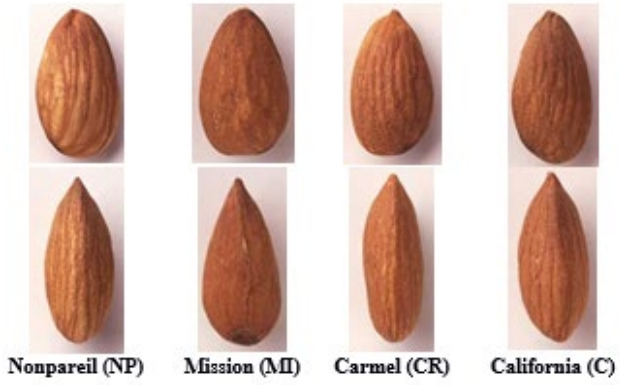

Fig. 2. The varieties almonds Kernel [47].

3.3. Feature extraction

GNU Octave image processing toolbox was used for image analysis (i.e., extract features from each kernel image). In pre-processing, first, the median filter is applied to the kernel image to remove noise. The filter has a 3x3 focus window on the considered pixel and enables the detection of the optimum threshold for image segmentation. The second is image segmentation accomplished using Otsu's algorithm, which separates the foreground and background with the 0.75 as a global threshold value (i.e., pixels with values below 0.75 treated as kernel i.e., binary image and otherwise background) [30-32].

3.3.1. Almonds kernel feature extraction (morphological)

Due to irregular shapes of the agricultural/food products, length (L) and width (W) calculations are more complicated than the area (A) and perimeter (P). The diameter of the Ferret is most commonly used in size by computing the object's major axis and the minor axis [33, 34] shown in Fig. 3. It deals with kernel images, extracting measurable information from different image areas. After obtaining the boundary of the selected area, properties such as the major axis length as L, minor axis length as W, area, and perimeter were determined.

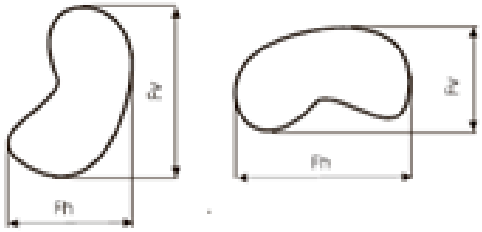

Fig. 3. The object horizontal (*Fh*) and vertical (*Fv*) Feret diameters [42].

3.3.1.1. Size:

The mean and standard deviation, describing simple intensity information, and are given by mean(μ) = $\frac{1}{n}\sum_{x=1}^{n} f_{xy}$ and standard deviation (σ) = $\sqrt{\frac{1}{n} \sum_{x=1}^{n}(f_{xy} - \dot{\mu})^2}$, where f_{xy} represents the pixels in each of the segmented intensity images, and n the total number of evaluated pixels in the segmented image. Figure 4 illustrates the length versus width of almond kernel variety. The X-axis represents the almond variety and Y-axis represents the pixel counts, which is the number of pixels in a segmented image of the kernel. The size distribution of almond's kernel variety is listed in Table 1.

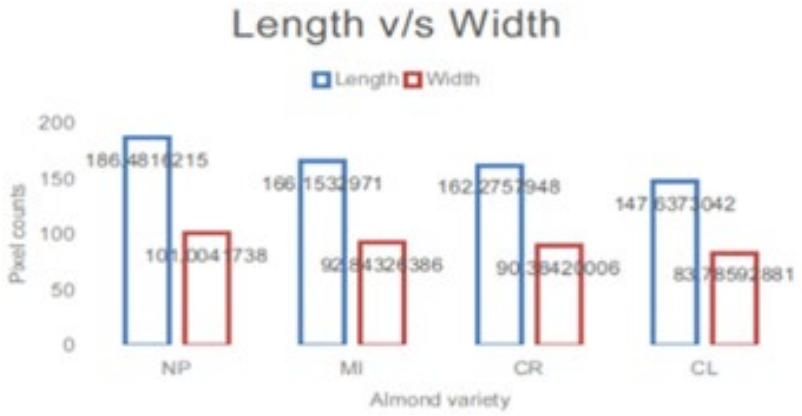

Fig. 4. Demonstrate the length and width of almond kernel variety.

Table 1. The mean and standard deviation of the almond kernels size distribution.

Variety	L	W	A	P
Nonpareil (NP)	3.12 ± 0.16	1.77 ± 0.13	4.08 ± 0.44	6.60 ± 0.39
Mission (MI)	3.27 ± 0.15	1.59 ± 0.09	4.22 ± 0.30	6.72 ± 0.34
Carmel (CR)	2.90 ± 0.14	1.50 ± 0.09	3.05 ± 0.27	5.40 ± 0.32
California (CL)	3.01 ± 0.16	1.38 ± 0.10	3.45 ± 0.31	6.34 ± 0.38

3.3.1.2. Shape:

The shape features were divided into size-dependent measurement (SDM) and size independent measurement (SIM)

i. SDM:

The Shape Factors (SF) describe the shape of a kernel and are calculated from the values of length, width, and area [35]. They are given by:

$$SF1 = L/A \quad (1)$$

$$SF2 = A/L^3 \quad (2)$$

$$SF3 = A/((L/2)2 \times 3.142) \quad (3)$$

$$SF4 = A/((L/2) \times (W/2) \times 3.142) \quad (4)$$

The natural diversity in the morphology of kernel types makes classification a complex work. So, we calculated the shape descriptors such as Elongation (*E*) and Roughness (*R*) [17, 36], and are given by:

$$E=L/W \tag{5}$$

$$R=A/L^2 \tag{6}$$

The distribution of SDM of an almond's kernel is listed in Table 2.

Table 2. Mean and standard deviation of almond kernels SDM.

Variety	Elongation (E)	Roughness (R)
Nonpareil	1.80 ± 0.42	0.35±0.12
Mission	1.83 ± 0.50	0.39 ± 0.09
Carmel	1.74 ± 0.54	0.41 ± 0.10
California	1.70 ± 0.48	0.45 ± 0.12

ii. SIM:

Due to irregularity in shape, the SDM is insufficient to characterize an almond kernel. So, four SIM were calculated to obtain shape information of the almond kernel, regardless of the image size and position. So, Hu moment invariant-based features [17, 36]of the kernel are given by:

$$\varphi_1 = \eta_{20} + \eta_{02} \tag{7}$$

$$\varphi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{8}$$

$$\varphi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{9}$$

$$\varphi_4 = (\eta_{30} - \eta_{12})^2 + (\eta_{21} - \eta_{03})^2 \tag{10}$$

A set of four invariant moments [37, 38] derived from the complex moments of the segmented image and are important for object (i.e., almond kernel) recognition. The bar chart, as shown in Fig. 5, illustrates the distribution of invariant moment features of the almond kernel type.

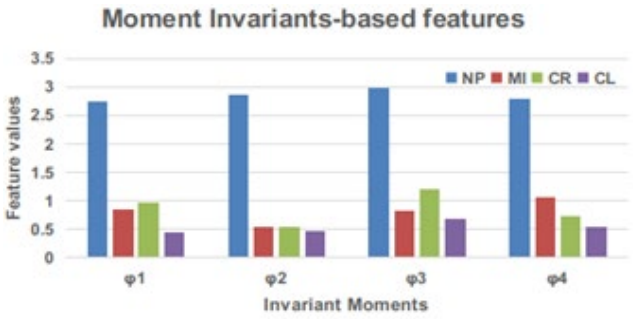

Fig. 5. Illustrate the first four Hu moments of the almond kernel variety.

3.3.2. Colour

In the real world, colour is the most significant and straight-forward feature that humans perceive an image. All colours are seen as various combinations of the three primary colours, namely, Red (R), Green (G), and Blue (B). The primary colours are

added to produce secondary colours such as cyan (green and blue), yellow (red and green), and the like. Widely, different colour spaces are used in CV-based systems today, namely,

RGB, HSL (hue, saturation, and brightness), and CIE-*L*a*b** colour space, etc. The characteristics (i.e., luminance (L), chrominance (C), Hue (H), and Distance metric) are used to distinguish one colour from another. Agriculture and food products such as cashews, peanuts, and chestnut, etc. differ in their colours. Hence, we have tested the suitability of RGB, HSL, and CIE *L*a*b** colour features to recognize, classify and grade the almond kernels [5, 30-32, 39-42].

i. RGB Colour space [36]

Initially, we proposed to separate the R, G, and B components from the kernel image (f_{xy}) and then, calculating L, C, and distance metric using Eq. (11) through Eq. (15).

$$L = 0.2126R_{component} + 0.7152G_{component} + 0.0722B_{component} \quad (11)$$

$$\text{Chrominance of blue component}(Cb) = B_{component} - L \quad (12)$$

$$\text{Chrominance of red component}(Cr) = R_{component} - L \quad (13)$$

$$\text{Hue angle}, H^{\circ} = \begin{cases} \frac{1}{360^{\circ}}[90^{\circ} - tan^{-1}\frac{F}{\sqrt{3}} + 0^{\circ}] \; G_{component} < B_{component} \\ \frac{1}{360^{\circ}}[90^{\circ} - tan^{-1}\frac{F}{\sqrt{3}} + 180^{\circ}] \; G_{component} > B_{component} \end{cases} \quad (14)$$

where $F = (2R_{component}\text{-}G_{component}\text{-}B_{component})/(G_{component}\text{-}B_{component})$

$$\text{Distance metric}, \Delta E_{RGB} = \sqrt{(\Delta E_R)^2 + \quad (\Delta E_G)^2 + \quad (\Delta E_B)^2} \quad (15)$$

where, $\Delta E_R = (\mu_{Red\_component} - R_{component})$,

$\Delta E_G = (\mu_{Green\_component} - G_{component})$ and

$\Delta E_B = (\mu_{Blue\_component} - B_{component})$

ii. Statistical analysis

The colour images are recognized by quantifying the distribution of colour throughout the image, change in the colour about average/mean, and the difference between the highest and the lowest colour values. This quantification is obtained by computing the mean, variance, standard deviation, and range for a given colour image. Since these features represent global characteristics for an image, so we have adopted the mean, variance, standard deviation, and range colour features in this work. Eq. (16) through Eq. (19) are used to evaluate the mean, variance, standard deviation, and range of image samples [36, 41, 43].

$$\text{Mean } (\mu) = \mu_x \sum_{x,y=0}^{n-1} x * (f_{xy}) \; and \; \mu_y \sum_{x,y=0}^{n-1} y * (f_{xy}) \quad (16)$$

$$\text{Variance } (\sigma^2) = \sum_{x,y=0}^{n-1} (f_{xy})(x - \mu)^2 \quad (17)$$

$$\text{Standard Deviation } (\sigma) = \sqrt{\sigma^2} \quad (18)$$

$$\text{Range} = Max(f_{xy}) - Min(f_{xy}) \quad (19)$$

We have extracted 14 colour features ($\mu_{Red\_component}$, $\sigma_{Red\_component}$, $\text{Range}_{Red\_component}$, $\mu_{Green\_component}$, $\sigma_{Green\_component}$, $Range_{Green\_component}$, $\mu_{Blue\_component}$,

$\sigma_{Blue\_component}$, $Range_{Blue\_component}$, L, Cb, Cr, H°, and $\Delta ERGB$). The RGB colour distribution of almonds kernel variety is shown in Fig. 6.

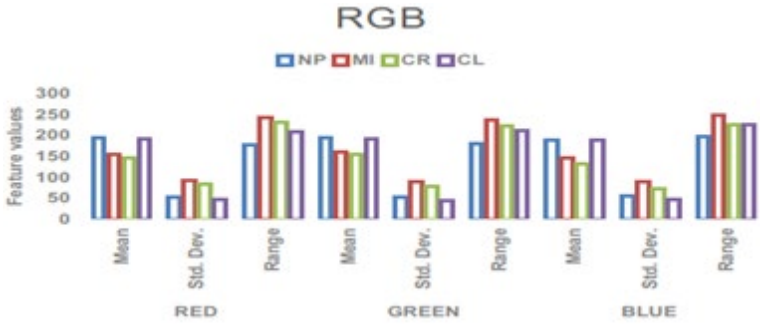

Fig. 6. Illustrates the distribution of the mean, standard deviation, and range of RGB colour image of almond kernel variety.

iii. HSL Colour Space

The values of RGB colour components are in the range [0, 1] and H, S, and L components are extracted from these RGB components. Eq. (20) through Eq. (23) are used to evaluate H, S, and L [5, 44].

$$H_{component}\begin{cases} 60^{\circ} * \left(\frac{G`-B`}{C} mod6\right), C_{max} = R` \\ 60^{\circ} * \left(\frac{B`-R`}{C} mod6\right), C_{max} = G` \\ 60^{\circ} * \left(\frac{R`-G`}{C} mod6\right), C_{max} = B` \end{cases} \quad (20)$$

where C_{max} = Maximum(R`, G`, B`), $C_{min}$ = Minimum(R`, G`, B`), C = $C_{max} - C_{min}$

$R` = R_{component}/255$, $G` = G_{component}/255$ and $B` = B_{component}/255$

$$S_{component} = \begin{cases} 0, & C = 0 \\ \frac{C}{1-|2L-1|}, & C <> 0 \end{cases} \quad (21)$$

$$L_{component} = (C_{max} + C_{min}) / 2 \quad (22)$$

$$\Delta E_{HSL} = \sqrt{(\Delta E_H)^2 + (\Delta E_S)^2 + (\Delta E_L)^2} \quad (23)$$

where, $\Delta E_H = \mu_{Hue\_component} - H_{Component}$,

$\Delta E_S = \mu_{saturation\_component} - S_{Component}$ and

$\Delta E_L = \mu_{Lightness\_component} - L_{Component}$

The statistical analysis is done on H, S, and L colour components by using Eq. (16) through Eq. (19). So, ten colour features ($\mu_{Hue\_component}$, $\sigma_{Hue\_component}$, $Range_{Hue\_component}$, $\mu_{Saturation\_component}$, $\sigma_{Saturation\_component}$, $\text{Range}_{Saturation\_component}$, $\mu_{Luminosity\_component}$, $\sigma_{Luminosity\_component}$, $\text{Range}_{Luminosity\_component}$ and ΔE_{HSL}) are extracted. The almond kernel type HSL colour distribution is shown in Fig. 7.

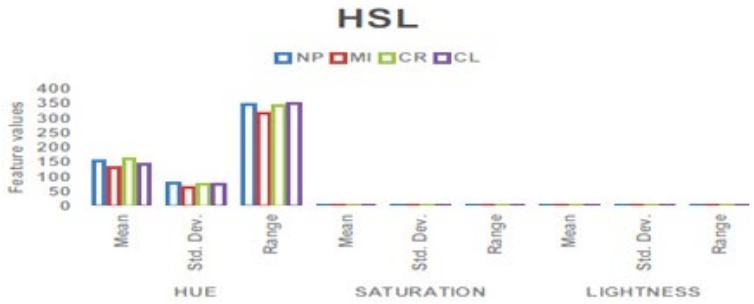

Fig. 7. Illustrate the mean, standard deviation, and HSL colour range of the almond kernel type.

iv. CIE L*a*b*

According to CIE (1976), XYZ and $L^*a^*b^*$ are designed to quantity colour changes continuously with different observed colours. The developed colour model is considered homogeneous because the distance between two colours in linear color space corresponds to the observed differences between them. As a result, it allows the objective colour representation, and its use is critical for applications where the results are consistent with human perception [36, 42, 45, 46], Eqs. (24) through (30) are used to evaluate CIE $L^*a^*b^*$ colour values [5, 44].

$$L^*_{component} = 116f\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - 16 \tag{24}$$

$$a^*_{component} = 500\left[f\left(\frac{X}{X_n}\right)^{1/3} - f\left(\frac{Y}{Y_n}\right)^{1/3}\right] \tag{25}$$

$$b^*_{component} = 200\left[f\left(\frac{Y}{Y_n}\right)^{1/3} - f\left(\frac{Z}{Z_n}\right)^{1/3}\right] \tag{26}$$

The colour intensity and propagation of the almond kernel are determined in the specified area. Nielsen [45] has very explicitly stated that certain colour features, such as chroma, colour score (CS), white index (WI), and colour distance metric contributes to kernel classification. Therefore, we tried extracting these features using the CIE $L^*a^*b^*$ color space.

Chroma (C) =(a^2+b^2)1/2 (27)

Color Score (CS) =bL/a (28)

White Index (WI) = 100-[(100-L^*)^2+a^{*2}+b^{*2}]1/2 (29)

$$\Delta E_{L*a*b*} = \sqrt{(\Delta E_{L^*_{component}})^2 + (\Delta E_{a^*_{component}})^2 + (\Delta E_{b^*_{component}})^2} \tag{30}$$

where, $\Delta E_{L^*_{component}} = \mu_{L^*_{component}} - L^*_{component}$,

$\Delta E_{a^*_{component}} = \mu_{a^*_{component}} - a^*_{component}$ and

$\Delta E_{b^*_{component}} = \mu_{b^*_{component}} - b^*_{component}$

The statistical analysis is done on L^*, a^*, and b^* color component by using Eq. (16) through Eq. (19). A thirteen CIE $L^*a^*b^*$ color features ($\mu L^*_{component}$, $\sigma L^*_{component}$, Range $L^*_{component}$, $\mu a^*_{component}$, $\sigma a^*_{component}$, Range $a^*_{component}$, $\mu b^*_{component}$, $\sigma b^*_{component}$,

Range $b^*_{component}$, C, CS, WI, and $\Delta EL^*a^*b^*$) are extracted. The CIE $L^*a^*b^*$ color distribution of the almond kernel types is shown in Fig. 8.

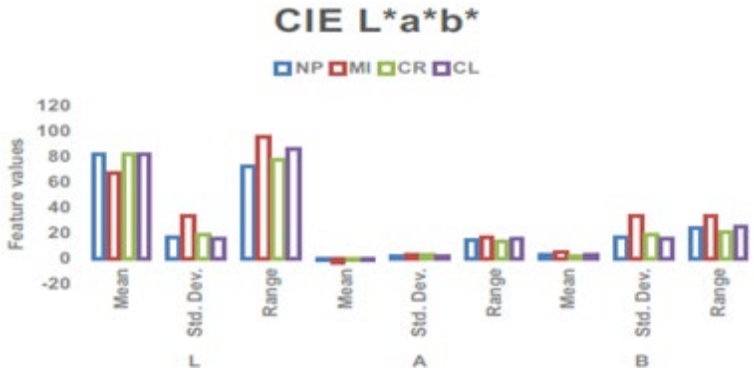

Fig. 8. Describes the mean, standard deviation, and range of CIE $L^*a^*b^*$ color space for the almond kernel type.

3.3.3. Texture

Texture information is extracted from the distribution of the intensity values based on Haralick [47] approach They are computed using grey-level co-occurrence matrices representing the second-order texture information (the joint probability distribution of intensity pairs of neighbouring pixels in the image), where the mean and range-for different pixel distances in four directions of the following variables are measured: mean, variance, standard deviation, contrast, correlation, the angular second moment, energy, dissimilarity, entropy, homogeneity, cluster shade, cluster performance, smoothness, the third movement, and maximum probability [36, 48]. The fifteen texture features are extracted, and the bar graph (as shown in Fig. 9.) shows the distribution of some of the texture features.

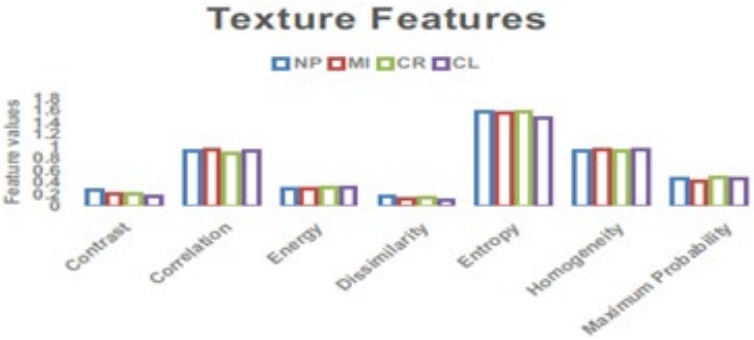

Fig. 9. Illustrates the distribution of some of the textural features using the GLCM of almond kernel variety.

3.4. Feature selection and classification

In this study, each almond kernel variety image is representing 66 features (4-Geometric, 6-SDM, 4-SIM, 14-RGB, 10-HSL 13-CIE $L^*a^*b^*$, and 15-Texture). The n (i.e., 66) extracted features for sample i are arranged in the ith row of matrix F: [Fi1, ….Fin] that corresponds to a point in the n-dimensional measurement feature space.

The features are normalized yielding $N \times n$ matrix W, which elements are defined as $Wij=(Fij-\mu i)/\sigma j$ for $i = 1 \ldots N$ and $j = 1 \ldots n$, where Fij denotes the jth feature of the ith feature vector, N is the number of samples, and μi and σj are the mean and standard deviation of the jth feature, i.e., the jth column of F. Thus, the normalized features have 0 mean and a standard deviation equal to one. Those constant features and highly correlated features can be eliminated because they do not give relevant information about the kernel evaluation quality.

After extraction and normalization, it is necessary to select the best features to train the classifier. It leads to an increase in the classification performance and efficiency in the classification model. In the feature selection, a subset of m features ($m \leq n$) that leads to the smallest classification error is selected. The selected m features are arranged in a new matrix X with N x m elements obtained from m-selected columns of the large set of normalized feature W [49-51]. The features can be selected using several state-of-the-art algorithms reported in the literature, In the proposed work, feature selection algorithms such as principal component analysis (PCA) and Modified Sequential Floating Forward Selection (MSFFS) are implemented using the GNU octave and have achieved high performance.

i. Principle Component Analysis (PCA):

To reduce the feature vector, so applied the dimensionality reduction techniques as PCA. In PCA, first, prepare the correlation matrix from the quantified data set X (dimension-d). After that estimate the eigenvectors and eigenvalues from the correlation matrix. Next, the eigenvalues are arranged in descending order. Choose the k eigenvectors corresponding to the largest eigenvalues of k, where k is the number of new feature subspace dimensions ($k \leq d$), then construct projection matrix (W). Finally, $Y = X \times W$, where Y is the feature subspace/ subset which contains a total of 46 features [52, 53].

ii. Modified Sequential Floating Forward Selection (MSFFS):

The feature ranking and selection are the two stages in this method. In the first stage, for all feature subsets, evaluate the rank of each feature f using Eq. (31).

$$rank(f) = \sum_{S \in CGS(f)} Acc(S)/|CSG(f)| \tag{31}$$

where accuracy rate $Acc(S)$ is calculated from the Nearest Neighbour (NN) method, $CSG(f)$ is the collection of all generated subsets that include f, and $|CSG(f)|$ is the number of these subsets. The larger the $rank(f)$ is, the more important the feature f. In the second stage, K (i.e., 4) classes in the dataset generate K binary sub-classifications, where each binary sub-classification (class j for instance) involves the dataset of class j and class *non-j*. Class non-j represents all other data patterns not belonging to class *j*. Data pattern *w* to each independent sub-classified to calculate the corresponding membership (j). The formulas are listed below Eq. (32) through Eq. (34).

$$membership(j) = n_{non-j}/n_j + n_{non-j} \tag{32}$$

where,

$$n_j = \min_{n_l \in Class\ j} \|w - w_l\| \tag{33}$$

$$n_{non-j=} \min_{n_p \in Class\ non-j} \|w - w_p\| \tag{34}$$

where n_j is the shortest distance between *w* and the data pattern belongs to class *j* in the jth sub-classified, and n_{non-j} is the shortest distance to class *non-j*. The bigger the membership, the higher the probability that *w* belongs to class *j*. Select the class with the largest membership as the classified class to data pattern w using Eq. (35).

$$j* = Arg \max_{j=1,2,3,4} membership(j) \tag{35}$$

where j^* is the classified class to data pattern *w*.

In the experimental simulation in this study, we used MSFFS. The dataset for the experiment includes 4-Geometric, 6-SDM, 4-SIM, 14-RGB, 10-HSL 13-CIE $L^*a^*b^*$, and 15-Texture. The total samples of 2000 kernel varieties (NP, MI, CR, and C) are divided into training (i.e., total 1200 samples of which 300 from each variety) and testing (i.e., total 800 samples of which 200 from each variety). To extract feature subsets using MSFFS, we used the NN method as the classification to build the classifier. Table 3 shows the datasets generated for simulation. And simulation is done on a personal computer (Hewlett Packard), which is having a 7th Generation Intel® Core™ i3 processor, 4 GB DDR4-2400 SDRAM (1 x 4 GB), 1 TB 7200 rpm SATA.

Table 3. Simulation results from the dataset using the MSFFS method.

Dataset (i.e., Subset)	No.'s of features selected	Accuracy of test data (%)	Computational time (minutes)
Geometric(4)	02	90.04	16
SDM(6)	03	89.03	22
SIM(4)	01	86.45	19
RGB(14)	09	91.04	53
HSL(10)	06	88.21	40
CIE *L*a*b(13)**	08	92.45	50
Texture(15)	11	87.98	60
Total	40		

Finally, 40 features are in subspace/subset. In this work, we have tested two feature selections algorithms (PCA & MSFFS) to obtain the highest performance with classification. The idea is to obtain the highest accuracy defined as the proportion of true results.

3.5. Classification

A supervised learning approach was used to train the pattern classification algorithm. Supervised classes, known as labels, were based on four categorical classes according to international standards (UNECE and USDA), where each acquired almond kernel variety (NP, MI, CR, and C). In the proposed work, the following well-known classifier has achieved high performance. Initially, we tried with the Backpropagation Neural Network classifier and obtained less accuracy, when compared with the Decision Tree classifier. So decided to continue with the Decision Tree.

Decision Tree (DT): It is a form of hierarchical classifier and classification that uses a series of simple decision functions, commonly binary to determine the class of unknown patterns. The decision tree model starts from the root node, and the branches

pass through the internal nodes toward the terminal nodes. Terminal nodes called leaf nodes represent different classes. During the decision tree construction, attribute selection steps are used to select the best partitioning attribute into tuples. Popular criteria for attribute selection are Information gain, Gain ratio, and the Gini index [54]. In this work, the decision tree is constructed from the training data labelled using the Gini index, $Gini(D) = 1 \sum_{i=1}^{m} p_i^2$ where pi is the probability that a tuple in D belongs to class Ci and is estimated by |Ci, D| / |D|. Each leaf node corresponds to an almond kernel grade label. Figure 10 illustrates the length as a root node.

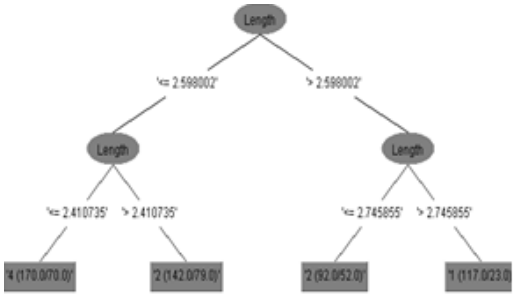

Fig. 10. DT is considered for length as a root node.

4. Results and Discussion

Confusion matrices are used to evaluate the performances of classifiers as the best method for assessing multiclass predictive models. Each column of the matrix contains patterns in the predicted class, but each row contains examples of the actual type (ground truth). The cell of the confusion matrix can be true positive (TP), true negatives (TN), false positive (FP), and false negatives (FN). All the correct predictions on the diagonal of the matrix. Formulas of the performance parameters are provided in Eqs. (36) and (37) [29].

$$\text{Accuracy (for a certain class)} = \frac{TP}{TP+FP+FN} \tag{36}$$

$$\text{Overall Accuracy} = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \tag{37}$$

After normalization, we are considered the subset of extracted features (geometric, shape, color, and texture) of the almond kernel for classification with/without feature section (PCA and MSFFS). Due to unnecessary features, the obtained classification accuracy was low. Therefore, a whole feature set is considered to achieve better accuracy. So, we used three strategies to develop a proposed system and obtained an accuracy of the DT classifier is listed in Table 4 for strategy-1, Table 5 for strategy-2, and Table 6 for strategy-3.

Table 4. Confusion matrix for DT classifier using splitting rule Gini diversity index (without PCA and MSFFS feature selection).

Variety	NP	MI	CR	CL	Total	The success rate in %
NP	**179**	12	9	0	200	94.75
MI	56	**111**	33	0	200	77.75
CR	9	61	**187**	43	200	71.75
CL	0	26	58	**116**	200	79.00
Total					800	**80.81**

Table 5. Confusion matrix using DT classifier with splitting rule Gini diversity index of strategy-2 (with PCA).

Variety	NP	MI	CR	CL	Total	The success rate in %
NP	**170**	30	0	0	200	92.50
MI	65	**135**	0	0	200	83.75
CR	0	16	**156**	28	200	89.00
CL	0	0	8	**192**	200	98.00
				Total	800	**90.81**

The developed method results are compared with the reported literature [55, 56]. Artificial Neural Network (ANN) to identify five classes are normal almond (NA), broken almond (BA), double almond (DA), wrinkled almond (WA) and shell of almond (SA) with 18-7-7-5 topology. The extracted features were shape (8), HIS color (45), and texture (162) from the almond kernel. The obtained accuracy of the ANN classifier for each class was NA (98.92%), BA (99.46%), DA (98.38%), and SA (100%) with PCA. But, in this paper, we considered the almond variety (NA, MI, CR, C) grade and classification using 66 features with feature selection (PCA and MSFFS) methods. The obtained accuracy of DT classifier is 80.81% (strategy-1) listed in Table 4, 90.81% (strategy-2) listed in Table 5, and 97.13% (strategy-3) listed in Table 6. The proposed method was developed on a personal computer (Hewlett Packard), which is having a 7th Generation Intel® Core™ i3 processor, 4 GB DDR4-2400 SDRAM (1 x 4 GB), 1 TB 7200 rpm SATA. The processing of an almond kernel takes less than one minute for the proposed method. In real practice, the manual method is being followed to identify the grades of the almond kernel. This proposed method will be helpful for the industry to maintain quality control.

Table 6. Confusion matrix using DT classifier with splitting rule Gini diversity index of strategy-3 (with MSFFS).

Variety	NP	MI	CR	CL	Total	The success rate in %
NP	**196**	4	0	0	200	99.00
MI	9	**176**	15	0	200	94.00
CR	0	3	**190**	7	200	97.50
CL	0	0	8	**192**	200	98.00
				Total	800	**97.13**

5. Conclusion

The classification of almond products has an essential role in promoting the export of this valuable product. In this research, almonds kernel variety used by decision tree (DT) and image processing techniques based on the size, shape, color, and texture features, and the application of feature selection methods (PCA and MSFFS) is to reduce the dimensionality of the features, for better accuracy, and then almonds classified into four distinct categories based on UNECE (2009) standards. Feature selection methods are used to select interactive and efficient features from all extracted features. Because of the feature vector high dimensionality, PCA and MSFFS were used to reduce the ratio of the feature vector. Confusion matrix and statistical parameters show that using the combination of size, shape, color, and texture features from RGB, HSL, and CIE $L^*a^*b^*$ color space and applying DT to classify almond products is useful and successful. The classification accuracies of the three strategies are compared, and it was found that DT with MSFFS has an overall 97.13% efficiency across the considered almond kernel variety (NP, MI,

CR, and CL). This approach extends to a real-time range and sorting machines. Work is currently underway in this direction.

Acknowledgment

The authors want to thank the University of Agricultural Sciences, Dharwad, Karnataka, India, and Central Institute of Temperate Horticulture, Srinagar, Jammu & Kashmir, India for providing relevant information on the identification and variety of almonds. The authors thank the Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, INDIA-576104 for providing the necessary resources.

Nomenclatures

$a^*_{component}$	a* component of CIEL*a*b* color space
$b^*_{component}$	b* component of CIEL*a*b* color space
$B_{component}$	The blue component of RGB color space
C	Chrominance or Chroma
Cb	Chrominance of blue component
Cr	Chrominance of red component
$G_{component}$	The green component of RGB color space
$H_{component}$	Hue component of HSL color space
Ho	Hue angle
L	Luma
$L_{component}$	Lightness component of HSL color space
$L^*_{component}$	L* component of CIEL*a*b* color space
$R_{component}$	The red component of RGB color space
$S_{component}$	Saturation component of HSL color space

Greek Symbols

$\Delta E_{a^*_{component}}$	Distance metric of a* color
$\Delta E_{b^*_{component}}$	Distance metric of b* color
ΔE_B	Distance metric of blue color
ΔE_G	Distance metric of green color
ΔE_H	Distance metric of hue color
ΔE_{HSL}	Distance metric of HSL color space
ΔE_L	Distance metric of lightness color
$\Delta E_{L^*_{component}}$	Distance metric of L* color
ΔE_{L*a*b*}	Distance metric of CIEL*a*b* color space
ΔE_R	Distance metric of red color
ΔE_{RGB}	Distance metric of RGB color space
ΔE_S	Distance metric of saturation color
σ^2	Variance
σ	Standard Deviation
μ	Mean

References

1. Food and Agriculture Organizations(FAO). (2018). Statistical Databases. Retrieved August 12, 2019, from http://www.faostat.fao.org.

2. Cakmak, Y.S.; and Boyaci I.H. (2010). Quality evaluation of chickpeas using an artificial neural network integrated computer vision system. *International Journal of Food Science and Technology*, 46(1), 194-200.
3. Du, C.J.; and Sun, D.W. (2004). Recent developments in the applications of image processing techniques for food quality evaluation. *Trends Food Science Technology,* 15(5), 230-249.
4. Patel, K.; Kar, A.; Jha, S.N.; and Khan, M.A. (2012). Machine vision system: A tool for quality inspection of food and agricultural products. *Journal of Food Science and Technology*, 49(2), 123-141.
5. Lawless, H.T.; and Heymann, H. (2010). *Color and appearance. Sensory evaluation of food. Food science text series.* New York: Springer.
6. Teimouri, N.; Omid, M.; Mollazadeand, K.; and Rajabipour, A. (2016). An artificial neural network-based method to identify five classes of almond according to visual features. *Journal of Food Process Engineering*, 39(6), 625-635.
7. Poonnoy, P.; Yodkeaw, P.; Sriwai, A.; Umongkol, P.; and Intamoon, S. (2014). Classification of boiled shrimp's shape using image analysis and artificial neural network model. *Journal of Food Process Engineering*, 37(3), 257-263.
8. Tautho, C.C.; Satonero, L.A.; Tautho, Y.C.; and Lariosa, E.A. (2002). Production, postharvest handling, and marketing practices of mango growers in Bukidnon [Philippines]. *Philippine Journal of Crop Science*, 27(1), 37.
9. UNECE. (2009). UNECE Standard DDP-21 concerning the marketing and commercial quality control of blanched almond kernels. Retrieved October 11, 2014, from http://www.unece.org/trade/agr/standard/dry/ddp-standards.html.
10. Chen, X.; Xun, Y.; Li, W.; and Zhang, J. (2010). Combining discriminant analysis and neural networks for corn variety identification. *Computers and Electronics in Agriculture*, 71, 48-53.
11. Choudhary, R.; Paliwal, J.; and Jayas, D.S. (2008). Classification of cereal grains using wavelet, morphological, colour, and textural features of non-touching kernel images. *Biosystem Engineering*, 99(3), 330-337.
12. Kiratiratanapruk, K.; and Sinthupinyo, W. (2011). Color and texture for corn seed classification by machine vision. *Proceedings of IEEE International Symposium on Intelligent Signal Processing and Communication Systems* (*ISPACS*). Chiang Mai, Thailand, 7-9.
13. Zapotoczny, P. (2014). Discrimination of wheat grain varieties using image analysis and multidimensional analysis texture of grain mass. *International Journal of Food Properties*, 17, 139-151.
14. Donis-Gonzalez, I.R.; Guyer, D.E.; Leiva-Valenzuela, G.A.; and Burns, J. (2013). Assessment of chestnut (Castaneaspp.) slice quality using color images. *Journal of Food Engineering*, 115(3), 407-414.
15. Mollazade, K.; Omid, M.; and Arefi, A. (2012). Comparing data mining classifiers for grading raisins based on visual features. *Computer and Electronic in Agriculture*, 84, 124-131.
16. Yu, X.; Liu, K.; Wu, D.; and He, Y.(2012). Raisin quality classification using least squares support vector machine(LSSVM) based on combined color and texture features. *Food Bioprocess Technol*ogy, 5(5), 1552-1563.

17. Namias, R.; Gallo, C.; Craviotto, R.M.; Arango, M.R.; and Granitto, P.M. (2012). Automatic grading of green intensity in soybean seeds. *Proceedings of the* 13*th Argentine Symposium on Artificial Intelligence* (*ASAI*), 96-104.
18. Soedibyo, D.W.; Ahmad, U.; Seminar, K.B.; and Subrata I, D.M. (2010). The development of automatic coffee sorting system based on image processing and artificial neural network. *Proceedings of the International Conference on the Quality Information for Competitive Agricultural-Based Production System and Commerce*, 272-275.
19. Kumar, M.; Bora, G.; and Lin, D. (2013). Image processing technique to estimate geometric parameters and volume of selected dry beans. *Journal of Food Measurement and Characterization*, 7(2), 81-89.
20. Hanbury, A. (2002). The taming of the hue, saturation, and brightness color space. *In CVWW'*02-*Computer Vision Winter Workshop*, 234-243.
21. Zhong-Zhi, H.; and You-gang, Z. (2009). A method of detecting peanut cultivars and quality based on the appearance characteristic recognition. *Journal of the Chinese Cereals and Oils Association*, 5, 123-126.
22. Chen, H.; Wang, J.; Yuan, Q.; and Wan, P. (2011). Quality classification of peanuts based on image processing. *Journal of Food, Agriculture and Environment*, 9(3 and 4), 205-209.
23. Jackman, P.; Sun, D.W.; Du, C.J.; Allen, P.; and Downey, G. (2008). Prediction of beef eating quality from color, marbling, and wavelet texture features. *Meat Science,* 80(4), 1273-1281.
24. Okamura, N.K.; Delwiche, M.J.; and Thompson, J.F. (1993). Raisin grading by machine vision. *Transactions of the ASAE*, 36(2), 485-492.
25. Pearson, T.C.; and Toyofuku, N. (2000). Automated sorting of pistachio nuts with closed shells. *Applied Engineering Agriculture*, 16, 91-94.
26. Paliwal, J.; Visen, N.S.; Jayas, D.S.; and White, N.D.G. (2003). Cereal grain and dockage identification using machine vision. *Biosystem Engineering*, 85(1), 51-57.
27. Wang, S.; Liu, K.; Yu, X.; Wu, D.; and He, Y. (2012). Application of hybrid image features for fast and non-invasive classification of raisin. *Journal of Food Engineering,* 109(3), 531-537.
28. Pazoki, A.R.; Farokhi, F.; and 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.
29. Zou, Q.; Fang, H.; Liu, F.; Kong, W.; and He, Y. (2010). Comparative study of distance discriminant analysis and backpropagation neural network for identification of rapeseed cultivars using visible/near-infrared spectra. *Proceedings of the Computer and Computing Technologies in Agriculture IV* - 4*th IFIP TC* 12 *Conference*, *CCTA* 2010. Nanchang, China, 124-133.
30. Narendra, V.G.; and Hareesh K.S. (2016). Recognition and classification of white wholes (WW) grade cashew kernel using artificial neural networks. *Acta Scientiarum Agronomy*, 38(2), 145-155.
31. Narendra, V.G.; and Priya, K. (2017). Intelligent classification models for food products basis on morphological, color, and texture features. *Acta Agronomica*, 66(4), 486-494.

32. Narendra, V.G.; Dasharathraj, K.S.; and Hareesh, K.S. (2012). Computer vision system for cashew kernel area estimation. *Proceedings of the International Conference on Computing Communication and Networking Technologies* (*ICCCNT*)*, IEEE Conference*. 1-6.
33. Mendoza, F.A.; Dejmek, P.; and Aguilera, J.M. (2006). Calibrated color measurements of agricultural foods using image analysis. *Postharvest Biology and Technology*, 41(3), 285-295.
34. Zheng, C.; Sun, D.W.; and Zheng, L. (2006). Recent developments and applications of image features for food quality evaluation and inspection-A review. *Trends in Food Science and Technology*, 17(12), 642-655.
35. Symons, S.J.; and Fulcher, R.G. (1988). Determination of wheat kernel morphological variation by digital image analysis, I Variation in eastern Canadian milling quality wheat. *Journal of Cereal Science*, 8(3), 211-218.
36. Sun, D. (2008). *Computer vision technology for food quality evaluation*. Elsevier Inc.
37. Hu, M.K. (1962). Visual pattern recognition by moment invariants. *IRE Transaction on Information Theory*, 8(2), 179-187.
38. Mercimek, M.; Gulez, K.; and Mumcu, T.V. (2005). Real object recognition using moment invariants. *Sadhana*, 30(6), 765-775.
39. Du, C.; and Sun, D. (2005). Comparison of three methods for the classification of pizza topping using different color space transformations. *Journal of Food Engineering*, 68(3), 277-287.
40. Plantaniotis, K.N.; and Venetsanopoulas, A.N. (2000). Color image processing and applications. *Springer-Verlag*, 237-277.
41. Sangwine, S.J. (2000). Color in image processing. *Electronics and Communication Engineering Journal*, 12(5), 211-219.
42. M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta (1996). A standard default color space for the internet - sRGB. http://www.color.org/srgb.html.
43. Savakar, D.G.; and Anami, B.S. (2009). Recognition and classification of food grains, fruits, and flowers using machine vision. *International Journal of Food Engineering*, 5(4).
44. Wen, C.Y.; and Chou, C.M. (2004). Color image models and its applications to document examinations. *Forensic Science Journal*, 3(10), 23-32.
45. Nielsen, S.S. (2003). *Food analysis*(3rd ed.). Kluwer Academic Publishers.
46. Maenpaa, T.; and Pietikainen, M. (2004). Classification with color and texture: Jointly or separately. *Pattern Recognition*, 37(8),1629-1640.
47. Harlick, R.M. (1979). Statistical and structural approaches to texture. *Proceedings of IEEE*, 67(5), 786-204.
48. Basavaraj, S.; Anami; and Viswanath, C.B. (2009). Texture based identification and classification of bulk sugary food objects. *ICGST-GVIP Journal*, 9(4) 3-16.
49. Li, G.Z.; Yang, J.; Liu, G.P.; and Xue, L. (2004). Feature selection for multi-class problems using support vector machines. *Lecture Notes in Computer Science,* 3157, 292-300.

50. Thiemjarus, S.; Lo, B.; and Yang, G. (2004). Feature selection for wireless sensor networks. *Proceedings of the* 1st *International Workshop on Wearable and Implantable Body Sensor Networks.* London, UK.
51. Mery, D.; Pedreschi, F.; and Soto, A. (2012). Automated design of a computer vision system for visual food quality evaluation. *Food Bioprocess Technology,* 6, 2093-2108.
52. Haykin, S. (1999). *Neural networks: A comprehensive foundation*. New York: Prentice-Hall.
53. Papoulis, A.; and Pillai, S.U. (2002). *Probability, random variables, and stochastic processes*. New York McGraw-Hill
54. Ian, H.; Witten; Frank, E.; and Mark, A. (2011). *Data mining practical machine learning tools and techniques*(3rd ed.). Elsevier Inc.
55. Shetty, D.K.; Acharya, D.U.; Prajual, P.J.; Malarout, N.; and Narendra, V.G. (2019). Calculation of area and perimeter of guntur and byadagi chilli images-A Fourier transformation. *International Journal of Recent Technology and Engineering*, 8(3), 4816-4819.
56. Shetty, D.K.; Acharya U.D.; Narendra V.G.; and Prajual, P.J. (2020). Intelligent system to evaluate the quality of DRC using image processing and then categorize using artificial neural network (ANN). *Indian Journal of Agricultural Research*, 54(6), 716-723.