Lecturer Tanmay SARKAR, PhD
E-mail: tanmay@wbscte.ac.in; tanmays468@gmail.com
Malda Polytechnic, West Bengal State Council of Technical Education
Food Technology and Biochemical Engineering, Jadavpur University
Assistant Professor Alok MUKHERJEE, PhD Student
E-mail: alok@gcect.ac.in, alokmukherjee.ju@gmail.com
Government College of Engineering and Ceramic Technology
Assistant Professor Kingshuk CHATTERJEE, PhD
E-mail: kingshukchatterjee@gcect.ac.in; kingshukchaterjee @gmail.com
Government College of Engineering and Ceramic Technology
Professor Tanupriya CHOUDHURY, PhD
E-mail: tanupriya@ddn.upes.ac.in
Computer Science, University of Petroleum and Energy Studies

DETECTION OF EDIBILITY OF AMLA (*Emblica officinalis*) THROUGH PCA BASED IMAGE ANALYSIS

Abstract.*Identification of edibility of fruit samples is very essential, as well as difficult. This is more applicable in places where bulk fruits are used in different automated factories, where investigation of each fruit manually is an impossible task. In this work, we have proposed a principal component analysis (PCA) based threshold classifier scheme for the identification of edibility of amla fruits. We have analyzed only the hue histogram of the image samples using PCA to segregate the samples into Good, Intermediate or Bad classes. Use of analysis like PCA reduces the computational burden many folds compared to the other supervised learning models involving variants of neural network, or mathematically heavier transform based models like wavelet or Fourier transforms. The model is validated using different sample images. High accuracy of 98.33% for classification of samples is achieved in this work. Most importantly, low computational burden and non requirement of any other pre-filtering method to the sample images highlights the effectiveness of this algorithm.*

Keywords*: Multivariate analysis; Principal Component Analysis (PCA), Hue histogram, Smartphone image; Consumer acceptability; Edibility classifier; Image analysis.*

JEL Classification: L66; C63; C33; O13

DOI: 10.24818/18423264/56.2.22.06

1. Introduction

Identifying fresh fruits in market places for consumption is a difficult challenge for most people. Hence, the analysis of these images of different fruits, vegetables and other edible agricultural products bear a major significance; and here lies the major importance of different image-processing methods for the analysis and quality assessment of the agricultural products. In this work, we focus on developing an algorithm to identify fresh amla (*Emblica officinalis*) from its image. This task is difficult, especially due to the different varieties of the same fruit and vegetable which exhibit difference in colour, shape and texture with variation of their ripeness; as well as, due to the dissimilarity of image acquisition methods (Dubey and Jalal, 2015).

The preliminary challenge with fruit image analysis is the vast variation in the texture and colour attributes. Seasonal variation, level of maturity, climate condition, soil composition and cultivar are the key parameters determine the colour variation of fruits and vegetables, even within the same batch. This variability ultimately causes a restriction in automated detection system. Thus expert dependency and manual verification of each fruit is obvious; both in industries as well as in the domestic household. Several unsupervised methods are developed by researchers for defect detection; but the development of unsupervised method for quality evaluation of native subtropical fruit is still less explored.

The mobile diagnostics in the field of agriculture, health care and environmental monitoring is gaining attention, due to its exclusive features like on-site availability, portability and rapid extraction of information. In order to avoid expensive instrument-based analysis, static nature of the laboratory equipment and time intensive routine test, instant and unsupervised system is required to maintain the food safety standard. The significantly large number of smartphone users may contribute to food standard analysis (Novotna, 2020). Smartphone enables the wide accessibility; fast communication, internet connectivity, real-time notification facility and geo-location features. It is portable, easy to use and feasible to deploy in diverse conditions. Smartphone may facilitate acquisition of information in a relatively easier way, compared to complex, lengthy and laborious food testing protocols executed by experts. Hence, Smartphone based applications; especially colour-based image analysis for food standard judgment is gaining popularity. Most importantly, image capturing features, non requirement of additional hardware for data collection and information management enable personal Smartphone users to develop an idea about the quality of food sample. This may initiate a paradigm shift in the sector of food safety and shelf-life detection.

In this work, we have developed a Hue based analysis (HSV) for the quality assessment of amla. Amla has extreme importance in the traditional medicinal system from ancient days. Ayurvedic medicine (India) as well as Tibetan, Chinese and herbal medicine have utilized amla as an ingredient (Baliga and Dsouza, 2011). It is evident from different studies that amla contains lots of essential nutrients which are effective in fighting many diseases. It acts as an antioxidant, anti-microbial, anti-inflammatory agent and exhibits hypoglycemic, hypolipidemic activities (Kapoor et al., 2020; Diaconeasa et al., 2019). Due to the perishable nature of the fruit, it cannot be consumed after a few days without any preservation. Proper storage technique and preservation method can extend its shelf life (Goraya and Bajwa, 2015). There are few indications by which one can judge the freshness of the fruit like browning, bruises and change in shape of the fruit though these are not enough to understand the quality of the fruit by eye estimation. Continuous softening of the fruit cell wall primarily regulates the shelf-life of fruits. Deterioration continues during transportation and handling; outer surface is softened and become vulnerable to microbial and pest attack (Bahramsoltani and Rahimi, 2020). Textural attributes and firmness are the most important factor responsible for consumer acceptance. This degradation in surface texture is reflected from the hue spectrum of the fruits.

Several researches have been carried out for the assessment of the quality of fruits and vegetables. Apart from detection of rotten samples, this analysis helps in detection of different species of these and their varieties (Dubey and Jalal, 2015).This is more helpful in automatic price scheduling in a supermarket. Hence, the importance of digital image processing has become increasingly popular. Most of the researches use either colour or texture analysis; either in binary mode or as a multiclass detection method. Some other works also analyzes the shape, firmness, absence of bruises as well for classification and quality assessment fruits.

Many of these analyze the red-green-blue (RGB) levels of the image or the hue-saturation-intensity (HSV) level. Here, we have adopted a binary method of faulty fruit detection using hue histogram analysis which has high application in research (Danti et al. 2012), rather than another popular choice of colour histograms used in red-green-blue (RGB) space (Dubey and Jalal, 2012; Antonelli et al., 2004). Hue and saturation parameters bear high importance since these are less prone to variation due to the surrounding illumination level and variation in sensor of the capturing device, compared to red-green-blue (RGB) parameters. More importantly, we have analyzed the hue histograms using PCA, instead of direct hue spectrum analysis. We have also shown that application of PCA improves the detection of fruit quality. PCA is one of the major statistical methods used for analyzing multivariate data and identifying the major directions of variation, in the descending order of importance. The hue intensity histograms are analyzed using the proposed PCA algorithm to obtain the PC score index for each image individually. These PC features are further analyzed using threshold based classifier scheme to segment the bad or non-edible fruit images from the good and intermediate quality fruit images. The good and intermediate quality fruit images together are denoted as the edible set. The qualities of fruits were determined by food experts based on Hedonic scale on three parameters: colour, shape and texture.

In the proposed work we have explored the hue levels of different sets of fruit samples; with the help of principal component analysis (PCA). PCA uses an advance analysis of the statistical covariance and the associated eigenvectors and eigenvalues, to yield a number of principal directions in the descending order of importance. PCA is extremely effective in identifying the key features from a signal or image; simultaneously, reducing the dimension of a multivariate data set; hence, used extensively in agricultural researches, especially for defect detection of different fruits and vegetables like orange and mandarin (López-García et al., 2010), cucumbers (Liu et al., 2006), mushrooms (Gowen et al., 2008), peaches (Sun et al., 2018), apple (Zhu et al., 2007) and several more. Other methods of have also been practiced where different quality parameters like moisture, instrumental colour, antioxidant profile etc. have been investigated using PCA (Patras et al., 2011). In several of these works, PCA plays a vital role in reduction of the image dimension; although, retaining the most important directions of variation. The authors of (Geladi and Grahn, 2006; Wu et al., 1997) have described the effective ways of multivariate analysis; as well as, shown the use of kernel PCA, especially for multivariate image analysis (MIA). Baranowski et al., 2012; Zhao et al. 2010; Ariana et al., 2006 and Xing et al., 2005 have proposed efficient bruise detection methods using a multi-tool model including PCA. Thus, altogether, PCA has a major role in this field of research; and we have explored this method further in this work.

In this paper, we have described the fruit sample and methods of image acquisition in the initial sections; followed by analysis of the methodology of designing the classifier with case studies and histogram and PCA features. Finally we have described the validation of the model and concluded the paper explaining the key features and outcomes of the proposed algorithm.

2 Experimental arrangements

2.1 Material and method

Amla were bought from Agri Horticultural Society of India, Kolkata, India (22°31'41.52" N and 88°19'59.88" E) harvested on the same day of purchase. Out of 240 fresh fruits 12 samples were choose randomly. All the 12 fruits were mature, sound (87.3% moisture, 9.6% carbohydrate, 1.2% protein, and 0.6% fat) free from any bruise or scar with 15 to 25 g of weight and 2.7 to 3.6 cm of diameter (Sarkar et al., 2021).

2.2 Image acquisition

The samples were captured at different angles with Redmi Note 9 Pro smartphone device (48 megapixel camera with Samsung Isocell GM2 sensor) and the operating system was Android v10 (Q). The distance between the mobile camera and the fruit was 25 cm. The size of each image was 8000 × 6000 pixels. The physical aperture was F1.79 (Mukherjee et al., 2021).

2.3 Colour, shape and texture evaluation

The semi-trained panel (35 male participants and 28 female participants) with 21-48 years of age bracket was selected in coherence with the guidelines of

ISO 8586-1 (1993) to evaluate the colour, shape and textural changes of the sample. ISO 4120:2004, ISO 5496:2005 and ISO 10399:2004 were considered for selection of panel lists in the basis of triangular discrimination, proficiency of colour identification and competency of the members in differentiating two fruit depending on their colour, shape and texture. The colour, texture and shape of fruit samples were apprised with nine point hedonic scale (where, 9 = excellent, 8 =very good, 7 = good, 6 = satisfactory, 5 = neither good nor bad, 4 = bad, 3 = moderately poor, 2 = poor, 1 = worst). Samples were evaluated at 25±5 °C and ambient conditions. The hedonic scores were recorded and reported as the average numeric value.

2.4 Data set description

Amla samples are kept at the same position for the next few days. We have tested the fruits with Hedonic scale on each day and thrice daily as mentioned earlier. It is observed from the Hedonic scale values that the quality of fruits falls below 5 point from seventh day onwards. In this work, we have found that fruit samples on day 1 and day 2 are at the most fresh condition. Hence, we have segmented the class of fruits in such a way that when the Hedonic scale level is above 5 we are calling it as *'Edible'* class and when the level falls below 5 we are calling it *'Non-Edible'* class. Hence the fruit images taken from 7th day onwards are treated as *'Non-Edible'* class, and prior to that we are denoting those by *'Edible'* class.

We have further segmented the images of the *'Edible'* class into two more categories; i.e., the images taken on day 1 and day 2 are treated as the freshest condition which is treated as *'Good'* condition and the images taken during day 3 to day 6 are treated as *'Intermediate'* conditions. We have observed five fruits on each day thrice daily at three different times: morning, afternoon and night. Hence, the total data set contains (5 samples × 3 times × 8 days), i.e., 120 images, out of which 60 images are considered for training and the rest of the 60 images are used for testing the algorithm. As per this analysis, the good data set contains (5 samples × 3 times × 2 days), i.e., 30 number of image samples; the intermediate set contains (5 samples × 3 times × 4 days), i.e., 60 number of images and the *'Non-Edible'* class of the *'Bad'* class contains (5 samples × 3 times × 2 days), i.e., 30 number of samples. Out of these three classes: *'Good'*, *'Intermediate'*, and *'Bad'*, which thereby contains 30, 60 and 30 number of fruit sample images respectively, we have randomly picked up half of the images from each of these classes and used for training. We have further used the remaining set of data of exactly the half number of images for validating the proposed scheme. Hence, the training and test classes contain 60 images each, containing all three classes. Further we have made an attempt to identify the robustness of the scheme by cross validating the algorithm.

2.5 Morphological changes in fruits with progression of time

Rotting of fruit is a natural process. With progress in time, the cell walls begin to break down and the fruits become less solid, which is reflected in its colour, shape, texture etc. These modifications are analyzed by researchers using several soft computational schemes to detect the faulty samples.

3. Methodology of analysis

3.1 Cropping the image

The raw image taken by the observer is cropped first using the maximum proportion of the fruit. Since, the fruit taken as the sample in this work is having a round shape; we have tried to crop the image using an approximate square crop region inside the circular shape of the fruit. The square region is taken in such a way that each vertex of the square resides approximately on the peripheral border of the fruit. This ensures almost the maximum area of crop, i.e., the maximum area of fruit under the cropped area, which is used for further analysis. These fruits tend to lose their freshness and develop tiny regions of rottenness all over its surfaces with days of progression. These areas are interpreted as discolouration and differ considerably from the images of the fresh fruits: in terms of colour, texture, as well as, shape. Figure 1(a) shows the full images of the three progressive conditions of a single sample of fruit and Figure 1(b) shows the cropped form of the respective images of Figure 1(a).

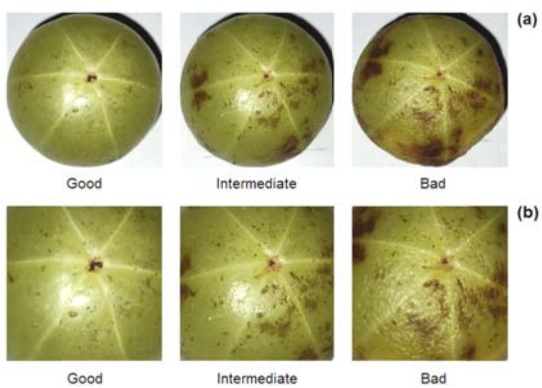

Figure 1 – (a) A sample fruit specimen under three different stages: Good or the freshest condition; Medium condition with minor degradation in quality; and Bad or poor condition, when several segments of the same has already started to degrade and (b) Centre-cropped image of (a)

3.2 Analysis of Hue levels

In the proposed work, we have explored the hue levels of the image. The images are converted from the Red-Green-Blue colourmap (RGB) to hue-saturation-value (HSV) colourmap. Hue level is less affected by the light intensity of the image; whereas, RGB analysis is affected by the variation of light intensity level. Hence, we have judiciously discarded analysis using RGB map. We have obtained the histogram of hue level in this work.

3.3 Application of Principal Component Analysis (PCA)

We have analyzed the hue intensity map so obtained using principal component analysis (PCA). PCA is a broad extension of identifying the interrelation or the correlation between a set of variables in a multivariate statistical model. PCA primarily emphasizes on the variance of a data set rather than the covariances and correlations. For a vector x containing p number of random variables, PCA tries to identify a linear function $a^T_1 x$ with highest variation. This may be expressed as:

$$a_1^T x = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p = \sum_{j=1}^{p} (a_{1j}x_j) \qquad (1)$$

Here, a_1 is a vector of the constant coefficients of the linear equation and a^T_1 denotes the transpose of a_1. In this work, we have analyzed the histogram frequency of each image using PCA; hence, the value of p is 256. Further, another linear function $a^T_2\, x$ is investigated, which is uncorrelated with $a^T_1\, x$ with maximum variance. This is continued till k-th stage to obtain a set of uncorrelated linear function $a^T_1\, x$, $a^T_2\, x$, ... ,$a^T_k x$. Thus, the k-th derived variable is termed as the k-th principal component (PC). Out of the maximum p-th PC, maximum variation is obtained using the ordered sequence of only m PCs, and definitely m is practically much smaller than p. This ensures high reduction in dimension of the data. In most of the cases, the unknown $\sum$ is replaced by the corresponding covariance matrix S, given by:

$$S = \frac{1}{n-1} X^T X \qquad (2)$$

for a set of n observations given by x_1, x_2, ... , x_n, and X is an $(n \times p)$ matrix with (i, j)-th element is given by $(\tilde{x}_{ij} - \bar{x}_j)$ and,

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{ij}, \quad j = 1, 2, \ldots, p \qquad (3)$$

Thus, the (j, k)-th element of S is given by,

$$\frac{1}{\text{n}-1}\sum_{i=1}^{n} (\tilde{x}_{ij} - \bar{x}_j)(\tilde{x}_{ik} - \bar{x}_k) \qquad (4)$$

$\tilde{z}_{i1}$ is defined from the above observations as $\tilde{z}_{i1} = a_1^T x_i$, and $i = 1, 2, \ldots, n$. The vector of coefficients a_1^T is chosen in order to maximize the variance given by:

$$\frac{1}{\text{n}-1} \sum_{i=1}^{n} (\tilde{z}_{i1} - \bar{z}_1)^2 \qquad (5)$$

Considering normalization constraints $a_1^T a_1 = 1$. Similarly, $\tilde{z}_{i2}$, $\tilde{z}_{i3}$ ... are obtained; where, $a_1^T x$ is the k-th PC as defined earlier. The k-th largest eigenvalue of S, which is the sample covariance matrix for the above series x_1, x_2, ... ,x_n, and $a_k$is eigenvector corresponding to $k = 1, 2, \ldots, p$; as defined by the expression (1).

Now, if the two *(n × p)* matrices $\tilde{X}_{ik}$ and $\tilde{Z}_{ik}$ are defined such a way that *(i, k)*-th element is equal to the k-th element $\tilde{x}_{ik}$ of x_i and $\tilde{z}_{ik}$respectively. Thus, $\tilde{X}_{ik}$ and $\tilde{Z}_{ik}$ are related to each other using $\tilde{Z}_{ik} = \tilde{X}A$, where, A is a *(p × p)* orthogonal matrix, k-th column of which is defined by the coefficient vector a_k. Thus, the k-th PC is denoted using $z_k = a_k^T x$, and a_k is an eigenvector of $\sum$ of its k-th largest eigenvalues denoted by λ_k. If, a_k is of unit length, i.e., if $a_k^T a_k = 1$, then, variance of z_k is equal to λ_k.

PCs are concerned on maximizing the variance; i.e., maximizing$z_k = a_k^T x = a_k^T \sum a_k$, following the constraint$a_k^T a_k = 1$; and this is executed using the technique of Lagrange multipliers as a standard method. Thus it maximizes the function

$$(a_k^T \sum a_k) - \lambda(a_k^T a_k - 1) \tag{6}$$

where, λ is a Lagrange multiplier. Differentiation is performed for maximizing the same function with respect to a_k, which gives,

$$\sum a_k - \lambda a_k = 0;\ \text{i.e.,} \left(\sum - \lambda I_p\right) a_k = 0; \tag{7}$$

where, $I_p$is an identity matrix with a dimension of *(p × p)*. Thus it gives λ and a_k as a set of eigenvalue and corresponding eigenvector of Σ respectively. The eigenvector with maximum variance of $(a_k^T x)$ is obtained by maximizing the term $(a_k^T \sum a_k)$, which in turn, leads to

$$,a_k^T \sum a_k = a_k^T \lambda a_k = \lambda a_k^T a_k = \lambda; \text{since, } a_k^T a_k = 1 \tag{8}$$

Thus, maximization of eigenvalue λ is required, i.e., the eigenvector $a_k$corresponding to the highest eigenvalue $\lambda_k$would lead to the highest PC $(a_k^T x)$, where, variance $\left(a_k^T x\right) = \lambda_k$; thus, k = 1 corresponds to the first PC and so on.

In this work, we have used PC1 only for extracting the image quality in terms of the PCA feature; thus making the analysis light in computational. These are further used to develop the threshold based classification model.

3.4 Metrics for performance analysis

The performance of the prediction or the classification efficiency through PCA is measured using the metrics like accuracy, precision, recall (or sensitivity), F-measure (or F-score), and specificity according to the equations as follows:

Accuracy (%): (TP+TN)/(TP+FP+FN+TN) (9)

Precision (%) = TP/(TP+FP) (10)

Recall (%) = TP/(TP+FN) (11)

Specificity (%) = TN/(TN+FP) (12)

F-measure (%) = 2*(Recall * Precision) / (Recall + Precision) (13)

Where, TP = true positive; TN = true negative; FP = false negative; FN = false negative; TN = true negative classes and obtained from the confusion matrix generated.

4. Results and analysis

4.1 Hue intensity plot

The hue levels obtained from the cropped images of fruit have distinct dissimilarity while considering images of different qualities of fruit samples. The hue histograms of the different quality fruit images are shown in Figure 2(a) and Figure 2(b) is a magnified view of Figure 2(a). We have shown the histogram plot of 15 samples only in these figures; although, we have used 15, 30 and 15 samples of *'Good'*, *'Intermediate'*, and *'Bad'* classes.

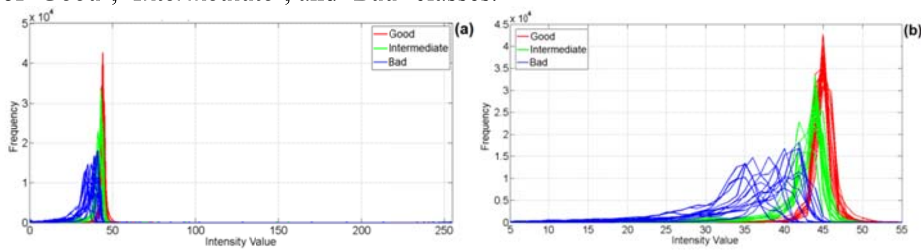

Figure 2 – (a) Hue intensity maps of different fruit images, (b) Magnified view

4.2 Analysis using PCA

The hue levels are further analyzed using Principal Component Analysis (PCA). We have observed that the hue intensity maps of the different quality of samples are significantly different. In this work, training has been carried out using a total of 60 number of fruit images comprising of three classes. Each image is represented by the corresponding histogram intensity map, which is further fed to the PCA algorithm to obtain the PC score index. Histogram count for each image contains the intensity plot of 0 to 255 levels. Hence, each histogram map is a vector of length 256, with intensity levels corresponding to each level. Thus the training matrix for the PCA algorithm takes the following shape:

$$[Tr] = \begin{bmatrix} h1_1 h2_1 \ldots hk_1 \ldots h60_1; \\ h1_2 h2_2 \ldots hk_2 \ldots h60_1; \\ \ldots \quad \ldots \qquad \ldots \qquad \ldots \qquad ; \\ h1_i h2_i \ldots hk_i \ldots h60_i; \\ \ldots \quad \ldots \qquad \ldots \qquad \ldots \qquad ; \\ h1_{256} h2_{256} \ldots hk_{256} \ldots h60_{256} \end{bmatrix}_{256 \times 60}$$

where, k defines the sample index; hence, k = 1, 2, … , 60; and k = 1, 2, … , 15 defines the *'Good'* condition fruits, k = 16, 17, … , 45 defines the *'Intermediate'* class, and k = 46, 47, … , 60 defines the *'Bad'* or *'Non-Edible'* class of fruits. Thus, altogether, k = 1, 2, …, 45 indicate the *'Edible'* category of fruits.

This training matrix [Tr] is fed to the PCA algorithm, which identifies the principal directions of variation in the descending order of importance. We have considered only the first or the most important direction in this work. Hence, only a single principal component index (PCI) is obtained as a representation of each vector. The training matrix is reduced to a single vector with 60 elements; each element representing one vector of the hue values of a single image. Thus, the PCI matrix is obtained as:

$[P] = [p1\ \ p2\ \ p3\ldots\ pk\ \ \ldots\ p60]_{1\times 60}$

This PCI vector is further segmented into '*Good*', *'Intermediate'*, and *'Bad'* classes. Thus, the PCI values of the three classes are found as:

$[P_G] = [p1\ \ p2\ \ p3\ldots\ p15]_{1\times 15}$
$[P_I] = [p16\ \ p17\ \ p18\ldots\ p45]_{1\times 30}$
$[P_B] = [p46\ \ p47\ \ p48\ldots\ p60]_{1\times 15}$

where, P_G , P_I and P_B denotes the '*Good*', *'Intermediate'*, and *'Bad'* classes of fruit images respectively. It is found from Figure 2 that the hue levels of 'Good', 'Intermediate' and 'Bad' fruit images, as denoted by the red, green and blue lines respectively, are separated. The hue levels of 'Bad' images are lesser in magnitude, as well as, shifted more towards the lower intensity values; whereas, the hue levels of *'Good'* images are higher in magnitude and shifted more towards the higher intensity values. These are readily observed from Figure 2(b). We have further used PCA to express these inferences mathematically. These indices are plotted in a two dimensional graph, where the independent axis represent the sample index of the fruit image and the dependent axis represent the first principal component index values (PCI). This plot is shown in Figure 3.

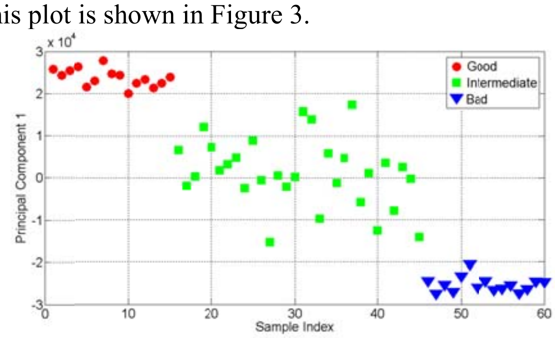

Figure 3 – Principal Component Score index plot

It is observed from Figure 3 that PCA works very efficiently to segment the hue levels corresponding to different qualities of samples, as observed from Figure 2(b). Figure 3 indicates that each cropped image of the three different conditions is represented by a single marked point using the proposed algorithm. The three different qualities of samples are also coloured separately in the same figure. This indicates clear distinction among the three different phases of progression using the score index plot. Besides, no overlap of the PCA index points is observed; although, there is partial overlap of the hue curves for the three quality levels as indicated by the partially overlapping red, green and blue curves of Figure 2(b). Thus, the effect of the partial overlap of the hue curves is nullified using PCA, rightly justified by the completely non-overlapping PCI.

4.3 Threshold based class segmentation

The PC1 score index plot, shown in Figure 3, reveals distinct segmentation of the three levels of quality. Thus, a threshold based analysis is thought of and is implemented here. Figure 3 shows that *'Bad'* class points could be separated from the rest using a horizontal line drawn through the intermediate region between *'Bad'* and *'Intermediate'* classes: marked as *'Threshold Line 1'* in Figure 4. The *'Good'* and *'Intermediate'* samples are broadly considered together as *'Edible'* category; and the *'Bad'* quality of samples are considered *'Non-Edible'* category. Hence, a second threshold is perceived in order to separates the *'Good'* class of fruit images from the *'Intermediate'* ones. The objective of the second threshold is to identify the freshest fruits from the *'Edible'* class, so that, the best quality samples could be identified while purchasing; although, this second threshold is lesser significance compared to the first threshold. We have marked this second threshold as *'Threshold Line 2'* in Figure 4 using another dotted red horizontal line. It is observed that the margin of clearance about this *'Threshold Line 2'* for both the corresponding classes: *'Good'* and *'Intermediate'*, is very low. It is observed from Figure 4 that *'Threshold Line 1'*segments the *'Edible'* and *'Non-Edible'* using the first principal component only. This method also enables reduced use of memory, as it considers the first PC only. Besides, the proposed method helps in direct identification of the test image between *'Edible'* and *'Non-Edible'* classes.

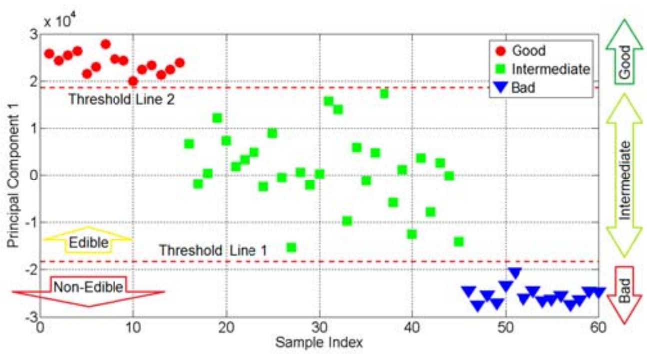

Figure 4 – Segmentation of Edible and Non-Edible sets using Principal Component Score index plot

The distribution of points along the PC1 axis is further represented using boxplots in Figure 5, which undoubtedly exposes the broad distinction between the

'Edible' and *'Non-Edible'* classes. The median levels of the two classes are also found largely apart. More importantly, the few outliers are also found to remain within the respective boundaries. The proposed steps of algorithm design are presented in Figure 6.

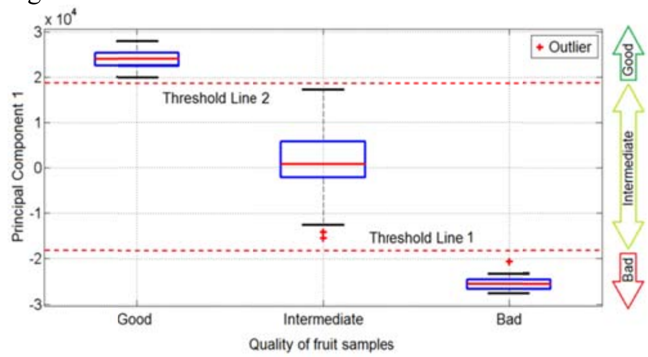

Figure 5 – Analysis of the Principal Component 1 score index values to develop segmentation of Edible and Non-Edible sets

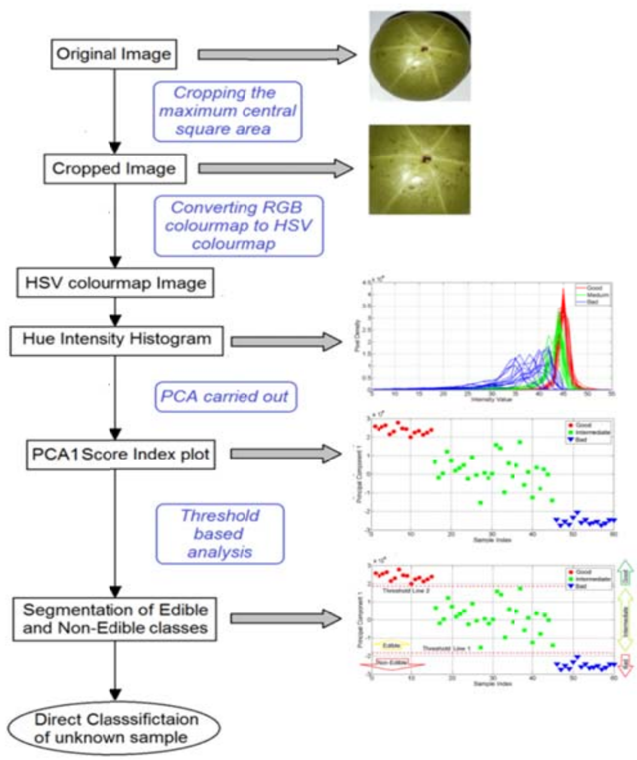

Figure 6 – Flowchart of the proposed scheme

4.4 Classifier outcomes

Validation of the proposed model has been carried out using the rest of the 60 number of samples. On an overall analysis, the proposed method is found to produce a very high accuracy as shown in table 1. It is observed from Table 1 that proposed model is able to classify the test image samples with 98.33% accuracy.

Table 1 – Result of the proposed classifier

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">Classifier Result</td></tr>
<tr><td>Good</td><td>Intermediate</td><td>Bad</td></tr>
<tr><td rowspan="3">Actual Result</td><td>Good</td><td>15</td><td>0</td><td>0</td></tr>
<tr><td>Intermediate</td><td>0</td><td>29</td><td>1</td></tr>
<tr><td>Bad</td><td>0</td><td>0</td><td>15</td></tr>
<tr><td colspan="5">Overall Accuracy: 98.33%</td></tr>
</table>

4.5 Metrics for performance analysis

Table 2 showed the performance of the proposed methodology. The accuracy of classification signifies only the presence or absence of errors. It is basically measured as accurate classification (TP+TN) efficiency of the proposed model on the overall dataset. The high accuracy 98.33% of the proposed model indicates the absence of errors in the model. Researchers often describe model robustness with specificity and recall. The high numeric value for recall and specificity stands for the flawless classification of the target classes. The recall is often denoted as the rate of true positive while the rate of true negative is termed as specificity. Both the specificity and recall provide the indication that how competently the fresh fruits were distinguished.

Table 2 – Result for the performance analysis of the proposed classifier

Parameter	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F-measure
Good	100	100	100	100	50
Intermediate	98.33	96.67	100	96.77	49.15
Bad	98.33	100	93.75	100	48.39

4.6 Effectiveness of using PCA

The effectiveness of using PCA in the proposed work is further justified from Figure 7 and Figure 8. Figure 7 shows an average representation of the frequency distribution of the hue intensity histogram, separately for the three major classes and Figure 8 is the representation of the distribution of the hue histogram frequency three classes of fruit samples using boxplots.

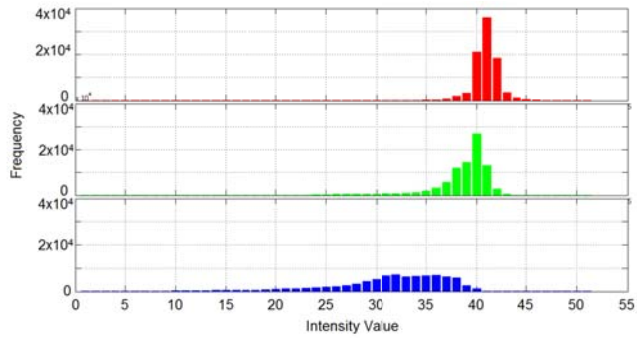

Figure 7 – Average representation of frequency distribution of the hue intensity histogram

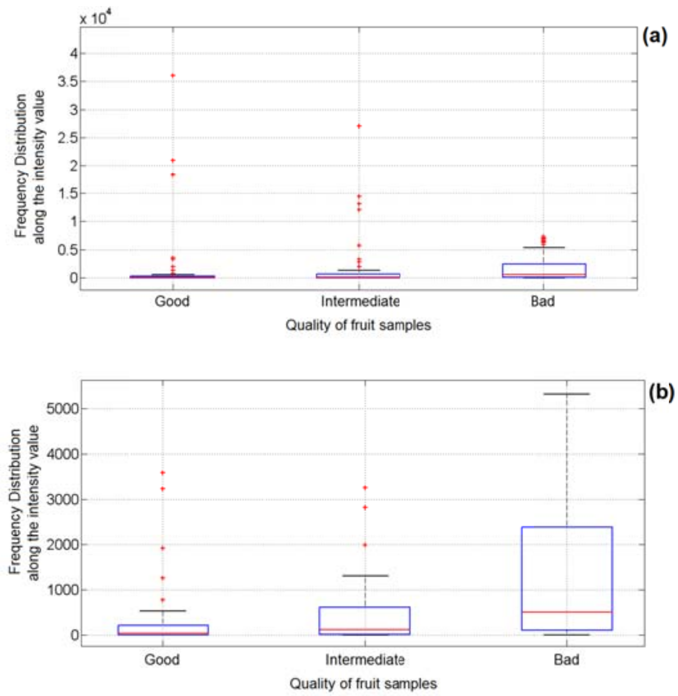

Figure 8 – (a) Boxplots representing the distribution of the hue histogram frequency three classes of fruit samples, (b) Magnified view of Figure 8 (a).

It is easily observed that there is considerable level of overlap in the range of intensity values; although, the centre of gravity of the area shifts towards the lower end of intensity values as the fruit belongs more towards the *'Bad'* class of image from the *'Good'* level. Besides, the distribution of magnitudes of the frequency plots although differs as visible from the three plots, but are not significant enough to develop distinctly separable thresholds to segregate the three classes. This distribution is further emphasized in Figure 8(a) which shows a boxplot of the frequency distribution from the hue histogram directly and Figure 8(b) is a magnified view of Figure 8(a). A comparative analysis of this boxplot with that of the same obtained using the principal component 1 values as shown in Figure 5, shows that the direct hue intensity boxplots of Figure 8 has large overlap in values between the three classes; although, the medians are separable. Thus, the above analysis proves that use of direct hue intensity plot doesn't provide distinctive separation among each of the three categories, whereas, application of PCA over these hue intensity features largely improves the classification accuracy. This proves the effectiveness of the proposed PCA-threshold based method for segregating the three classes.

4.5 Discussion

The most important features of the proposed works are described as follows:

- The present method is simple in analysis as it used PCA as the only tool for extracting the features from the hue histogram of the image. Only the first principal component is used in this analysis for identifying the key features of the hue intensity diagram; thus, reducing the computational burden, as well as requirement of large memory for computation.
- The proposed method is simpler in computation compared to several other popular techniques like the supervised learning models involving neural networks, which require large and diverse data set for accurate training; or the mathematically heavier transform based algorithms like wavelet or Fourier transforms, which demand for intricate analysis of signals. Hence, the proposed method is easier to implement in real life applications.
- Hue intensity map is analyzed here since it is less affected by the variation in light intensity level; which is quite common considering practical circumstances. In most of the cases, light intensity of the fruit image are prone to variation as the images are mostly taken in different condition of daylight, including variation due to sunny or overcast, indoor or outdoor condition etc.
- Thresholding method has been applied over the PC score based index plot, which is found to segregate the three classes distinctly. This shows the effectiveness of the proposed scheme.
- No specific pre-processing is required in this scheme for the fruit images. The images so taken are centre cropped only, followed by direct histogram analysis. No other filter especially meant for equalization of colour or light intensity is applied in this work. This reduces the cost computation to a large extent.
- The proposed method uses the images of fruit taken with Smartphone only, which is a more user friendly and convenient option, especially considering on-site availability and portability of the device.

- Only the hue level of each image is analyzed here out of the three layers of the image: hue, saturation and intensity. Analysis of only a single layer reduces the image layer depth to 33%; thereby, reducing time of computation.
- Finally, a high classification accuracy of 98.33% proves the high efficiency of the proposed method in identifying the quality of fruit among a set of different qualities of such samples.

5. Conclusion

A pattern identification algorithm is proposed in this paper for the detection of rotten members from a large set of amla (*Emblica officinalis*) images, captured using smart phones to avail their advantages of on-site availability and portability. The method uses only hue intensity histogram of the unprocessed cropped images and analyzed these features using principal component analysis. A threshold based implementation has been modeled using the PCA score based index features. The proposed method is simple as it avoids use of any complex mathematical, statistical analysis, or supervised learning models incorporating neural network based models. Non use of any specific pre-filter allows for reduced computational complexity. Besides, the hue intensity map, used in this analysis, is less prone to variation of ambient luminance; which is a key feature of practical robustness of the proposed scheme. Finally, a threshold based classification accuracy of 98.33% to detect the rotten specimen of fruit samples proves the effectiveness of the proposed scheme; as well as displays its applicability for real life application.

REFERENCES

[1] **Ariana, D. P., Lu, R. & Guyer, D. E. (2006),** ***Near-infrared Hyperspectral Reflectance Imaging for Detection of Bruises on Pickling Cucumbers;*** *Computers and electronics in agriculture*, *53*(1), 60-70;

[2] **Antonelli, A., Cocchi, M., Fava, P., Foca, G., Franchini, G. C., Manzini, D., & Ulrici, A. (2004),** ***Automated Evaluation of Food Colour by Means of Multivariate Image Analysis Coupled to a Wavelet-Based Classification Algorithm;*** *Analytica Chimica Acta*, 515(1), 3-13;

[3] **Bahramsoltani, R. & Rahimi, R. (2020),** ***An Evaluation of Traditional Persian Medicine for the Management of SARS-CoV-2;*** *Frontiers in pharmacology*, 11;

[4] **Baliga, M. S. & Dsouza, J. J. (2011),** ***Amla (Emblica officinalis Gaertn), A Wonder Berry in the Treatment and Prevention of Cancer.*** *European Journal of Cancer Prevention*, 20(3), 225-239;

[5] **Baranowski, P., Mazurek, W., Wozniak, J. & Majewska, U. (2012),** ***Detection of Early Bruises in Apples Using Hyperspectral Data and Thermal Imaging.*** *Journal of Food Engineering*, 110(3), 345-355;

[6] **Danti, A., Madgi, M., & Anami, B. S. (2012),** ***Mean and Range Color Features Based Identification of Common Indian Leafy Vegetables.*** *International Journal of Signal Processing*, Image Processing and Pattern Recognition, 5(3), 151-160;
[7] **Diaconeasa, M. C., Popescu, G. & Boboc, D. (2019),** ***Sustainable Food Consumption in the Web of Science Abstracts;*** *Economic Computation and Economic Cybernetics Studies and Research,* 53(1), 299-307; *ASE Publishing;*
[8] **Dubey, S. R. & Jalal, A. S. (2012),** ***Robust Approach for Fruit and Vegetable Classification.*** *Procedia Engineering,* 38, 3449-3453;
[9] **Dubey, S. R. & Jalal, A. S. (2015),** ***Application of Image Processing in Fruit and Vegetable Analysis: A Review.*** *Journal of Intelligent Systems,* 24(4), 405-424;
[10] **Geladi, P. & Grahn, H. F. (2006),** ***Multivariate Image Analysis.*** *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation;*
[11] **Goraya, R. K. & Bajwa, U. (2015),** ***Enhancing the Functional Properties and Nutritional Quality of Ice Cream with Processed Amla (Indian gooseberry).*** *Journal of food science and technology,* 52(12), 7861-7871;
[12] **Gowen, A. A., O'donnell, C. P., Taghizadeh, M., Cullen, P. J., Frias, J. M. & Downey, G. (2008),** ***Hyperspectral Imaging Combined with Principal Component Analysis for Bruise Damage Detection on White Mushrooms (Agaricus bisporus);*** *Journal of Chemometrics: A Journal of the Chemometrics Society,* 22(3-4), 259-267;
[13] **Kapoor, M. P., Suzuki, K., Derek, T., Ozeki, M. & Okubo, T. (2020),** ***Clinical Evaluation of Emblica Officinalis Gatertn (Amla) in Healthy Human Subjects: Health Benefits and Safety Results from a Randomized, Double-Blind, Crossover Placebo-Controlled Study.*** *Contemporary clinical trials communications,* 17, 100499;
[14] **Liu, Y., Chen, Y. R., Wang, C. Y., Chan, D. E. & Kim, M. S. (2006),** ***Development of Hyperspectral Imaging Technique for the Detection of Chilling Injury in Cucumbers; Spectral and Image Analysis;*** *Applied engineering in agriculture,* 22(1), 101-111;
[15] **López-García, F., Andreu-García, G., Blasco, J., Aleixos, N. & Valiente, J. M. (2010),** ***Automatic Detection of Skin Defects in Citrus Fruits Using a Multivariate Image Analysis Approach;*** *Computers and Electronics in Agriculture,* 71(2), 189-197;
[16] **Novotna, M. (2020),** ***The Impact of Industry and Business Entity Type on Corporate Survival;*** *Economic Computation & Economic Cybernetics Studies & Research*, 54(1);
[17] **Patras, A., Brunton, N. P., Downey, G., Rawson, A., Warriner, K. & Gernigon, G. (2011),** ***Application of Principal Component and Hierarchical Cluster Analysis to Classify Fruits and Vegetables Commonly Consumed in Ireland Based on in Vitro Antioxidant Activity;*** *Journal of Food Composition and Analysis*, 24(2), 250-256;

[18] **Sun, Y., Wei, K., Liu, Q., Pan, L. & Tu, K. (2018),** ***Classification and Discrimination of Different Fungal Diseases of Three Infection Levels on Peaches Using Hyperspectral Reflectance Imaging Analysis;*** *Sensors,* 18(4), 1295;

[19] **Wu, W., Massart, D. L. & De Jong, S. (1997),** ***The Kernel PCA Algorithms for Wide Data. Part I: Theory and Algorithms;*** *Chemometrics and Intelligent Laboratory Systems,* 36(2), 165-172;

[20] **Xing, J., Bravo, C., Jancsók, P. T., Ramon, H. & De Baerdemaeker, J. (2005),** ***Detecting Bruises on 'Golden Delicious' Apples Using Hyperspectral Imaging with Multiple Wavebands;*** *Biosystems Engineering,* 90(1), 27-36;

[21] **Zhao, J., Ouyang, Q., Chen, Q. & Wang, J. (2010),** ***Detection of Bruise on Pear by Hyperspectral Imaging Sensor with Different Classification Algorithms;*** *Sensor Letters,* 8(4), 570-576;

[22] **Zhu, B., Jiang, L., Luo, Y. & Tao, Y. (2007),** ***Gabor Feature-Based Apple Quality Inspection Using Kernel Principal Component Analysis;*** *Journal of Food Engineering,* 81(4), 741-749.

[23] **Sarkar, T., Mukherjee, A., Chatterjee, K. (2021), Supervised Learning Aided Multiple Feature Analysis for Freshness Class Detection of Indian Gooseberry (Phyllanthus emblica),** Journal of The Institution of Engineers (India): Series A, 102(3), 247-261.

[24] **Mukherjee, A., Sarkar, T., Chatterjee, K. (2021), Freshness Assessment of Indian Gooseberry (Phyllanthus emblica) Using Probabilistic Neural Network,** Journal of Biosystems Engineering, 46(4), 399-416.