Quality Analysis of Fruits and Vegetables using Machine Learning Techniques

Richa Shah[1], Pooja Gujarathi[2], Parth Unadkat[3], Rijoosinh Mulik[4], A. N. Bandal[5]
[1,2,3,4]*Student, Department of Computer Engineering, Sinhgad Institute of Technology, Pune, India*
[5]*Professor, Department of Computer Engineering, Sinhgad Institute of Technology, Pune, India*

***Abstract*—The Quality of fruits or vegetables plays an important role in consumer consumption and thereby affecting its sales. In India most of the population survival is based on agricultural products. All the business and organizations that make, display, transport or prepare food for sale they will need to check food quality. If we are able to identify the maturity of fresh fruit, then it will be very beneficial to farmers as they can optimize their harvesting. This ability will help them avoid harvesting under-matured or over-matured fruits. It makes an attempt to use image processing techniques to extract colour, size and other attributes of the image forming training dataset. Then using supervised and unsupervised learning we form trained data from training dataset. Further the new image of a fresh fruit whose quality is to be predicted undergoes image processing later we apply machine learning algorithm (Deep Neural Network) on the extracted attributes, referring its outcome and trained dataset quality will be predicted.**

Index Terms—Machine Learning

I. Introduction

Fruits and vegetables contain lots of essential vitamins and minerals that are not found in other type of foods and they may also contain more of these nutrients than other foods.

The economic value of plant derived food depends on its quality and how it is preserved over the whole production chain, until it reaches the final customer. The concept of quality is wide and covers several aspects such as external appearance, nutritional aspects, and presence of health-related compounds, safety and security.

II. Literature Survey

In earlier times, many types of image analysis techniques have been applied for analysis of the images in field of agriculture such as fruits and vegetables, for classification and recognition purposes.

In 2009, Woo Chaw Seng, Seyed Hadi Mirisaee[7] proposed a Fruit Recognition System. The system was applied on fifty image samples. The mean color values of these images are computed. The fruit area and perimeter are chosen as features to distinguish one fruit from another. Their system mainly consisted of five main processing modules, which are, fruit input selection module, fruit color computing module, fruit shape computing module, fruit size computing module, and fruit classification or recognition module. It used the KNN algorithm for classification and recognition of the input fruit. The recognition results were accurate up to 90%.

Saswati Naskar and Tanmay Bhattacharya[1] have suggested a method for fruit recognition using three main features. The features used here are texture, color and shape. One hundred and fifty images have been selected for analysis. Neural networks have been used for classification purpose. Through this approach most of the results are true without any confusion. As per study carried out at EC Department, G.H. Patel College of Engineering & Technology, Vallabh Vidyanagar, by Rohit R. Parmar, Kavindra R. Jain, Dr. Chintan K. Modi [5], their paper presents a unified approach for quality evaluation of food using image processing and machine vision. They have captured images of various food products by vision system to identify, analyze and assess the quality. According to their reviews on fruits and vegetable analysis, they found that color plays a major role in identifying the age of fruits. P. Sudhakara Rao et al, proposed a new method for inspection of color of apples named as Improved Radius Signature[5]. This method compares many samples against a reference shape for the purpose of sorting and grading using correlation coefficient technique, Graphical analysis and Fourier transformation technique. After their studies they have observed that parameters like area, shape, size, color, shape compactness are directly affecting the quality & black moles, brown undaubed shapes are inversely proportional to the quality parameter of fruits & vegetables. Hence, they propose a generalized formula with weights of different features. The weights are set as per the food quality requirement. The constants c1 and c2 play a major role in defining final grades from the quality. An automated framework has been proposed by Amol Bandal and Mythili Thirugnanam[8] to predict food quality by using a multi-sensor network.

A. AL-Marakeby, Ayman A. Aly, Farhan A. Salem[6] proposed a fast system for inspection of food products using computer vision. This system was applied on four products namely apple, tomatoes, eggs and lemons. A dataset comprising of about 1000 images (250 for each system) is used to train and test the vision-based sorting system. In their system, the captured images are sent to the computer to be processed and analyzed in real time. The decision, "pass" or "fail", is sent as an electronic signal to the interfacing circuits. During classification, they have considered color as the major factor. A neural network is trained to give the final results if the product is accepted or not. The back-propagation algorithm is used to train the neural network. D. Surya Prabha and J. Satheesh Kumar[3] have developed a maturity classification algorithm based on color and size and using image processing methods. Their works are confined to assessment of maturity of the banana fruit. Sample collection and image acquisition are major tasks in determining maturity of bananas. The banana fresh fruit bunches can be broadly classified into three main categories namely: (a) ‘under-mature’ with the age of less than

12 weeks after flowering and having dark green color. (b) 'mature' with the age of 12–17 weeks after flowering characterized by pale green color and (c) 'over-mature' with the age of <17 weeks after flowering characterized by yellowish green color. They captured about 120 images, 40 from each category. 60 images were used as calibration images which were used to develop the algorithm and remaining 20 each were used as validation images. The color mean intensity value, area, perimeter, major axis length and minor axis length data were collected from 20 CI in each category. Two classifier algorithms were developed from the datasets of color mean intensity value and area. The datasets were analyzed using box and whisker plot technique. The maturity determination of banana was automated using the Graphical User Interface Development Environment (GUIDE) of matlab7.10. Mean color intensity and area algorithms were developed and tested for the accuracy on determination of fruit maturity. Testing of these two classifier algorithms indicated that mean color intensity algorithm was more accurate with overall accuracy of 99.1 %. The area algorithm was accurate up to 85 % for differentiating under-mature banana, but unsuccessful to distinguish between mature and over-mature category.

TABLE I
LITERATURE SURVEY

Title	Methodology	Fruit/Vegetable	Accuracy
A Fruit Recognition Technique using multiple features and Artificial Neural Network (2015)	Artificial Neural Network	Orange, apple, pineapple	90%
Assessment of Banana Fruit Maturity by Image Processing Technique (2013)	Mean colour Intensity Algorithm and Area Algorithm	Bananas	99.1% in classifying fruit maturity and 85% under mature fruit
Fast Quality Inspection of Food Products using Computer Vision (2013)	A neural network is trained based on these features to give the final decision if this product is accepted or not	Apples, lemons, tomatoes, eggs	97% with speed up to 200 images/minute
Unified Approach in Food Quality Evaluation using Machine Vision	Artificial Neural Networks	Peanuts, Rice	92%
A New Method for Recognition System	KNN	Apple, Strawberry, banana, lemon	90%

III. PROPOSED MODEL

A) Phase I: Preprocessing phase

We collected about 50-60 image samples of the same item, all of varying values of attributes from a fixed distance and using the same device. These images serve as an input to image processing through which we obtained training dataset by extracting various attributes of images, such as area, roundness, skewness, kurtosis, color intensity etc. This phase includes processing on image such as background removal, fruit image extraction, color value extraction, as well as size extraction.

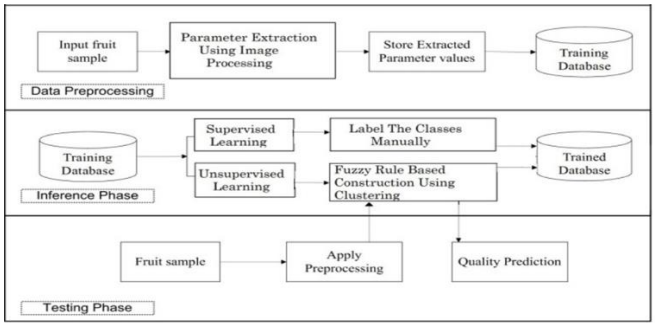

Fig. 1. Architecture Diagram

Working Paradigm:

At the very first the polygon of the image is clipped.

Fig. 2. Working paradigm

This clipped image undergoes following processes.

1) Background Removal:

Background was removed from the image using a threshold value and converted the RGB image into binary image.

The converted binary image had 0's in the banana region and 1's in the background region.

2) Extracting fruit region:

The binary image is converted into the complement image with 1's in the region of banana and 0's in the background region.

Complement image is used to identify the boundary of banana region.

While creating the training dataset we will extract all the possible attributes/features of the images. The relationship among these attributes, knowing which attributes are useful to

extract to decide the quality of food item from the next coming input image will be decided. This decision will be made using the algorithm.

Multivariable regression model:

In this model there are number of independent variables (Xs) in the RHS of equation and one dependent variable in the (Y) in the LHS of the equation.

$$y = m_1X_1 + m_2X_2 + m_3X_3 + m_4X_4 + ... + m_nX_n + c \quad (1)$$

Here X_1, X_2, X_3, X_4, ..., X_n are all the possible attributes of the image, 'c' is the constant and Y's value tells us the about the quality of the image.

Regression analysis include a list of steps to be followed:

Step1: Generate list of potential independent and dependent variables.
Step 2: Collect the data to form a dataset. Here we have taken about 50 sample data in a csv file format.
Step 3: Check relationship between each independent variable and dependent variable using scatterplots and correlation matrix.

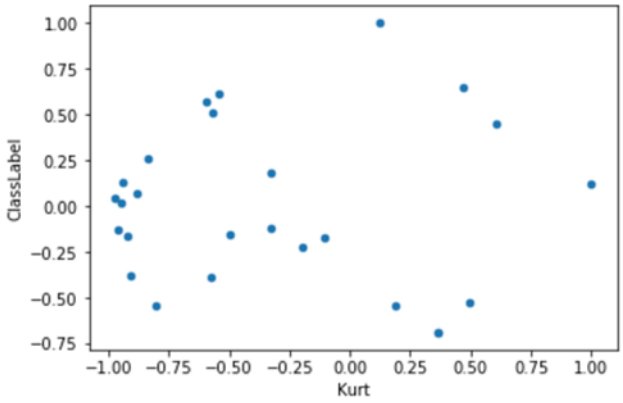

Fig. 3. Kurt vs. Class label

The scatterplot above shows that kurtosis (X) is not showing any relation with class label(Y).

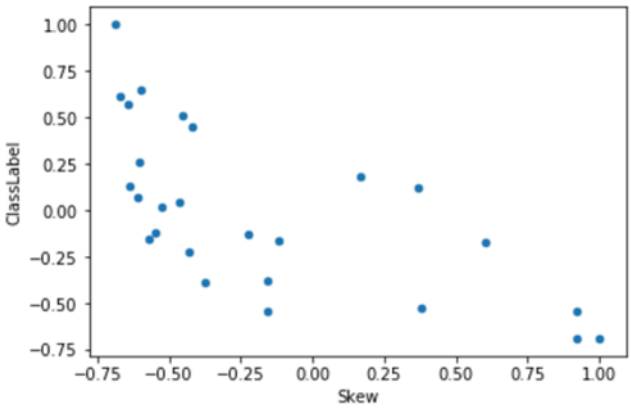

Fig. 4. Skew vs. Class label

Here we observe a kind of relationship between the two variables. From this we can conclude that skewness (X) can be used to determine Class Label(Y)

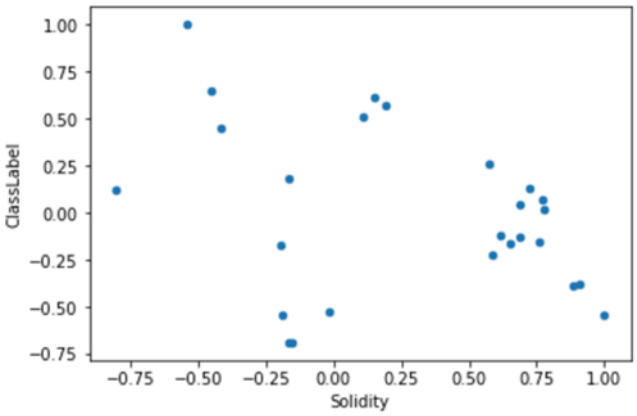

Fig. 5. Solodity vs. Class label

Step 4: Check Relationships among independent variables using scatterplots and correlations.

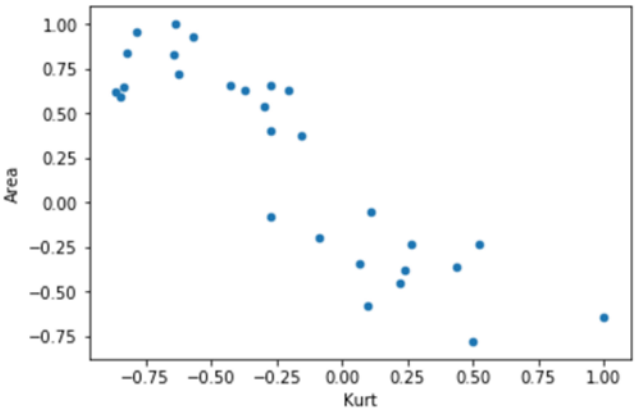

Fig. 6. Kurt vs. Area

Here we can observe a relationship between the two variables which means that we can use any one of these to uniquely determine the independent variable.

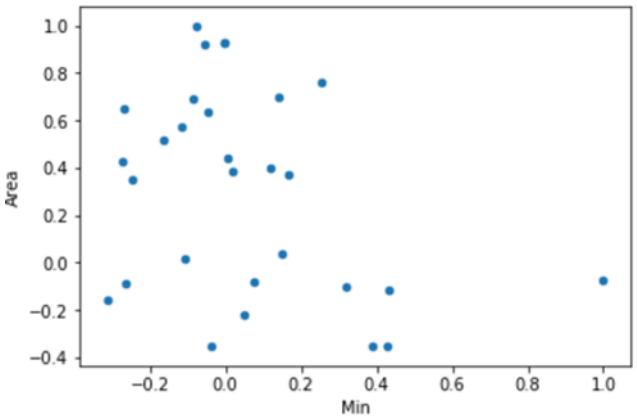

Fig. 7. Min. vs. Area

Since there is no relation between attributes in the above diagram, we cannot eliminate any of these two as they both separately determine the output variable.

The attributes include:

Mean of Intensity:

Mean is used to measure the average intensity value of the pixel distribution.

$$y = \sum_{n=0}^{L-1}(x_n t(x_n)) \quad (2)$$

'L' is the number of possible intensity values, 'x_n' is the discrete variable represents intensity level in an image and 'y' is the mean. t (x_n) is the probability estimation of occurrence of 'x_n'.

Variance:

Variance is used to measure how wide the pixels spread over from the mean value.

$$\mu_2 = \sum_{n=0}^{L-1}((x_n - y)^2\, t(x_n)) \quad (3)$$

Smoothness:

Image with constant intensity had a smoothness of '0' and the image with irregular intensity had a smoothness of '1'.

$$R = (1 - (1 / (1 + \mu_2(x)))) \quad (4)$$

Skewness:

Skewness is used to measure asymmetric shape of the distribution.

$$\mu_3 = \sum_{n=0}^{L-1}((x_n - y)^3\, t(x_n)) \quad (5)$$

Kurtosis:

Kurtosis is used to measure the peak of the distribution.

$$\mu_4 = \sum_{n=0}^{L-1}((x_n - y)^4\, t(x_n)) \quad (6)$$

Roundness:

It measures how closely the shape of an object approaches that of a mathematically perfect circle.

Solidity:

The proportion of number of pixels in the convex hull that are also in the region.

Solidity = Area/ (Convex Area) (7)

Area: Actual number of pixels in the region.
Convex Area: Number of pixels in convex area.

While preprocessing the new image whose quality we have to predict, after performing the activity of background removal and fruit region extraction we will extract only those attributes that we figured out using multi-regression model.

B) Phase II: Analysis phase

At the very first the retrieved attributes which is the training dataset will be used to create trained dataset.

The trained data will be created by following two methods:

Manually classifying the dataset with the labels:

a) Eatable
b) Non-eatable
c) May or may not be eatable

This is supervised learning where we have input variables (image attributes) and output variable (labels) we will use algorithm to derive the mapping function between them.

Clustering:

Clustering is a machine learning algorithm to quickly predict groups in data set. Predictions in this case are based on :-

1) Number of cluster centers present(k)
2) Nearest mean values (Euclidean Distance)

When using K-means-

1) Scale your variables.
2) Look at the scatterplot or data table to estimate the number of centroids or cluster centers, to set k parameters in model.

In our model, we use those attributes of the images which we have obtained as a result of regression analysis. The sample of results obtained from clustering are:

TABLE II
CLUSTERING

	index	Area	Skew	Kurt	MinFeret	AR	Round	Solidity	cluster	x	y
0	0	3475512	0.948	0.146	1618.604	2.693	0.371	0.746	2	-163.051367	0.548200
1	1	4076744	0.814	-0.405	1749.346	2.535	0.394	0.771	0	-32.303742	-0.827437
2	2	2872120	1.081	0.475	1600.456	2.304	0.434	0.732	2	-181.198626	0.381998
3	3	3904000	0.820	-0.111	1637.732	2.717	0.368	0.792	2	-143.923487	0.679653
4	4	3595635	1.019	0.101	1693.268	2.596	0.385	0.725	2	-88.387697	0.782611
5	5	3424080	1.092	0.279	1595.142	2.694	0.371	0.752	2	-186.513277	0.440902
6	6	3624555	0.669	-0.219	1681.807	2.865	0.349	0.709	2	-99.848953	0.847954
7	7	9734400	0.071	-0.989	2340.000	1.778	0.563	1.000	1	558.345095	2.056162
8	8	3333739	0.437	-0.429	1880.254	2.152	0.465	0.663	0	98.604262	-0.447337
9	9	3660266	0.346	0.601	2041.864	2.117	0.472	0.644	1	260.210327	0.716055
10	10	3536406	-0.145	-0.255	1912.000	2.278	0.439	0.658	0	130.350096	-0.395147
11	11	3982883	0.782	0.503	1739.436	2.704	0.370	0.726	0	-42.213917	-1.101609
12	12	3295359	0.778	0.243	1779.477	2.474	0.404	0.654	0	-2.172779	-0.922038
13	13	3822289	0.638	-0.537	1960.782	2.139	0.468	0.681	0	179.131693	-0.171437
14	14	3335459	0.524	0.206	1593.734	2.797	0.357	0.715	2	-187.921210	0.458269
15	15	3572111	0.692	0.075	1590.893	2.992	0.334	0.741	2	-190.762602	0.497540
16	16	3497341	0.619	0.286	1872.563	2.282	0.438	0.672	0	90.913039	-0.660611

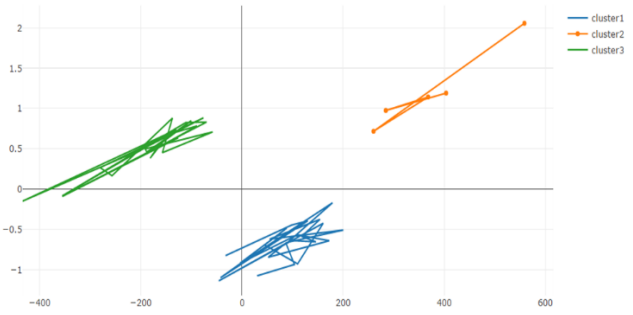

Fig. 8. Clustering

Classification using Machine Learning algorithm into above classes:

This is an unsupervised learning process where we have

input variables (attributes) only and we don't have output variable (Label) and model the distribution in data set.

Further when we need to predict the quality of any new fruit/vegetable, using this machine learning algorithm we will identify the appropriate class by using trained data set as well. The algorithm that we will be using is Deep Neural Network.

Deep Neural Network:

It is a type of feed-forward artificial neural network that is simulates human brain neural system that how neurons communicate, and information is processed.

The neural network includes three kinds of layers:

Input layer: It consists of as much number of neurons as number of input parameters. In our case the input parameters will be input image attributes.

Hidden Layers: These are the middle layers that accept input passed by input layer, transform them using some function and sends this output as input to the next hidden layer.

Output Layer: This layer contains as much number of neurons as number of classes/labels.

In our case we have three labels so three neurons are there in output layer.

IV. Conclusion

The proposed framework aims to predict the quality of fruits and vegetables to a great accuracy. The quality of fruits or vegetables plays an important role in consumer consumption and thereby affecting its sales. In India most of the population survival is based on agricultural products. All the business and organizations that make, display, transport or prepare food for sale they will need to check food quality. If we are able to identify the maturity of fresh fruit, then it will be very beneficial to farmers as they can optimize their harvesting. This ability will help them avoid harvesting under-matured or over-matured fruits. This study attempted to use image processing techniques to extract colour, size and other attributes of the image forming training dataset. We use regression analysis to identify dependencies among variables. Here we use multivalued regression models that considers multiple independent variables and one dependent variable. Through this we can find out which attribute best describe the image analysis. Then using supervised and unsupervised learning we form trained data from training dataset.

This involves manual labelling of images to classify them into classes. This comes under supervised learning. Further we also use clustering which is a part of unsupervised learning to again form clusters. These clusters aim at maximizing the intra-cluster similarity and minimizing inter-cluster similarities. These clusters are matched with the manually labelled dataset which will then help us to identify the attributes of fruits and vegetable images belonging to different classes. Further the new image of a fresh fruit whose quality is to be predicted undergoes image processing later we apply machine learning algorithm (Deep Neural Network) on the extracted attributes, referring its outcome and trained dataset quality will be predicted.

References

[1] S. Naskar and T. Bhattacharya, “A Fruit Recognition Technique using Multiple Features and Artificial Neural Network,” *International Journal of Computer Applications*, vol. 116, no. 20, pp. 23-28, April 2015.

[2] N. B. Ahmad Mustafa, F. Bakri and S. K. Ahmed, "Identification of image angle using Projective Transformation: Application to banana images," *2014 IEEE REGION 10 SYMPOSIUM*, Kuala Lumpur, 2014, pp. 408-413.

[3] D. S. Prabha and J. Satheesh Kumar, “Assessment of Banana Fruit Maturity by Image Processing Technique,” *Journal of Food Science and Technology*, vol. 52, no. 3, pp. 1316-1327, 2015.

[4] L. Wang, X. Tian, A. Li and H. Li, "Machine Vision Applications in Agricultural Food Logistics," *2013 Sixth International Conference on Business Intelligence and Financial Engineering*, Hangzhou, 2013, pp. 125-129.

[5] R. R. Parmar, K. R. Jain and C. K Modi, “Unified Approach in Food Quality Evaluation using Machine Vision,” *in Communications in Computer and Information Science* July 2011.

[6] A. AL- Marakeby, A. A. Aly and F. A Salem, “Fast Quality Inspection of Food Products using Computer Vision,” *International Journal of Advance Research in Computer and Communication Engineering*, vol.2, no. 11, pp. 4168-4171, November 2013.

[7] W. C. Seng and S. H. Mirisaee, “A New Method for Fruits Recognition System,” *2009 International Conference on Electrical Engineering and Informatics*, 5-7 August 2009, Selengore, Malaysia.

[8] A. Bandal and M. Thirugnanam, “Quality Measurements of Fruits and Vegetables using Sensor Network,” in *Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC – 16')*, pp. 121-130.

[9] A. Sanaeifar, A. Bakshipour and M. Guardia, “Prediction of Banana Quality Indices from Color Features using Support Vector Regressions,” *Talanta*, vol. 148, pp. 54-61, February 2016.

[10] C. Xie, B. Chu and Y. He, “Prediction of Banana Color and Firmness using a Novel Wavelengths Selection method of Hyperspectral Imaging,” *Food Chemistry*, vol. 245, pp. 132-140, April 2018.