



A Machine Learning Approach to Determine Maturity Stages of Tomatoes

KAMALPREET KAUR* and O.P. GUPTA

Department of School of Electrical Engineering and IT, COAE&T, PAU Ludhiana, Punjab, India.

Abstract

Maturity checking has become mandatory for the food industries as well as for the farmers so as to ensure that the fruits and vegetables are not diseased and are ripe. However, manual inspection leads to human error, unripe fruits and vegetables may decrease the production [3]. Thus, this study proposes a Tomato Classification system for determining maturity stages of tomato through Machine Learning which involves training of different algorithms like Decision Tree, Logistic Regression, Gradient Boosting, Random Forest, Support Vector Machine, K-NN and XG Boost. This system consists of image collection, feature extraction and training the classifiers on 80% of the total data. Rest 20% of the total data is used for the testing purpose. It is concluded from the results that the performance of the classifier depends on the size and kind of features extracted from the data set. The results are obtained in the form of Learning Curve, Confusion Matrix and Accuracy Score. It is observed that out of seven classifiers, Random Forest is successful with 92.49% accuracy due to its high capability of handling large set of data. Support Vector Machine has shown the least accuracy due to its inability to train large data set.



Article History

Received: 14 July 2017

Accepted: 30 July 2017

Keywords

Tomato Classification, Machine Learning, Image Processing, Classifiers, Python, Learning Curve, Confusion Matrix.

Introduction


Agriculture in India constitutes the major part of the economy. Tomato is globally cultivated as protective fruit because of its better nutritive value. So, farmers are compensated only for the good quality fruits and vegetables. Tomatoes are consumed directly as raw vegetables or used in a number of processed products like ketchup, juices, soup, paste, etc. Therefore, maturity checking has become mandatory

for the food industries as well as for the farmers so as to ensure that the fruits are not diseased and are ripe. Another additional benefit of evaluating ripeness stage of tomato is to decide its shelf life, so as to use it for stocking purpose⁵.

Present day, in some cases manual inspection takes place which is time consuming for the farmers. Not only this, it also has some additional disadvantages like human error due to large quantity, unripe fruits

CONTACT Kamalpreet Kaur ✉ kamalpreet-seeit@pau.edu 📍 Department of School of Electrical Engineering and IT, COAE&T, PAU Ludhiana, Punjab, India.

© 2017 The Author(s). Published by Enviro Research Publishers

This is an  Open Access article licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (<https://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits unrestricted NonCommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

To link to this article: <http://dx.doi.org/10.13005/ojcs/10.03.19>

or vegetables might decrease the production. This would result in bad quality products that will create great loss to the farmers as well as to the manufacturing concerns¹¹. Visual appearance is the primary source of information and can be broken down by image analysis to parameters like size, color, shape, deformities, and abnormalities⁶. With the approach of quick and high precision machine learning, automation of the reviewing procedure is required to diminish work cost, enhance the proficiency and accuracy of the sorting process¹⁴. Color being a sign of tomato maturity, change from green to red-mature which is used to evaluate the stage at which it should be harvested for consumption¹. As per Jaramillo et al (2007) the tomato maturity is divided into five stages:

Stage 1 - Green mature

The complete surface of tomato is green.

Stage 2 - Breaking/Turning

A color other than green appears on the tomato surface. Between 10-30% of the tomato surface is pale yellow in color.

Stage 3 - Pink

Between 30-60% of the tomato surface shows a pink or light red color.

Stage 4 - Light red

Between 60-90% of the tomato surface is red in color.

Stage 5 - Red

More than 90% of the surface is red⁷.

Figure 1 shows all the maturity stages of tomato.



Fig. 1: Tomato Maturity Stages

Analysis of Tomato Maturity by Machine Vision

With recent advances in computer technologies, present day food industries manufacture has turned their consideration regarding machine-vision examination frameworks. The automated grading system not only speeds up the processing time, but also reduces the human error². Tomato classification through machine learning will be a step forward towards enhancing the practical knowledge for evaluating the maturity of tomato using image processing technique. It would reduce the human errors up to great extent and the software will help the repetitive work in an easy way. Likewise, a neural system based machine color vision is great investigation frameworks for tomatoes⁴.

More image features can produce more accurate results in this non-destructive fruit quality evaluation

system⁸. A high-quality image can help to reduce the time and complexity of the subsequent image processing steps, which can diminish the cost of an image processing system⁵.

Design and development of the proposed software has been implemented using Pycharm which is an IDE for many applications whose community edition is open-source for research and scientific developments. Python is a fast processing functional programming language that has superior data handling capabilities in comparison to MATLAB. It has huge library set and simple to use. Python and its modules like Numpy, Scipy, Matplotlib and other special modules provide the satisfactory functionality that enables us to deal with plenty of images with their features.

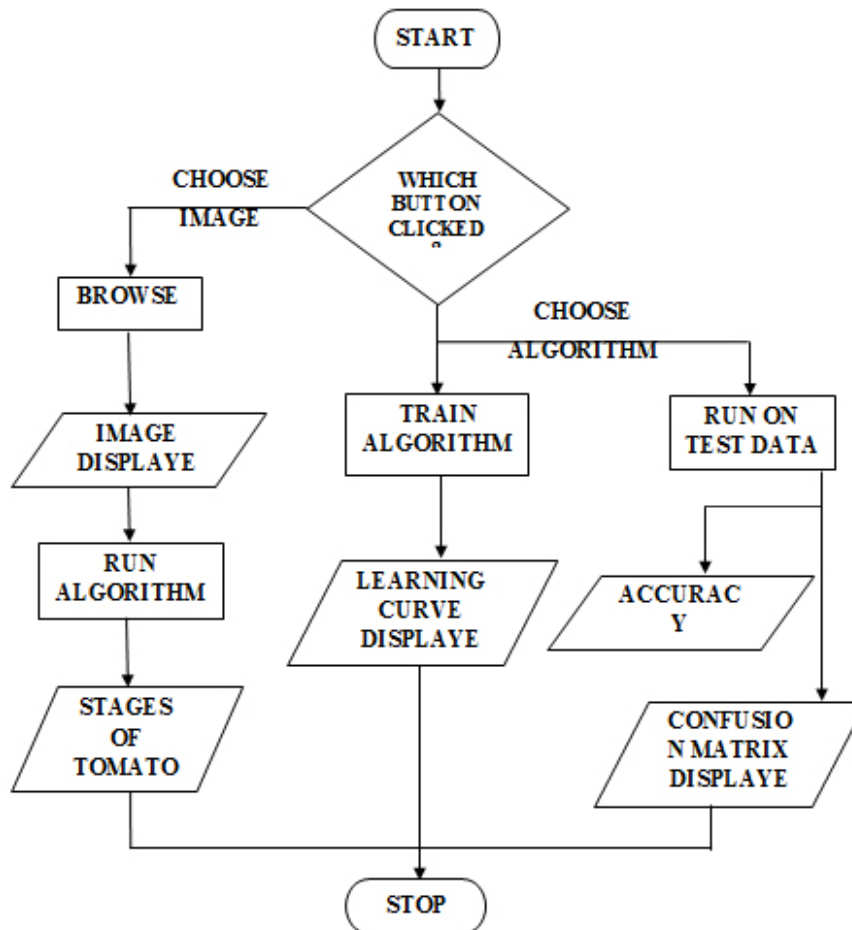


Fig. 2: Flow Chart of Interface Design

Materials and Methods

Proposed System

The main aim of the proposed approach is to provide an automated classification system for tomato ripeness evaluation through classifying the different maturity stages based on color features. The dataset used for this study is based on real sample images of 'Punjab Ratta' variety of tomato at different stages, which were collected from research field of Punjab Agricultural University, Ludhiana. The proposed approach consists of four phases namely Data collection, Pre-processing, Feature extraction and Classification phase.

- In the first step, sample images for each maturity stage of tomato is taken to generate more images by rotating/flipping/zooming etc. using Keras Library.
- In the second step, the average values of the 100th, 500th (horizontally & vertically) pixel is taken for every channel (RGB) in the image.
- In the next step, the average values is taken as a feature from step-2 & created a CSV file for each stage of the Tomato. So there are five CSV files (one for each class). Every CSV file has an extra column "label". Values of label column are 1, 2, 3, 4, and 5 indicating the different stages of tomato. For example, if we have taken the feature for Green stage in CSV file then label column of this CSV will have value 1. Similarly, for Breaker stage label column has value 2 and so on.
- In fourth step, each CSV file is combined in a single CSV named TomatoData.csv
- In the last stage, machine learning models are built using Scikit-learn package.

The results have been shown in the form of learning curve and confusion matrix that shows the accuracy of the predictive algorithm used. Learning curve compares the performance of a model on training and testing data over a number of training instances. The performance improves as sample training size increases. A confusion matrix shows a table that is used to calculate the accuracy of a classification model on a set of test data for which the true values of all the stages of tomato are known. By inputting the image, the system tells the maturity stage to which tomato belongs classified by these algorithms separately.

Implementation Plan

In implementation design, each component is sufficiently described to allow for its coding. This

design provides an ongoing plan for assessing components of the program. It is an action plan to achieve program outcomes and program learning outcomes. The implementation plan breaks each strategy into identifiable steps, suggests the ways for the completion of each step individually and then combining them in order to have a fully developed Tomato Classification system. This plan is developed during the design phase and updated during the development phase.

Figure 3 shows the design of complete methodology that is implemented in Tomato Classification system.

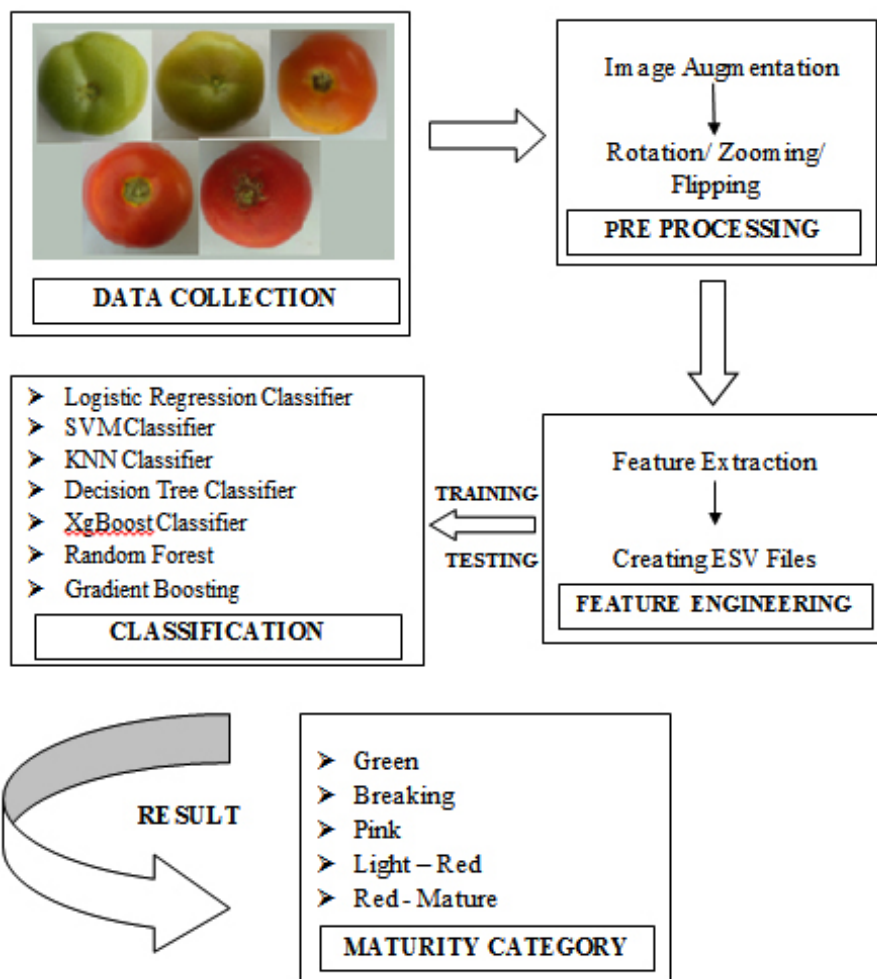


Fig. 3: Flow Chart of Methodology Implemented

Data Collection

The data being collected should not be biased and collected in appropriate quantity that is used to predict the exactness of the system. The more data an algorithm has, the more accurate it becomes. In the proposed system, if the training data is limited, it may not be able to support the model complexity needed for the problem. Improper data collection will reduce the ability to build predictive, accurate machine learning models.

For Tomato Classification, 100 images of each maturity stage are captured that makes total of 500 images. From these existing images, more number of images is generated in order to train the models well. For given 5 classes, around 1000-1200 images are generated for each class. Therefore total images generated become approximately 5000.

Image Augmentation

Some additional images are constructed from the existing raw data to generate a large training dataset as more number of training set will give more accuracy. The new images are generated to provide better information that is not clearly defined by the sample dataset. This process is done by using the Pandas Python package. The Tomato Classification data need to be loaded from the database into the Panda data frame and then it is processed further. The Pandas library in Python gives a huge set of data structures and analysis tools for data handling in Python programming.

From the given set of images more images are generated by using the Image Generator method of Keras which is a deep learning package for Python. Different techniques that are used to generate images are

- By rotation
- By flipping the image horizontally
- By zooming some random pixels of the image

Feature Extraction

In Tomato Classification system, the most important characteristic to evaluate tomato maturity is its surface color. This system uses High, Variance and Average values of 100th and 500th pixel for every RGB channel in the image. For this purpose, Scikit-learn library is used for feature extraction in order to

create a CSV file for each stage of tomato. Fifteen color features, five for each channel (R, G and B) (100H, 100V, 500H, 500V and avg) are computed. Then a CSV file is created for each maturity stage and combined into a single file for training and testing models.

Creating a CSV File

CSV (Comma Separated Values) is a file format that is used for a large set of data storage. In Tomato Classification system, there is a huge data set that is generated by rotating, flipping or zooming methods. For that purpose, the information is organized within a CSV file. The combined CSV file (TomatoData.csv) contains 4859 records including all the five maturity levels of tomato (Green, Breaker, Pink, Light Red, and Red-Mature). All the attributes signifies the feature that are extracted from the images and the value of 'label' column indicates the maturity stage which the tomato belongs to. For example, if the label column has the value 1, then this indicates that the entire record belongs to stage 1. Similarly, for Breaker stage, label has value 2 and so on. Later on, machine learning models are trained based on these features included in CSV file.

Training and Classification

After extracting relevant features from the dataset, now there is a need to train a model. To compare the performance of different models, seven classifiers have been used for classification of different maturity stages of tomato. The process of training a model contains provision of machine learning algorithm with training dataset to learn from. The data which is to be trained must have correct answer, which is a target attribute. The learning algorithm compares the input data to this target attribute (the right answer to be predicted) and outputs the machine learning model that grabs these patterns in it. Then these models are used to make predictions on future data.

Once the training has been completed, next step is to understand, recognize and differentiate various stages of tomato based on the training set of data. This process is called classification and the algorithm which implements classification is known as classifier. Classification generally refers to classifying images into distinct categories of tomato maturity in Tomato Classification system.

Testing and Validation

In order to estimate the wellness of the predictive model that has been trained that is dependent upon the data size that the system is fed, the testing has been performed on the test set (20% of the dataset) which is a set of examples that is used to evaluate the performance of fully trained classifiers. The models are adjusted accordingly to minimize the error rate of the classifier. Among all the classifiers used in Tomato Classification system, the best one is chosen and applied to the real world problem to get the results.

Results and Discussion

The performance of the learning model depends on the data set that has been provided to train it. The present section discusses the output of all the performance metrics that are used to evaluate the accuracy of the classifier. All the classifiers have shown different accuracy in the form of confusion matrix.

Figure 4 shows the layout of the proposed application.

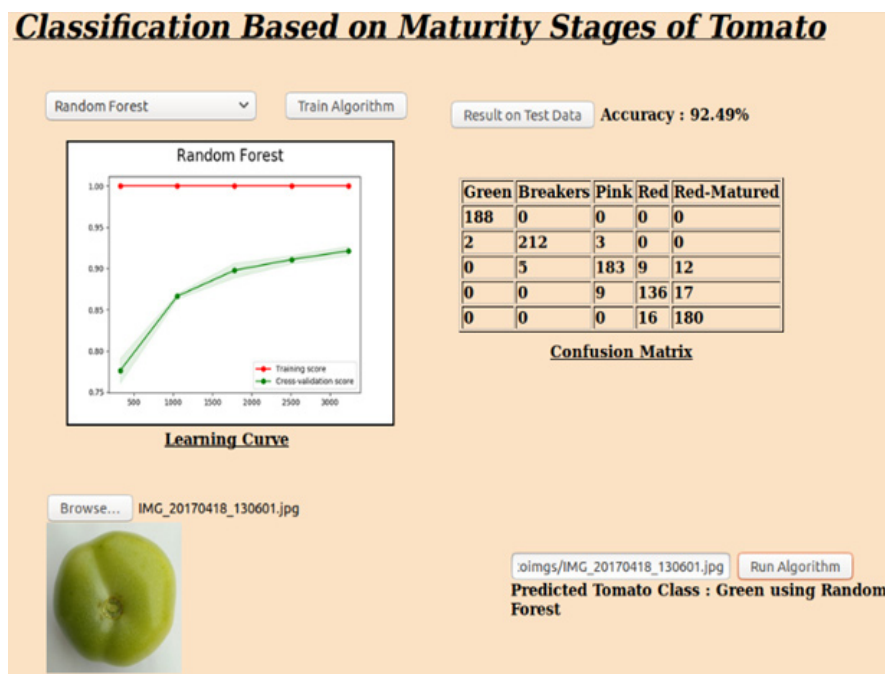


Fig. 4: Result Showing after Running the Algorithm

Comparison of Classifiers

On the basis of the learning process on training data set and accuracies on test data set, all the classifiers have shown different performance. Some of the classifiers fit best to the provided data of Tomato classification data set like Random Forest, Gradient Boosting, XgBoost and Decision Tree while rest of the classifiers found hard to fit the training data set like SVM, K-NN and Logistic Regression.

Table 1 has listed the accuracies of all the classifiers. The table shows that Random Forest has the highest accuracy among all the classifiers while SVM have shown worst accuracy on Tomato Classification data set. This is due to the reason that when there is large training data set, SVM is inefficient to train. For such

problems, Random Forest is the best model to train on large number of training samples.

Table 1: Accuracies of all the Classifiers

S.No.	Classifiers	Accuracy (In %)
1.	Decision Tree	87.55
2.	Gradient Boosting	91.98
3.	K-NN	81.07
4.	Logistic Regression	77.37
5.	Random Forest	92.49
6.	SVM	70.88
7.	XgBoost	90.33

Table 2 has listed the misclassification made by all the classifiers. The table shows that Random Forest has the least misclassification rate that is 0.0751 among all the classifiers while SVM have shown maximum of 0.2911 on Tomato Classification data set.

Figure 5 shows the bar graph comparing all the classifiers that have shown their performances in the form of accuracies.

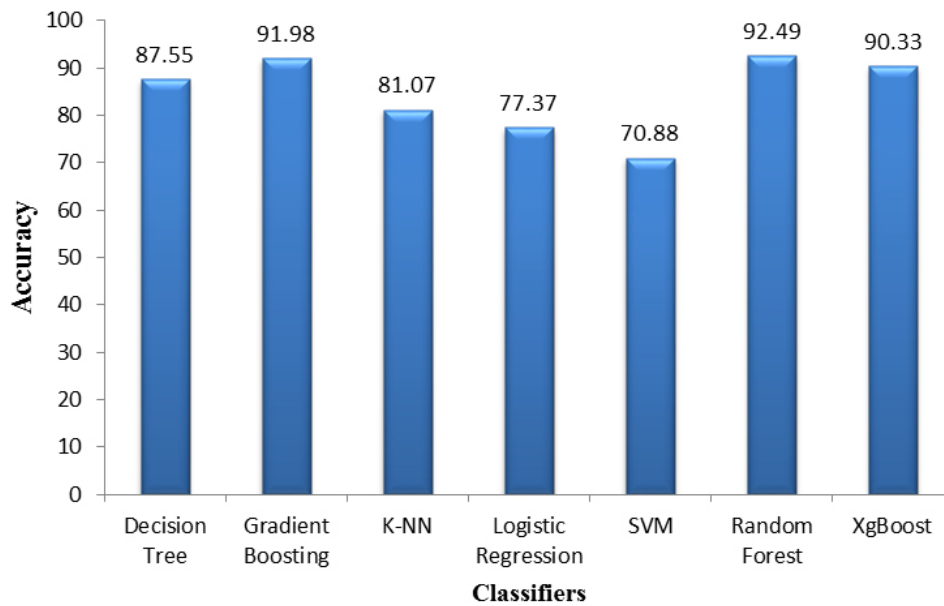


Fig. 5: Comparison of Different Classifiers Based on their Accuracies

Table 2: Misclassification by all the Classifiers

S.No.	Classifiers	Misclassification
1.	Decision Tree	0.1245
2.	Gradient Boosting	0.0802
3.	K-NN	0.1893
4.	Logistic Regression	0.2263
5.	Random Forest	0.0751
6.	SVM	0.2911
7.	XgBoost	0.0967

Conclusion

It is concluded from the results that the performance of the model depends on the following parameters:

- Increasing the size of data will increase the accuracy of the classifier.
- Selection of appropriate machine learning

algorithm will determine the success or failure.

- Features extracted from the dataset.
- Incorrect data collection and noise in the data will decrease the performance.
- Understanding and improving upon features selecting.
- Proper handling of training and testing data.

The Random Forest has shown the highest accuracy among all the classifiers while SVM have shown worst accuracy on Tomato Classification data set. This is due to the reason that when there is large training data set, SVM is inefficient to train. For such problems, Random Forest is the best model to train on large number of training samples. Predictive accuracies of Gradient Boosting and XgBoost are remarkably similar but Random Forest performed somewhat better than them.

References

1. Agrawal S, Jha S and Dewangan C (2016) Grading of tomatoes using digital image processing on the basis of color. *Int J Res Engg Tech* **5**: 138-40.
2. Arakeri M P and Lakshmana (2016) Computer vision based fruit grading system for quality evaluation of tomato in agriculture industry. *Procedia Comp Sci* **79**: 426-33.
3. Bendary N E, Hariri E E, Hassanien A E and Badr A (2014) Using machine learning techniques for evaluating tomato ripeness. *Exp Sys App* Pp 1-14, Egypt.
4. De Grano A. V. and Pabico J. P (2007) Neural network-based computer color vision for grading tomatoes. *Philippine J Crop Sci* **31**: 130-42.
5. Gunasekaran S (1996) Computer vision technology for food quality assurance. *Trends Food Sci Tech* **7**: 245–56.
6. Hashim N M, Mohamad N H, Zakaria Z, Bakri H and Sakaguchi F (2013) Development of tomato inspection and grading system using image processing. *Int J Engg Comp Sci* **2**: 2319-26.
7. Jaramillo, Rodriguez J, Guzman V, Zapata M and Rengifo T (2007) Good Agricultural Practices in the Production of tomato under protected conditions: Tech Manual. Food and Agriculture Organization, United Nations.
8. Kondo N, Ahmed U, Monta M and Murase H (2000) Machine vision based quality evaluation of lyokan orange fruit using neural networks. *Comp Elect Agric* **29**: 135-47.
9. Nakano K (1997) Application of neural networks to the color grading of apples. *Comp Elect Agric* **18**: 105-86.
10. Ohali Y A (2011) Computer vision based date fruit grading system: Design and implementation. *J Comp Inf Sci* **23**: 29-36.
11. Raut K and Brora V (2016) Assessment of fruit maturity using digital image processing. *Int J Sci Tech Engg* **3**: 273:79.
12. Rokunuzzaman M and Jayasuriya H P (2013) Development of a low cost machine vision system for sorting of tomatoes. *CIGR J Agric Engg* **15**: 173-80.
13. Rupanagudi S R, Ranjani B S, Nagaraj P, Bhat V G (2014) A cost effective tomato maturity grading system using image processing algorithm for farmers. *Int Conf Cont Comp Inf* Pp 7-12, Bangalore, India.
14. Sehgal P and Goel N (2016) Auto-annotation of tomato images based on ripeness and firmness classification for multimodal retrieval. *Int Conf Adv Comp Comm Inf* Pp 1084-91, Jaipur, India.