# MangoLDB: A Dataset of Mango Leaves RGB, Binary and Gray-Scale Image

**Hasna Hena, Ahmed Al Marouf, Rezwana Sultana**

***Abstract. This paper presents the very first image dataset of mango leaves of different species which are originated in Bangladesh. This dataset contains the RGB, binary and gray-scale; three versions of each images. Being the national tree of Bangladesh, mango is a sweet and juicy drupe having numerous species of trees. The scientific name for the commonly found mango species is Megnifera Indica. After the Jamdani Saree and Hilsha, different species of mango such as Khir sha, Langra, Aswina, Fazli, Haribhanga are the future geographical identification (GI) products of Bangladesh. Therefore, being highly demandable fruit, the identification of different species from its leaf images could be a challenging task. Agriculture specialists, farmers and general people may have difficulty to recognize samples just by observing the leaves. In this paper, we have formulated an image dataset of mango leaves of six different species namely, Aswina, Fazli, Gopalvog, Khirsha, Langra and Lokhna. After data retrieval, cleaning and processing, we have created an image dataset consisting of 7905 sample images. The images are collected using Smart phone having enough image pixel information for applying image processing tools. The dataset contains a minimal amount (1.5%) of blur and noise. The dataset could be considered as the basic ground-truth dataset for species recognition, disease recognition etc. from mango leaves in the area of computer vision.***

***Keywords: Computer Vision, Mango, Megnifera Indica, Mango Leaves Image Dataset, Species recognition.***

## I. INTRODUCTION

As declared on November 2010, mango is the national tree of Bangladesh, which shows the love and demand of this fruit in this country. Mango is a sweet and juicy drupe generally being produced in South-East Asian countries. Different species also being produced in South American countries such as Brazil. Being a very popular fruit in Bangladesh, only in Bangladesh it has more than 50 species. Each species has different features such as shape of the fruit, leaves size, leaves key features, tree size etc. Because of this diversity in mango leaves, it is difficult to recognize the species when there is no fruit in the tree. According to the rules of International Code of Botanical Nomenclature (ICBN), the scientific name of a plant consists of two names: the genus or generic name and the specific epithet or species name.

**Revised Manuscript Received on May 10, 2019**

**Most. Hasna Hena,** B.Sc and M.Sc degree with First Class Second position in the department of Information and Communication Engineering from the University of Rajshahi, Bangladesh.

**Ahmed Al Marouf,** currently pursuing M.Sc. Engg in Computer Science and Engineering (CSE) from Islamic University of Technology (IUT), Gazipur, Bangladesh.

**Rezwana Sultana** has completed her M.Sc. in CSE and B. Sc. in CSE from University of Dhaka.

Therefore, the genus for mango is "Magnifera" and species name changes for different species. For ease of use, in this paper, we are going to use the names used in local markets of Bangladesh such as Khirsha, Langra, Aswina, Fazli, Lokhna etc. Image processing, video processing and computer vision overlaps in the dimension of techniques in adopting a system and analyze the system. Images play the vital role in computer vision systems as the visual artifacts and effects of them on a vision-based system is significant. Finding visual descriptors and video frame descriptors are a challenging task in computer vision, therefore recognition or identification of a particular object from images got researchers' attention in recent years. It is evident in literature that image as stimuli have higher recognition rates than any other features. Therefore, to identify species or diseases of mango trees from the image of its leaves could be considered as a traditional image processing problem.

Image processing has been considered as one of the distinct and essential research fields of computer science from the very beginning of its introduction. Researchers' from many different area came together to solve various research problems in this field. Detection of tree species from its leaves is one of the image processing based problems that could be named as species recognition. For developing a species recognition system, the first thing required is a ground-truth image dataset. From the image datasets different types of image features such as shape-based features, size-based features, color-based features could be extracted. Various classifiers such as support vector machine (SVM), regression analysis could be applied on the extracted images to classify different species. To adopt such a recognition model, the first requirement is data collection and labeling. In this work, we have collected and formulated a RGB image dataset of mango leaves for recognizing different mango species. In this paper, we have presented the first ever mango leaves dataset containing RGB, binary and gray-scale images. Binary images contain the combination of 0's and 1's, which could lead to specific features. However, gray scale images are very useful for image enhancements and could provide enough features to detect diseases or particular species. The different gray levels could provide enough information regarding an image. The rest of the paper contains literature review in section II, dataset credentials in section III, data collection procedure in section IV, analysis of statistical information in section V and finally section VI concludes.

## II. Literature Review

This section illustrates the related works performed by the researchers' in the research area of image processing related to species recognition. Not only the researchers from computer science, but also from botany or microbiology involved in such research. Leaf dataset [1] is a ground-truth dataset presented by University of California Irvine (UCI) machine learning repository having 16 shape and texture features extracted from digital images of 40 different plant species. Apple leaf dataset [2] is publicly available dataset containing the genetic data and binary images which gives morphometrics about the complex and heritable apple leaf shapes.

**Table 1. Some existing plant image dataset**

| Name of the Dataset with citation | No. of Images | Available Features |
|---|---|---|
| Leaf dataset [1] | 340 | Class (Species), Specimen Number, Eccentricity, Aspect Ratio, Elongation, Solidity, Stochastic Convexity, Isoperimetric Factor, Maximal Indentation Depth, Lobedness, Average Intensity, Average Contrast, Smoothness, Third moment, Uniformity, Entropy |
| Apple Leaf dataset [2] | 9000 | Binary leaf image, morphological attributes. |
| Leaf Classification Data [3] | 1,584 | Binary black images against white backgrounds. Three set of features: a shape descriptor, texture histogram and fine-scale margin histogram. |
| Leafsnap Dataset [4] | 30,866 | RGB image, saturation-value extracted image and segmented image using expectation-maximization. |
| Model Dataset [5] | 150 | Region segmentation, Sobel Filter, Laplacian Filter |
| Sample Dataset [6] | 20 | Morphological Features (Leaf length, width, aspect ratio, area, perimeter, form factor) and Statistical Features (Entropy, constant, homogeneity, correlation) |

Leaf classification data [3] is published in 2016 as a Kaggle competition. The dataset contains 16 samples each of 99 species and then converted to binary images. The images are used for identifying 64 feature vectors which are kept for classification challenge. Leafsnap [4] database contains huge number of color images and the images are saturated & segmented using value extraction and expectation-maximization, respectively. N. Kumer et al. [4] proposed a computer vision based system for detecting plant species automatically. R. Nikam et al. [5] presented a methodology to determine the severity level of mango disease from leaf images. In their work they have used a model dataset of 150 color images and applied disease region segmentation, Sobel and Laplacian filters to detect the severity level. The paper analyzed the bacterial, fungal, viral and diseases due to insects.

From a small sample dataset having only 20 leaf images a method has been introduced by L. Dutta in [6]. Authors have utilized the different morphological & statistical features for feature extraction and Artificial Neural Network (ANN) for classification. Similar methods have been adopted by authors in [7-9]. Feature extractions and classification based supervised machine learning methods are common in literature to identify the mango tree species. Hardware instruments like Raspberry Pi have been used to capture real time image in [7] and k-means clustering has been used to detect the disease regions.

## III. Dataset Attributes

In this section, we have demonstrated the contents of the dataset. In this dataset, we have collected around 7900 images of 54 different trees. Some attributes of the dataset could be listed as following.

-Different mango species are kept in separate folders.

-Different trees of same species are kept in separate folders.

-Different leaves converted into three types of image are kept in the same folder.

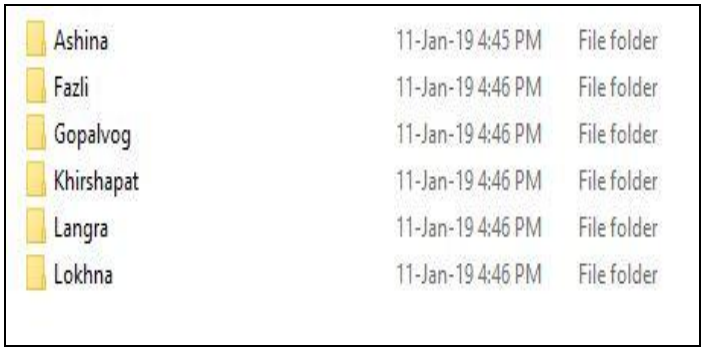



**Fig. 1. Mango Species Folder having tree-wise images inside.**

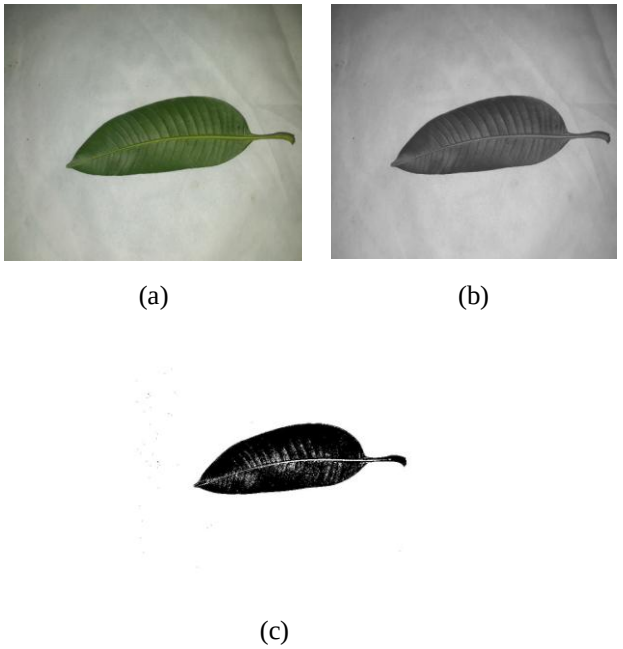

**Fig. 2. Tree-wise Langra Species folder having images inside.**


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The Fig. 1 and Fig. 2 illustrates the folders containing the images inside. Each of the species folder contains tree folders. Each tree folder contains three types of image inside. The Fig: .3. Shows the example of color, binary and gray-scale images which are included in our dataset. For each of the sample leaves, we have utilized MATLAB to convert it into a gray-scale and binary images.

The usefulness of gray-scale images can be found in [5-6] and binary images can be found in [6-8]. Again, for determining the statistical features in [6], the color image was converted to binary image. Edge detection mechanism was adopted to find features like area, perimeter etc.

For our dataset, we have organized the sample leaves into the given folder manner and named the files as "l1_t1_color.jpg", "l1_t1_grey.jpg", "l1_t1_bin.jpg" and so on.



(a) (b)

(c)

**Fig. 3. Sample leaves in (a) color, (b) gray-scale and (c) binary image.**

## IV. Data Collection Procedure

Mango trees in Bangladesh are blossomed from December to February. These grafted mango plants quickly start bearing blossoms after two or three years of planting. The owners of these trees try hard to nurture and support their young trees as well as the old for getting more mangoes. Mango is the leading seasonal cash crop of the northwestern region of Bangladesh and dominates the economy of these areas. Most of the people of these zones are employed for different jobs on the orchards such as nursing, harvesting and packing mangoes for transportation during those seasons in every year. Normally, the tree produces mango once in a year.

2635 field leaves from over 500 trees were collected from June, 2018 to July, 2018. About 10 to 30 different ages' leaves were collected from each tree and efforts were made to collect leaves from portions of the branch for giving variation. Seven volunteers had worked hard to collect the leaves from the branches and to keep them separately for taking images. Leaves were flattened and placed on white background, then images of both sides of each leaf were taken very carefully by smart phone, mostly used phone was Samsung. Some pictures of group of leaves were also taken for providing variation in the datasets.

## V. Statistical Analysis of Dataset

In this section, we have summarized the statistical properties of the proposed dataset. The quantitative properties of the dataset are in Table 2. 2635 RGB images were collected from about 60 trees and their binary as well as gray scale images are also available for the purpose of different types of research.

Table 3 demonstrates the species wise datasets. Six species: Aswina, Fazli, Gopalvog, Khirsha, Langra and Lakhna are available in our database now. In the near future, we are hoping to add more species and more samples. Species wise leaf datasets are displayed in Fig 1.

**Table 2.** Quantitative properties of MangoLDB

| Properties | Values |
|---|---|
| Total No of Images | 7905 |
| Total No of RGB Images | 2635 |
| Total No of Binary Images | 2635 |
| Total No of Gray-Scale Images | 2635 |
| No of Tree Samples | 54 |

The highest (678) and 2nd highest (559) number of samples are collected from Khirsha and Gopalvog, two most popular as well as delicious mangoes of Bangladesh. Different ages and different types of leaves were collected from eight to ten trees for each species. The average number of sample per species is 440, which is quite enough to run a supervised learning model for prediction or computer vision based works.

**Table 3. Species-wise Sample Statistics**

| Species | No of Sample Leaves |
|---|---|
| Aswina | 423 |
| Fazli | 339 |
| Gopalvog | 559 |
| Khirsha | 678 |
| Langra | 414 |
| Lokhna | 222 |
| Total | 2635 |


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

## VI. Applications of the Dataset and Future Scopes

In this section, we have analyzed the applications of the dataset and future scopes which may generate from our presented dataset. Species detection, disease detection, severity of disease level detection, computer vision and/or image processing based solutions could be focused application areas.

### *A. Species Detection*

Species detection is a trivial work which could be initiated with a very low number of image samples. The accuracy of these systems increases with the increment of number for training image samples. For these types of works our proposed MangoLDB could be used as a ground-truth dataset. As MangoLDB contains six types of species, using various feature descriptors, numerical features and multiclass classifiers species detection is very much possible.

### *B. Disease Detection*

In [5] R. Nikam et al. have presented a method to detect the severity of mango plant disease from the color images. In their work, they have converted the color images into gray-scale images as part of their pre-processing. As we have kept the gray-scale images in our dataset, it is very much possible to use our dataset instead of using random 150 samples. G. Kshirsagar et al. [8] proposed a plant disease detection method using raw image processing such as color segmentation in RGB and L*a*b* color space conversion. Similar disease recognition models could be depicted using our presented MangoLDB.

### *C. Disease Severity Measurement*

Disease severity detection is another highly demanding research area where the severity level of disease is detected so that the corresponding treatments could be offered to the plants. In [5, 6] different method has been proposed by authors regarding disease severity detection from leaf images. Similar experimental setup could be demonstrated using our proposed MangoLDB.

### *D. Computer Vision and Image Processing Models*

Computer vision and/or image processing based applications are very popular to utilize different types of image datasets for object detection, plant type detection or disease detection. V. B. Batule et al. [9] proposed image processing and support vector machine based technique to detect leaf disease. Similar work has been proposed by A. N. Rathod in [10] comparing different image processing techniques. Our dataset could be effectively used for these type of experiments.

## VII. Conclusion

In this paper, we have formulated and presented a leaf dataset particularly of Mango trees. According to the opportunities found in the literature, the proposed dataset could be utilized in many places effectively and in future we are going to implement some of the experiments using MangoLDB. To the best of our knowledge, the collection of 7905 image samples is the largest mango leaves dataset. As mango is one of the popular fruit among the South Asian peoples, the identification of particular species or diseases of mango trees would be challenging, yet worth research contribution. Apart from using for species and disease detection, this dataset could be used to model different machine learning systems and computer vision based systems. The research contribution of this paper is to set a benchmark dataset for the particular problem and achieve the future scopes at its best.

## REFERENCES


1. Pedro F.B. Silva, Andre R.S. Marcal, Rubim M. Almeida da Silva, “Evaluation of Features for Leaf Discrimination”, Springer Lecture Notes in Computer Science, Vol. 7950, pp. 197-204, 2013.
2. Zoë Migicovsky, Mao Li, Daniel H. Chitwood, Sean Myles, “Morphometrics Reveals Complex and Heritable Apple Leaf Shapes”, Frontiers in Plant Science, 2018.
3. Kaggle Leaf Classification Data [Online] https://www.kaggle.com/c/leaf-classification
4. Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida C. Lopez, João V. B. Soares, “Leafsnap: A Computer Vision System for Automatic Plant Species Identification,”, Proceedings of the 12th European Conference on Computer Vision (ECCV), October 2012.
5. R.Nikam, M. Sadavarte, “Application of Image Processing Technique in Mango Leaves Disease Severity Measurement”, National Conference on Emerging Trands in Computer, electrical and Electronics (ETCEE-2015), International Journal of Advance Engineering and Research Development (IJAERD), 2015.
6. K. Muthukannan, P. Latha, P. Nisha, R. Pon Selvi, “An Assessment on Detection of Plant Leaf Diseases and Its severity using image segmentation”, International Journal of Computer Science and Information Technology Research (IJCSITR), January-March 2015.
7. J. Sethupathy, Veni S., “OpenCV based Disease Identification of Mango Leaves”, International Journal of Engineering and Technology (IJET), Vol. 8 no 5, October-November 2016.
8. G. Kshirsagar, A. N. Thakre, “Plant Disease Detection in Image Processing using MATLAB”, International Journal on Recent and Innovation Trends in Computing and Comunication (IJRITCC), Vol. 6, issue. 4, April 2018.
9. V. B. Batule, G. U. Chavan, V. P. Sanap, K. D. Wadkar, “Leaf Disease Detection using Image Processing and Support vector machine (SVM)”, Journal for Research, Vol 2, issue 2, April 2016.
10. A. N. Rathod, B. Tanawal, V. Shah, “Image Processing Techniques for Detection of Leaf Disease”, International Journal of Advanced Research in computer Science and Software Engineering (IJARCSSE), Vol. 3, issue. 11, November, 2013.


## AUTHORS PROFILE

Most. Hasna Hena, She completed her B.Sc and M.Sc degree with First Class Second position in the department of Information and Communication Engineering from the University of Rajshahi, Bangladesh. She received gold medal Award and Merit Scholarship from Rajshahi University for her good result in the B.Sc (Honors) Examination. She also Received NSICT-2010 fellow award from Ministry of Science, Information and Communication Technology (MOSICT), Bangladesh, for her M. Sc. Research work. Her research area is Communication, Data Mining and Computer vision.



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

Ahmed Al Marouf is currently pursuing M.Sc. Engg in Computer Science and Engineering (CSE) from Islamic University of Technology (IUT), Gazipur, Bangladesh. He received his Bachelor degree from the Department of Computer Science and Engineering (CSE), IUT in 2014. He is a graduate researcher of Systems and Software Lab (SSL) in the CSE department of IUT. His research interest lies within Computer vision, Human-Computer Interaction, Biometric technologies. He is currently working as a lecturer in Department of Computer Science and Engineering (CSE) of Daffodil International University (DIU), Dhaka, Bangladesh. He also acts as the Technical Lead of DIU HCI (Human Computer Interaction) Research Lab.

Rezwana Sultana has completed her M.Sc. in CSE and B. Sc. in CSE from University of Dhaka. She has been working as a lecturer at Daffodil International University since 2015. She has high impact journals and conference papers. Her research area comprises Bioinformatics, Natural Language Processing and Computer Vision. She is currently acting as a core researcher in DIU HCI Research Lab and NLP Research Lab.