A Progress Report

on

**Quantification of infection patterns on tomato leaves due to various pathogens using image processing and machine learning techniques.**

*carried out as part of the course CS1634 Submitted by*

**Sarthak Sharma**

**189302091**

**&**

***Aahan Singh Charak***

***189301024***

***VI-CSE***

*in partial fulfillment for the award of the degree*

*of*

**BACHELOR OF TECHNOLOGY**

In

**Computer Science & Engineering**



**Department of Computer Science & Engineering,**

**School of Computing and IT,**

**Manipal University Jaipur,**

***April 2021***

Abstract

This project aims at quantification of infection patterns on tomato leaves due to various pathogens using image processing and machine learning techniques.Raters are prone to tiredness , which might eventually lead to inaccurate measurements.A lot of resources are required to constantly train the raters.They often require some reference to quantify disease severity e.g Standard Area Diagram. Raters can’t cover a large area and as such many plants are left with no inspection at all. Some plant diseases show symptoms after a long time. Till the time we are able to see the symptoms, most of the harm is done.So, the best alternative to visual estimation is using images to carry out the quantification. Images can be analysed in visual or non-visual spectrum. Visual spectrum is that part of the light spectrum in which humans can see objects. Other techniques include capturing images in the hyperspectral or multispectral forms. But it requires sophisticated sensor technology which is currently out of our scope and reach. So we will work with images captured using a camera in the visual range. By using image processing techniques we will segment out the diseased portion of the tomato-leaves and quantify the disease severity in them either using the nominal or percentage scale. We will mainly be using ROI segmentation and binary thresholding to achieve this. After thresholding quantification will be carried out, which involves calculating the fraction of the diseased pixels to the healthy pixels.This fraction can then be converted to a percentage value, which can be used to estimate the amount of disease present on the leaf. The real challenge in this research is the availability of accurate measurements of the diseased pixels in the image which we can refer to in order to validate our results. Ground truth validation is thus a challenge for us. Though our results might be correct, validation is still needed to confirm them.

Table of Contents

Pg. No

1. Introduction 1

2. Literature Review 4

3. Methodology and Framework 9

4. Work Done 15

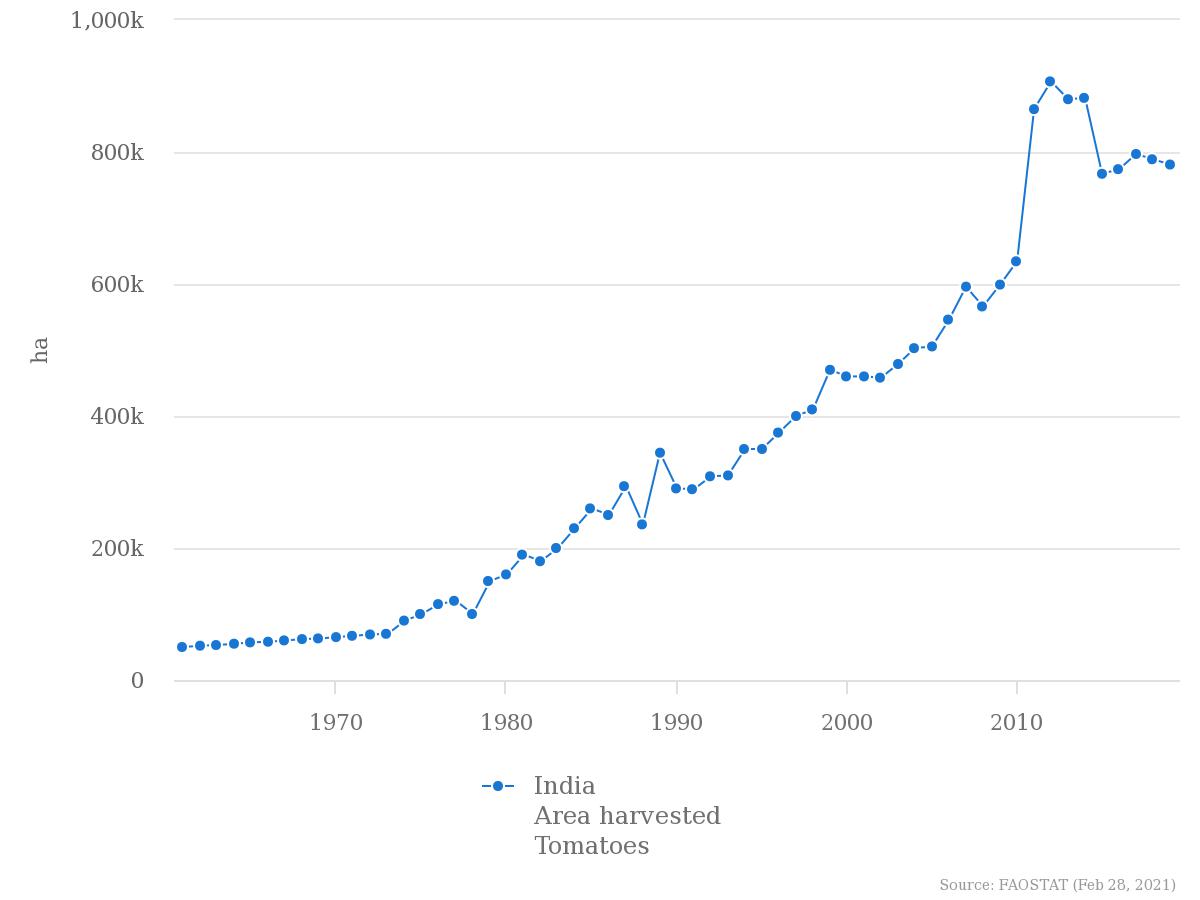
5. Conclusion and Future Plan 25

6. References 26

Introduction

Agriculture plays a very important role in our daily lives. From providing food to everyone, livelihood to the farmers and reducing the pollution levels , agriculture plays an important role in the survival of human beings.

Plants play a very vital role in nature. They are one of the forces keeping global-warming away from us. But plants aren’t resistant to diseases. So, it becomes necessary for us to study the diseases occuring in plants, in order to develop measures to prevent and cure them. Let’s take a look at the chart given below, showing the total area of tomato harvests vs the year of harvests in India.

***The graph can be constructed by visiting*** *:*  <http://www.fao.org/faostat/en/#compare>

From the above given trend we can clearly see that production takes a dip starting from around late 2014. One of the main factors behind this is diseases happening in plants, but if we are able to correctly identify them and calculate the severity of infection, we will surely be able to boost productivity. Being aware of the infection patterns and severity of infection well in advance can help to reduce the risk of disease spreading over a large area of crops. Diseases in plants are quite ubiquitous, so identifying them well in advance will be a great help to us.

1

Visual estimation is one of the ways to achieve this but requires a lot of hard work and human resources which isn’t cost efficient. Visual analysis of images for quantification can help estimate the severity of infection over a large area without much effort and is also cheap. Moreover, the market is filled with many softwares which make our jobs easier as we don’t have to develop our own algorithms to achieve this task and they also help us save a lot of precious time.

1.1 Motivation:

In most of the cases estimation about the severity of plant diseases is done by humans visually. These individuals who excel in the field of visual estimation of plant disease severity are called Raters. If they are trained effectively, they can make accurate estimations

about the plant disease severity. But there can be a lot of disadvantages in quantifying the plant diseases visually. According to Bock et al.(2013) there are many disadvantages of visual estimation. Some of them are discussed below :

Raters are prone to tiredness , which might eventually lead to inaccurate measurements.

A lot of resources are required to constantly train the raters.

They often require some reference to quantify disease severity e.g Standard Area Diagram.

Raters can’t cover a large area and as such many plants are left with no inspection at all.

Some plant diseases show symptoms after a long time. Till the time we are able to see the symptoms, most of the harm is done.

So, the best alternative to visual estimation is using images to carry out the quantification. Images can be analysed in visual or non-visual spectrum. Visual spectrum is that part of the visual spectrum in which humans can see objects. Other techniques include capturing images in the hyperspectral or multispectral forms. But it requires sophisticated sensor technology which is currently out of our scope and reach. So we will work with images captured using a camera in the visual range.

2

By using image processing techniques we will segment out the diseased portion of the tomato-leaves and quantify the disease severity in them either using the nominal or percentage scale. Example of a tomato leaf image in the visible spectrum is given below.



Image showing bacterial spot on an infected tomato leaf

Image can be found at [Plant Village image dataset at github.](https://github.com/spMohanty/PlantVillage-Dataset/blob/master/raw/color/Tomato___Bacterial_spot/00728f4d-83a0-49f1-87f8-374646fcda05___GCREC_Bact.Sp%206326.JPG)

3

Literature Review

**Quantification:**

Quantification is the process of finding out how much portion of a plant is diseased. It is also called the severity of disease in a plant. It basically tells us about the percentage of the diseased tissue in a plant.

In this project we are measuring the severity of infection on tomato leaves. There are many techniques to measure this but we are using visual spectrum image analysis technique. Image will be in the form of pixels (rows and columns) and each pixel will contain a specific portion of the leaf. The diseased pixels will be found out using image analysis. We will call them bad pixels. We will then calculate the percentage of bad pixels on the leaf segment. This percentage scale can now be converted to nominal in order to find the severity of infection on the leaf.

According to Barbedo (2013), the severity of infection on a leaf can be easily identified by the color pattern on the leaf.Most of the quantification algorithms require segmentation step.Every quantification analysis that we do by ourselves is not 100 percent accurate. All measurements diverge from the "true value" or "ground truth". The measurements that we make are not absolute and are bound to differ from the "ground truth". It's just an estimate that we make.

**Segmentation:**

Segmentation is one of the most important steps for disease quantification. Segmentation is an image processing concept which involves dividing a particular image into regions of interests. We can then study those regions of interest without worrying about unwanted pixels interfering with our measurements.

Thresholding is one of the most commonly used techniques used to segment the diseased portion of a leaf from the healthy portion. According to Barbedo (2013), the most simple thresholding requires separating the diseased portion of the leaf and then applying a correction factor. The correct- ion factor makes sure that the healthy pixels which were counted as diseased by mistake are counted out of the final value. Colored CCD(charged couple device) showed better results than black CCD in coffee leaves.

4

Quantification via image processing (using segmentation ) is found to be a better and efficient choice as compared to visual estimation.In two-stage thresholding we first segment the image from the back- ground. And in the second stage we segment diseased pixels from healthy. Then we calculate the portion of bad pixels divided by the whole leaf area pixels.

**Visible Spectrum Image Analysis:**

There exist many color spaces which can be used to analyze images. HSV,RGB ,La\*b\* are some of the color spaces which can be used to study and quantify leaf diseases in plants.

According to Barbedo(2013), visible spectrum image analysis has its own demerits. Some diseases are not visible to the naked eye, also some diseases begin to show symptoms at a much later stage, which is quite late to carry out analysis. Here we will deal with quantification of diseases using HSV and La\*b\* color spaces.

The good thing about visible spectrum image analysis is that it produces efficient results, if carried out under controlled conditions.

**Datasets:**

In order to carry out research properly datasets are a must.An algorithm can be made better and more useful by testing and evaluating results on a well defined collection of data ( Data Set ) that is compatible with our case study.Results after using the datasets will tell about how efficient our algorithm really is. Plant village data set is used which consists of 54303 healthy and unhealthy leaf images divided into 38 categories by species and disease.

**Nominal Scale and Percentage Scale:**

During experiments/research, results are recorded using predefined statistical scales. Nominal scale is one of those scales. Nominal data consists of non-numeric information. Nominal scale is used to record such kinds of data. Data like gender,hair-color etc.. are usually measured using the nominal scale.With the help of nominal scale we can not only classify different objects , we can also allot different numbers based on characteristics.For example we can allot numbers 1- 5 for disease severity , 1 indicating lowest concern and 5 indicating very severe disease .

Percentage scale on the other hand is simply the representation of diseased pixels in the form of percentage values ranging between 0 and 100.

5

**Standard Area Diagrams:**

Standard Area Diagram generally known as SAD is an important tool which can be used to visually estimate the severity of disease on a plant/leaf.

According to Bock et al. (2020) SAD’S are extensively tools used by raters to correctly quantify disease severity in plants.

The most typical SAD comprises five to eight black-and-white drawings of leaves, with severity increasing in a non-linear fashion. In a two-step SAD validation approach , linear regression is the preferred method.

**Color Spaces:**

RGB is mostly used for image capturing but is rarely used to segment images. HSV is used for image thresholding segmentation. In this we use a histogram of intensities which can be used to segment diseased portions from the leaf. RGB is not preferred as it just gives us the idea how human beings perceive the captured object, HSV on the other hand gives us an idea about the true color of the captured object, which at times is also called as pure color.

**Outcomes**

1. HSV and La\*b\* color spaces are the most useful when it comes to studying the extensive features of the plants.
2. Nominal or percentage scale can be used to represent severity of disease in numerical or alphabetical form.
3. Though standard area diagrams are effective to visually rate the severity of disease in the plants they are pretty hard to come by.
4. Raters are most efficient in visually estimating severity of plant disease but are costly and require a large amount of capital to train and keep up to date.
5. Thresholding is one of the most commonly used segmentation techniques which can be used to efficiently segment the diseased portion of a leaf/plant from the healthy portion.
6. One of the simplest quantification techniques is to find the fraction of diseased pixels to the healthy pixels and then accordingly convert the value to nominal or percentage scales.

6

1. There are some cases in which visual spectrum analysis can’t be used like in the case of hidden disease patterns. In such cases we use hyperspectral or multispectral image analysis techniques which are currently out of our scope.
2. Raters are prone to errors like lighting,illusions etc.. For such cases visible image spectrum techniques provide an alternative.
3. Shadows and darkness can cause errors while performing visual image spectrum image analysis.
4. Ground truth validation is also required to validate our results. Such sources of validation are very hard to come by and are mostly done by raters using standard area diagrams.

**Problem Statement**

**Objective**: Quantification of infection patterns on tomato leaves due to various pathogens using image processing and machine learning techniques.

**Description:**

Plant diseases affect agricultural related activities in all economies. It is necessary to identify plant disease at their earliest in order to avoid enormous amounts of damage. Visually estimating the severity of disease in leaves is easy but requires a lot of capital. Disease quantification using image processing is quite effective in getting the job done. Image processing techniques like segmentation,pixel quantification and color space analysis are productive tools which help us in estimating disease severity in plants. In this project we will mainly use these techniques on tomato leaves to quantify diseases on a vast dataset of images. Many tomato diseases like early-blight,late-blight,mites etc.. will be covered.We will use image processing using python and OpenCv to get this task done. This task can be achieved by calculating the bad pixels to good pixels ratio in the images. Bad pixel basically refers to the diseased portion of the leaf image and good pixels ratio refers to the healthy portion of the leaf.

7

**Research Objectives**

* To analyze the infection pattern on tomato leaves using various color spaces like HSV, La\*b\* etc..
* To segment diseased portion of the leaves from the healthy portion using image processing techniques.
* To quantify the disease portions using appropriate algorithms.
* Validating our results using ground truth validation.

8

Methodology and Framework

**DATASET ACQUISITION:**

Dataset acquisition is the first and foremost step in order to carry out research work. Plant village dataset from github was used to collect images for various kinds of tomato leaf diseases. Many different kinds of tomato leaf diseases like Mosaic virus, early blight, late blight were considered for our research. 4 different sample images for each disease were considered as of now.

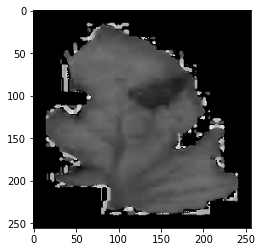
**Color Space Analysis:**

As our next step, we analysed the image samples using two different color space models, HSV and La\*b\* .

* **HSV Color Space Analysis:**

HSV stands for hue,saturation and value. Hue is the angle, the color makes with the color wheel (red is 0 degrees). Saturation basically tells the amount of greyness in a color. As we keep on mixing white light with the color, the greyness level keeps on increasing and eventually the color turns to white.It’s value ranges from 0 to 100 but since we used opencv to carry out our research work, the saturation value had to be converted to a range between 0 and 255. Value is the amount of brightness in an image. 100 value means 100 percent brightness, 0 value means 0 brightness or completely black color. Again for opencv the value had to be converted into a range of 0 to 255.

Next, we visualized the images in h,s and v channels separately, using grayscale. H channel gave us the best visualization of the diseased portion so it was considered to segment the image.



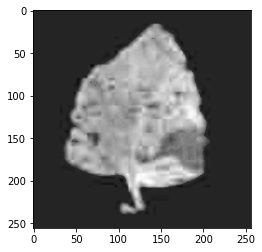


Tomato Early Blight Image Visualized in H,S and V channels respectively.

**9**

* **La\*b\* Color Space Analysis:**

For La\*b\* color space, we chose the L channel for segmentation.



Tomato Early Blight Image Visualized in L,a and b channels respectively.

**Contrast Enhancement:**

For contrast enhancement we used the algorithm as described by Barbedo (2016) . The algorithm is stated as:

**Resxy= ( Pxy- minim( P ) ) / ( maxim( P ) - minim( P ) ) \* 255**

Res=Resultant Pixel Value

Pxy= Current Pixel on which operation is being performed

maxim(P)= Maximum Pixel Intensity

minim(P)=Minimum Pixel Intensity

x and y are the indices of the current pixel and 255 is a factor multiplied to convert image’s scale to (0-255)

**ROI Segmentation:**

For ROI Segmentation we used H channel and A channel for HSV and La\*b\* respectively. We used the same algorithm as described by Barbedo (2016).

First we constructed an intensity histogram for both the channels. 100 bins were taken from 1 to 100 .We restricted the range upto 100 because after that most leaves didn’t have any significant amount of pixel intensities. Histogram construction was important because the healthier plant tissues tend to generate a peak at the right side of the histogram. The diseased tissues tend to generate peaks towards the left side of the histogram.

10

In grayscale images healthier plant tissues are represented as light shades of gray whilst on the other hand diseased tissues tend to be on the blacker side. First, peak intensity (global maxima) was identified as the bin corresponding to it was noted down as lets say **BIN.**

The intensity value of the maxima was noted down as **maxInt.**

**if ( BIN <= 40 )**

**REF = BIN value where Intensity > 0.2 \* maxInt and Intensity ! = maxInt**

**else**

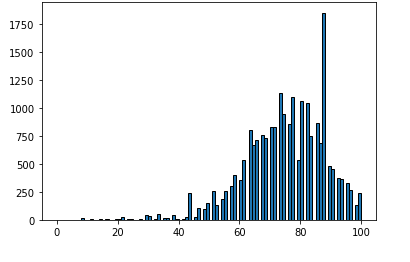
**REF = BIN value where Intensity > 0.5 \* maxInt and Intensity ! = maxInt**

**if no bin found for reference then REF = BIN**

If a peak is found at lower ranges, then that means that the plant is severely diseased. Also,if the peak is found at higher ranges, it means that the leaf is healthy as green and healthy pixels have the highest number in the image.

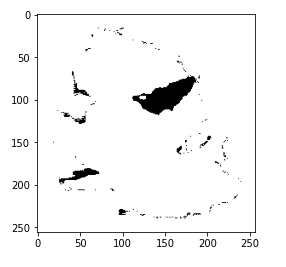
The best bin which separates diseased and healthy pixels can be given by

**TBIN = 2\* REF / 3**



An Intensity Histogram Example

**11**

****

Segmented Image of the leaf Example

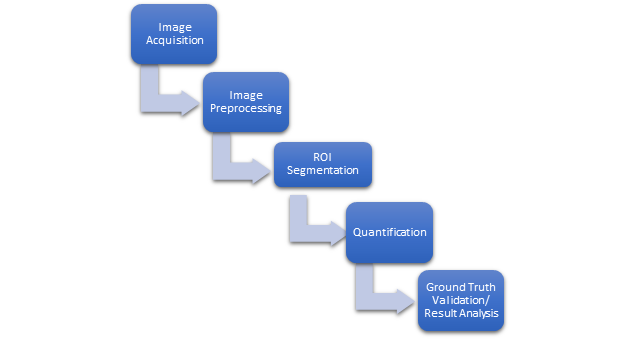
**Quantification :**

The quantification process was fairly easy. All we had to do was to calculate the total pixel area occupied by the leaf and the total pixel area occupied by the diseased pixels. Next the diseased pixels were divided by the healthy pixels to generate a percentage value. This process was repeated for different kinds of tomato leaf diseases and these values were noted down.

**Severity of Disease ( % ) = Nbad  / Ntotal \*100**

Nbad  = Number of diseased pixels

Ntotal = Total number of pixels occupied by the leaf



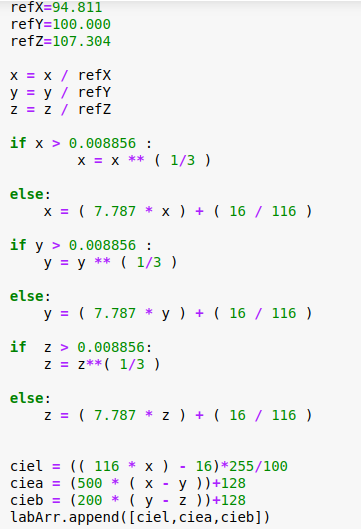
Flow Diagram For Methodology

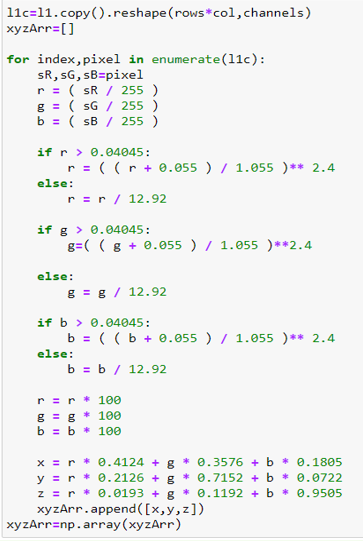
12

**Work Done**



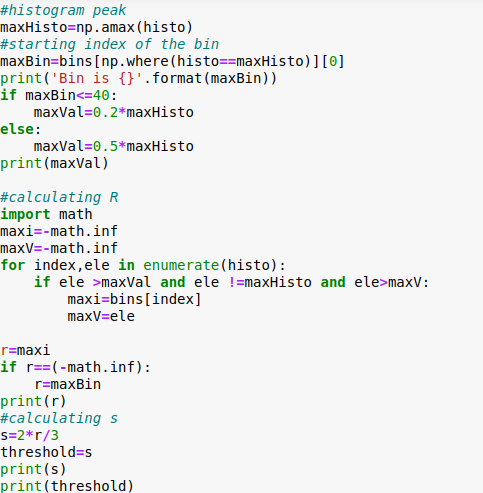
Algorithm to convert RGB to HSV



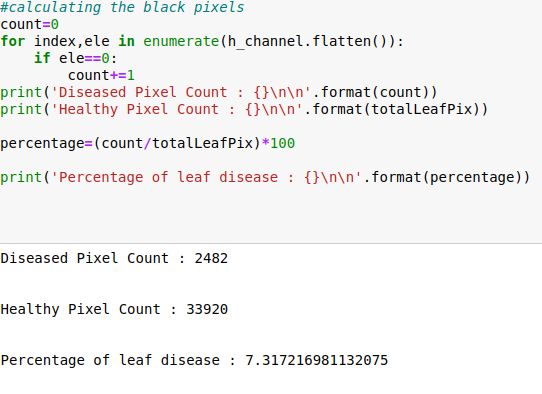


Algorithm to convert RGB TO La\*b\* color space

13



Thresholding Algorithm



Quantification Algorithm

14

**Results and Discussion :**

**Tomato Early Blight:**



Sample Early Blight-1

Diseased Pixel Count : 2482

Healthy Pixel Count : 33920

Percentage of leaf disease : 7.31



Sample Early Blight-2

Diseased Pixel Count : 516

Healthy Pixel Count : 29217

Percentage of leaf disease : 1.766



Sample Early Blight-3

Diseased Pixel Count : 8922

Healthy Pixel Count : 41447

Percentage of leaf disease : 21.526

15



Sample Early Blight-4

Diseased Pixel Count : 1660

Healthy Pixel Count : 28152

Percentage of leaf disease : 5.896

**Tomato Late Blight**



Sample Late Blight -1

Diseased Pixel Count : 748

Healthy Pixel Count : 33332

Percentage of leaf disease : 2.244



Sample Late Blight -2

Diseased Pixel Count : 8068

Healthy Pixel Count : 38418

Percentage of leaf disease : 21.005

16



Sample Late Blight -3

Diseased Pixel Count : 1362

Healthy Pixel Count : 26516

Percentage of leaf disease : 5.136



Sample Late Blight - 4

Diseased Pixel Count : 4851

Healthy Pixel Count : 35765

Percentage of leaf disease : 13.563

**Tomato Leaf Spot:**



Sample Leaf Spot -1

Diseased Pixel Count : 1901

Healthy Pixel Count : 27893

Percentage of leaf disease : 6.815

17



Sample Leaf Spot -2

Diseased Pixel Count : 2207

Healthy Pixel Count : 44716

Percentage of leaf disease : 4.935



Sample Leaf Spot - 3

Diseased Pixel Count : 5983

Healthy Pixel Count : 35226

Percentage of leaf disease : 16.984



Sample Leaf Spot - 4

Diseased Pixel Count : 1229

Healthy Pixel Count : 24725

Percentage of leaf disease : 4.970

18

**Tomato Bacterial Spot:**

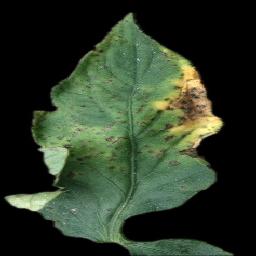


Sample Tomato Bacterial Spot -1

Diseased Pixel Count : 7155

Healthy Pixel Count : 29255

Percentage of leaf disease : 24.457



Sample Tomato Bacterial Spot - 2

Diseased Pixel Count : 6283

Healthy Pixel Count : 35474

Percentage of leaf disease : 17.7115



Sample Tomato Bacterial Spot - 3

Diseased Pixel Count : 3788

Healthy Pixel Count : 39803

Percentage of leaf disease : 9.5168

19



Sample Tomato Bacterial Spot - 4

Diseased Pixel Count : 4233

Healthy Pixel Count : 28325

Percentage of leaf disease : 14.944

**Spider Mites:**



Sample Spider Mites - 1

Diseased Pixel Count : 285

Healthy Pixel Count : 25380

Percentage of leaf disease : 1.122



Sample Spider Mites-2

Diseased Pixel Count : 619

Healthy Pixel Count : 24311

Percentage of leaf disease : 2.546

20



Sample Spider Mites - 3

Diseased Pixel Count : 521

Healthy Pixel Count : 28675

Percentage of leaf disease : 1.816



Sample Spider Mites - 4

Diseased Pixel Count : 260

Healthy Pixel Count : 26769

Percentage of leaf disease : 0.9712

**Mosaic Virus:**



Sample Mosaic Virus -1

Diseased Pixel Count : 215

Healthy Pixel Count : 23260

Percentage of leaf disease : 0.924

21



Sample Mosaic Virus - 2

Diseased Pixel Count : 275

Healthy Pixel Count : 21004

Percentage of leaf disease : 1.309



Sample Mosaic Virus - 3

Diseased Pixel Count : 182

Healthy Pixel Count : 19652

Percentage of leaf disease : 0.926



Sample Mosaic Virus - 4

Diseased Pixel Count : 137

Healthy Pixel Count : 18495

Percentage of leaf disease : 0.740

22

**Yellow Leaf Curl Virus:**



Sample YLCV -1

Diseased Pixel Count : 367

Healthy Pixel Count : 32190

Percentage of leaf disease : 1.140



Sample YLCV -2

Diseased Pixel Count : 166

Healthy Pixel Count : 38310

Percentage of leaf disease : 0.433



Sample YLCV - 3

Diseased Pixel Count : 271

Healthy Pixel Count : 40720

Percentage of leaf disease : 0.665

23

There’s no way to currently determine if our results are correct or not. We are still searching for some ground truth values in order to confirm that our algorithm is in fact correct. Moreover image erosion still isn’t applied so boundary pixel values might cause some errors in measurements.

**Contribution of individual members:**

**Aahan Singh:**

Aahan was responsible for quantifying diseases in HSV color space .

**Sarthak Sharma:**

Sarthak was responsible for quantifying diseases in the La\*b\* color space.

All algorithms and inferences were made with contribution from each member.

24

Conclusion and Future Plans

From the research we came to a conclusion that early-blight, late-blight, and bacterial spots tend to cause more disease severity in tomato leaves as compared to other diseases. This is not an absolute truth. Moreover there is still a need to add ground truth validation for the algorithm, so that we can be sure that it is in fact correct and only then we can come to a concrete conclusion. Moreover pixel erosion is still needed in order to remove small errors which might have happened due to them being present in the segmented image.

25

References

[[1](https://springerplus.springeropen.com/articles/10.1186/2193-1801-2-660)] Barbedo, J. G. A. (2013). Digital image processing techniques for detecting, quantifying and classifying plant diseases. *SpringerPlus*, *2*(1), 1-12.

[[2](https://link.springer.com/article/10.1007/s40858-016-0090-8)] Barbedo, J. G. A. (2016). A novel algorithm for semi-automatic segmentation of plant leaf disease symptoms using digital image processing. *Tropical Plant Pathology*, *41*(4), 210-224.

[[3](https://phytopatholres.biomedcentral.com/articles/10.1186/s42483-020-00049-8)] Bock, C. H., Barbedo, J. G., Del Ponte, E. M., Bohnenkamp, D., & Mahlein, A. K. (2020). From visual estimates to fully automated sensor-based measurements of plant disease severity: status and challenges for improving accuracy. *Phytopathology Research*, *2*, 1-30.

[[4](http://www.fao.org/faostat/en/#data)] FAOSTAT. (2021). Retrieved 1 March 2021, from <http://www.fao.org/faostat/en/#data>

[5] Gonzalez, R. C., & Woods, R. E. (2018). *Digital image processing* (4th ed.). Uttar Pradesh: Pearson India.

26