Tomato Leaf Disease Classification Proposal

Stephanie Cabanela, Akaash Venkat, Nicolas Loffreda

University of California, Berkeley
281 Computer Vision, Summer 2023

Abstract

According to a 2022 report by the United Nations, the world population is projected to increase from 8 billion today to 9.7 billion by 2050. Factors like global population growth and climate change contribute to the increasingly urgent issue of food security. Currently, infectious diseases reduce the potential yield of crops by an average of 40% and go as high as 100% for many crop growers in the developing world. The widespread distribution of smartphones among crop growers globally offers the opportunity to combat and prevent plant diseases by developing mobile disease diagnostic tools to algorithmically identify crop diseases through machine learning. We propose a project to use computer vision techniques to classify the leaves of tomato crops within the PlantVillage dataset into 10 categories (9 disease types/categories and 1 healthy category).

Dataset

Intended Classification Problem

The task is to categorize each tomato leaf into one of ten categories:

1. Bacterial Spot
2. Early Blight
3. Healthy
4. Late Blight
5. Leaf Mold
6. Septoria Leaf Spot
7. Spider Mites Two-spotted Spider Mite
8. Target Spot
9. Tomato Mosaic Virus
10. Tomato Yellow Leaf Curl Virus

Dataset Source

Link: https://www.kaggle.com/datasets/abdallahalidev/plantvillage-dataset?resource=download

The dataset which we'll use is the PlantVillage Dataset, a dataset from the online platform PlantVillage, which contains over 50,000 images on healthy and infected leaves of planted crops. This dataset can be downloaded from Kaggle via the link above. From this dataset, we have narrowed our focus into the healthy and infected leaves of tomato plants.

Dataset Description

The dataset provides images in three forms: grayscale, segmented, and color. Below in Figure 1, we can see an image of each category of healthy and infected tomato leaves, in the three forms provided by the dataset.

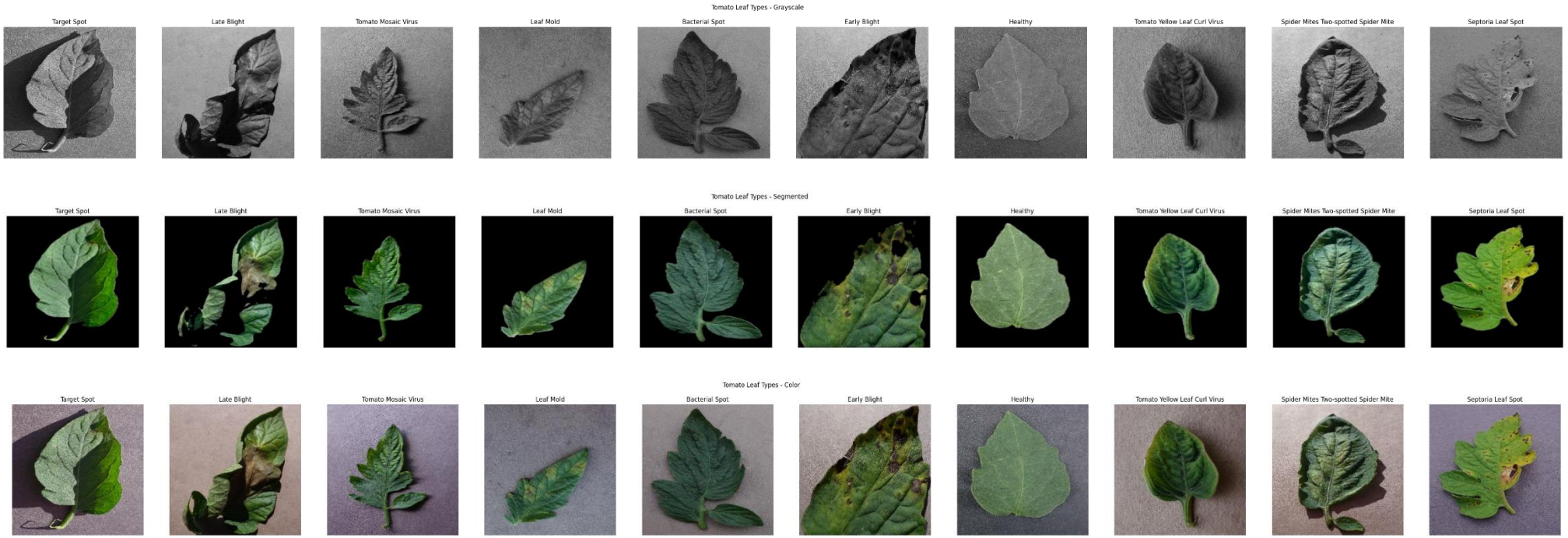

Figure 1: Example images of tomato plant leaves for each of the 10 disease categories. The top row contains sample images from the grayscale data subset, the middle row shows sample images from the segmented subset, and the bottom row contains images from the raw/color subset.

Some preprocessing has already been done for the images in the segmented subset: the background has already been removed and upon visual inspection, there seems to have been some light balancing applied since the leaves appear to be brighter and seem to have more vibrant color tones.

We plan to primarily use the segmented subset for our classification task given its cleaner images. However, given the knowledge we still have yet to gain from the class, there may be useful information in the other two subsets we aren't aware of yet (e.g. we hypothesize that the background in the color subset might give useful information about the lighting as we apply tone mapping). Therefore, we plan to use the segmented subset for now until we are better able to assess whether the color subset is more appropriate.

Description of Variation in the Dataset

In the segmented folder for the healthy and infected tomato leaves, there were a total of 18,160 images spread across the ten different categories. Of those images, all were of **shape (256, 256, 3)**, except for one image. To ensure that all images were of the same shape, we simply removed that one image that had a different shape, leaving us with **18,159 images**.

We then performed stratified sampling and split these images into splits of 80% train, 10% validation, and 10% test. This resulted in a **train set of 14,523 images**, a **validation set of 1,812 images**, and a **test set of 1,824 images**.

Tomato Leaf Category	***Train Dataset***	***Validation Dataset***	***Test Dataset***
Bacterial Spot	1701 images	212 images	214 images
Early Blight	800 images	100 images	100 images
Healthy	1272 images	159 images	160 images
Late Blight	1527 images	190 images	192 images
Leaf Mold	761 images	95 images	96 images
Septoria Leaf Spot	1416 images	177 images	178 images
Spider Mites Two-spotted Spider Mite	1340 images	167 images	168 images
Target Spot	1123 images	140 images	141 images
Tomato Mosaic Virus	298 images	37 images	38 images
Tomato Yellow Leaf Curl Virus	4285 images	535 images	537 images

Figure 2: Table of dataset distribution across 10 categories for train, validation, and test.

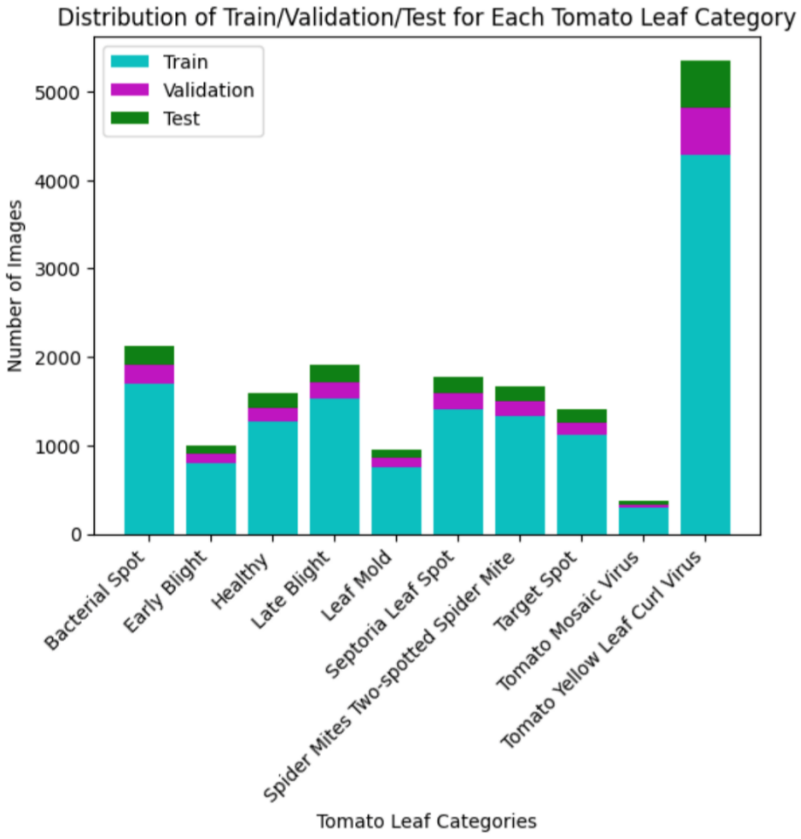

Figure 3: Stacked bar chart of dataset distribution.

Methodology

First, we believe that some preprocessing of the data will be necessary. In particular, we recognize that each picture is exposed to varying light conditions and that some sort of tone mapping, such as histogram equalization, may be needed to get a more homogeneous and truthful color palette of the leaves. We also see that the shadows and the color of the background may have some influence on the features. If we end up using the color subset instead of the segmented subset, we will attempt to remove the background so that only the leaf remains in each image.

As for the features, we believe that the color of the leaf may be a strong signal to be able to categorize each disease. Different shades of green, brown, and yellow can help identify which category the image should fall into. For this, we will be using a color histogram as one of our features.

At the same time, we recognize that only the color may not be sufficient. Each disease leaves different patterns of different colors on the leaf. For this, we will also try to identify blobs where the different colors are located to differentiate the shape and quantity of the different colors on the leaf (i.e., brown spots, big yellow circles, etc.).

We also see that some diseases deform the leaf in multiple ways, so we believe that using edge detection to get the contour of the leaf can also help to identify the different categories. A common algorithm for edge detection is Canny, which we will use as our first experiment. Also, a Histogram of Gradients (HOG) could be used for this purpose, but given that the plants have different orientations on the pictures, this may not work as well unless some pre-processing is applied to the images.

Some examples of the different colors and shapes generated by different diseases below:

Figure 4: Example tomato leaves of varying color and shape.