Proposal for Cassava Plant Disease Classification

Section 1

Group 2:
Clifton Harris, Jacob Jones, Rodigo Cochran, Tamar Brand-Perez

Dataset Link:
https://www.kaggle.com/datasets/mexwell/crop-diseases-classification?resource=download-directory

1. Dataset Information

The dataset used for this project is an image dataset of cassava plants, categorized into five classes based on different diseases. It consists of images stored in a folder named train_images and accompanying labels in a CSV file (train.csv) that maps the image id to their corresponding labels. The dataset also includes a JSON file (label_num_to_disease_map.json) that provides the name of the disease for each label. The dataset link can be found here.

First cultivated in the Yucatán by the Mayan people, cassava is cultivated worldwide to produce flour, alcohol, and tapioca[1]. Despite not being frost tolerant, the plant is highly drought resistant so it can be cultivated in various climates[2]. Cassava is vulnerable to viruses and bacteria that can attack the plant's roots, leaves, or tubers[2]. Keeping a healthy cassava crop is important for farmers as many common diseases spread from plant to plant and can be controlled if identified early[3].

2. Example Images for Each Class

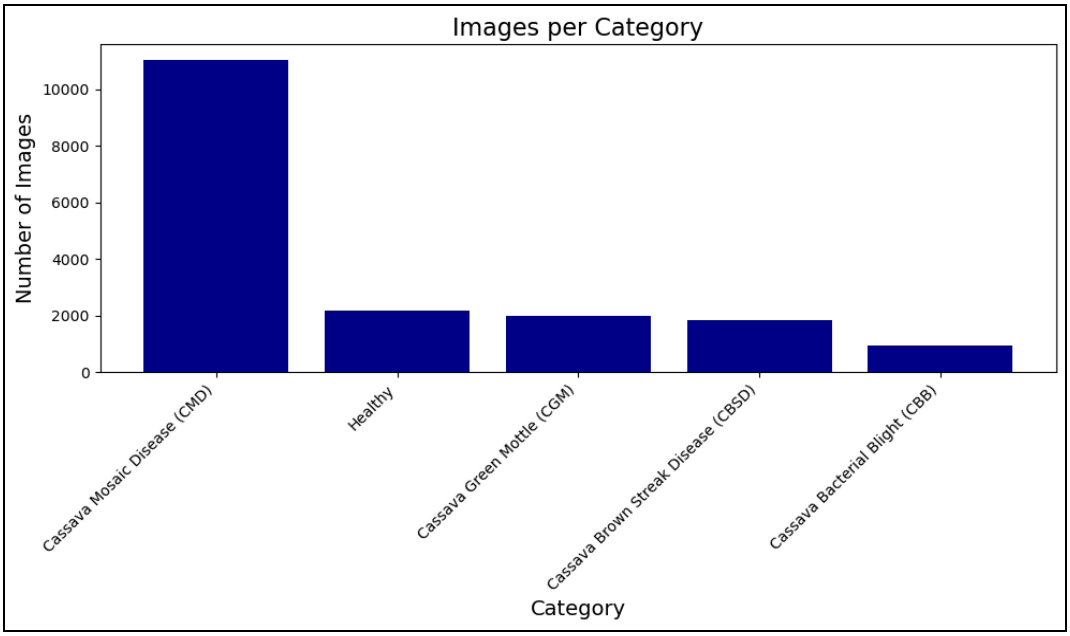
To illustrate the dataset, here are representative images for each of the five classes:

- Cassava Bacterial Blight (CBB) (label 0)
- Cassava Brown Streak Disease (CBSD) (label 1)
- Cassava Green Mottle (CGM) (label 2)
- Cassava Mosaic Disease (CMD) (label 3)
- Healthy (Label 4)

3. Description of Variation in the Dataset

- Categories: The dataset includes five distinct classes:
    - Cassava Mosaic Disease (CMD): 11,027 images (label 3)
    - Healthy: 2,166 images (label 4)
    - Cassava Green Mottle (CGM): 1,993 images (label 2)
    - Cassava Brown Streak Disease (CBSD): 1,831 images (label 1)
    - Cassava Bacterial Blight (CBB): 921 images (label 0)

[1] https://www.britannica.com/plant/cassava
[2] https://plants.usda.gov/DocumentLibrary/plantguide/pdf/pg_maes.pdf
[3] https://www.iita.org/wp-content/uploads/2016/06/Disease_control_in_cassava_farms_IPM_field_guide_for_extension_agents.pdf



This indicates that the dataset is skewed heavily towards "Cassava Mosaic Disease (CMD)," while "Cassava Bacterial Blight (CBB)" has the fewest images. Addressing this imbalance is necessary to improve classification performance.

- Size/Resolution:
    - All images have a consistent size of 800 x 600 pixels

4. Intended Classification Problem

The intended goal is to develop a classification model to distinguish between five categories of cassava plant health conditions. The output categories will be:

1. Cassava Bacterial Blight (CBB)
2. Cassava Brown Streak Disease (CBSD)
3. Cassava Green Mottle (CGM)
4. Cassava Mosaic Disease (CMD)
5. Healthy

Estimated Number of Images per Category: Based on the dataset information, the distribution across categories is highly imbalanced, which could impact the model's performance. The Cassava Bacterial Blight (CBB) category has the lowest number of images, with 921 samples. To address this imbalance without using resampling techniques, we plan to use 921 samples for each category. This approach ensures uniformity across classes while maintaining an adequate sample size for effective model training.

5. Useful Image Features for Categorization

The following types of image features are likely to be useful for classifying the cassava plant images:

- Edges and Contours: Since many of the diseases manifest through visible changes in leaf shape, edge detection could help distinguish between different classes.
- Color Histogram: The distribution of color variations in leaves, such as discoloration or mottling, will be an important feature for differentiating between diseases and healthy plants.
- Texture: The texture of the leaves might also indicate disease; for instance, some diseases cause rough patches or spotting that could be captured through texture analysis.
- Frequency: Differences in the visual patterns by decomposing the image into different frequency components may produce a unique pattern of visual disturbances.
- Vein structure: The pattern and prominence of leaf veins could be impacted by the diseases and may provide useful information.