Dataset Strategy-Project1

1. Project Overview:

We are building a dataset of grayscale leaf images from different plant species found on the MBMU campus. The goal is to train a deep autoencoder network that can reconstruct any random leaf image from this dataset.

2. Image Collection:

Collection Process

- **Location:** MBMU campus
- **Teams:** 8 teams participated
- **Species per team:** 2 species per team
- **Images per species:** 100-120 images
- **Species assigned to our team:**
    1. **Jamun**
    2. **Mango**
- **Lighting conditions:**
    - Sunlight
    - Artificial light
- **Image format:** JPG
- **Aspect ratio:** 1:1 (square images)
- **Camera settings:**
    - Ensure consistent focus and clarity.
    - Capture images from different angles for variety.
    - Maintain uniform exposure settings across images.

3. Dataset Upload & Collaboration:

- All teams will upload their collected images to a shared **GitHub repository**.
- Each team will organize their images into two separate folders (one per species).
- The dataset will be processed **only after all teams have uploaded their images**.
- Each team will independently carry out preprocessing steps on the entire dataset before training their models.

4. Preprocessing Steps:

Since images are collected in raw form, we need to apply preprocessing before training.

4.1 Organizing Images

- We will maintain **two main folders** initially—one for each species.
- The resolution (500x500 and 256x256) **will be changed during preprocessing**, not during collection.

4.2 Standardization & Cropping

- **Aspect ratio correction:** Ensure all images remain 1:1 by cropping out unnecessary space.
- **Resizing:** Images will be resized to **500x500** first, then to **256x256** for comparison.

4.3 Grayscale Conversion

- Convert all images from RGB to grayscale using OpenCV.

4.4 Noise Removal (Thresholding)

- Apply thresholding after grayscale conversion to remove shadows and unnecessary noise.
- Pixels above a certain value will be converted to **255 (white)**, and below it to **0 (black)**.
- **No smoothening will be applied** to retain the natural texture of leaves.

4.5 Data Augmentation (Optional)

- If required, we may apply:
  - Rotation
  - Scaling
  - Small transformations to increase dataset diversity

4.6 Intensity Normalization

- Adjust brightness and contrast to ensure all images have a similar intensity range.

5. Dataset Folder Structure:

Initially, the dataset will be structured as follows:

Dataset/
|— Jamun/
|— Mango/

Once preprocessing is done, we will introduce subfolders for different resolutions:

Dataset/
|— Jamun/
| ├— 256x256/

| ├── 500x500/
|── Mango/
| ├── 256x256/
| ├── 500x500/

6. Model Training Plan:

- We will train a **deep autoencoder** model using grayscale images.
- Training will first be done with **500x500** resolution images and later with **256x256** to compare results.
- The model will be evaluated based on **reconstruction error** (MSE, SSIM).

7. Next Steps:

1. Preprocess the dataset (grayscale conversion, noise removal, resizing)
2. .Augment data (optional).
3. Train the deep autoencoder with 500x500 images first, then with 256x256.
4. Evaluate and refine the model.

This approach ensures a well-structured dataset that is clean, uniform, and ready for deep learning applications.