

To submit the dataset strategy for project, you need to clearly outline the methods and steps you will use to collect, preprocess, and utilize the dataset for training the deep autoencoder. Here's a structured strategy you can follow:

Dataset Strategy for Grayscale Leaf Images:

1. Objective:

To create a high-quality dataset of grayscale leaf images from various plants/trees growing in MBMU campus. This dataset will be used to train a deep autoencoder that can reproduce any random leaf image from the campus.

2. Dataset Collection:

2.1 Identification

- Conduct a survey of the MBMU campus to identify all types of plants and trees present.

- Document the scientific/common names of plants for metadata purposes.

2.2 Image Collection

- Use a high-resolution camera to capture leaf images.
- Collect multiple samples of leaves from each plant/tree to account for variations (size, orientation, texture).
- Ensure diverse lighting conditions, but avoid extreme shadows or overexposure.
- Capture images from multiple angles to increase dataset diversity.

2.3 Labeling and Organization

- Label images with unique identifiers and include metadata (e.g., plant species, date of collection).
- Organize the dataset into a directory structure, e.g.:

Dataset/

Plant_1/

leaf1.jpg

leaf2.jpg

Plant_2/

leaf1.jpg

3. Preprocessing:

3.1 Conversion to Grayscale

- Convert all images to grayscale to remove color dependency.

3.2 Resizing

- Resize all images to a uniform dimension (e.g., 128x128 or 256x256 pixels) to standardize input for the autoencoder.

3.3 Normalization

- Normalize pixel values to the range [0, 1] for faster and more stable training.

3.4 Data Augmentation

- Apply augmentation techniques to enhance dataset diversity, such as:

- Rotation (e.g., ± 15 degrees).
- Horizontal and vertical flipping.
- Slight scaling.
- Noise addition.

3.5 Quality Check

- Ensure no corrupt or low-quality images remain in the dataset.

4. Dataset Splitting:

- Split the dataset into:
- Training Set (80%): Used for training the autoencoder.
- Validation Set (10%): Used for tuning and early stopping.

- Test Set (10%): Used for evaluating model performance.

5. Submission Format:

5.1 Dataset Files

- Save all images in a structured directory.
- Provide a metadata.csv file containing:
 - Image name
 - Plant/tree species
 - Collection details

5.2 Preprocessing Code

- Include the preprocessing scripts (e.g., Python code using OpenCV or PIL) in a separate directory.

5.3 Dataset Summary

- Document the dataset creation process in a report that includes:

- Total number of images.
- Number of unique plant/tree species.
- Preprocessing steps with sample images before and after preprocessing.

6. Dataset Submission Checklist:

1. Dataset Directory: Organized images in structured folders.
2. Metadata File: CSV file with relevant information.
3. Preprocessing Scripts: Code for reproducibility.
4. Dataset Report: Summary and methodology in a PDF or markdown file.

7. Tools and Libraries:

- Image Collection: DSLR/Mobile Camera.
- Preprocessing: Python, OpenCV, PIL, NumPy, Pandas.
- Autoencoder Framework: TensorFlow/Keras or PyTorch.