# Data Collection and Preprocessing Phase

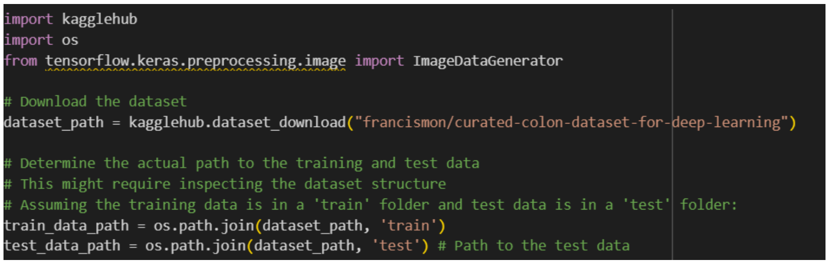
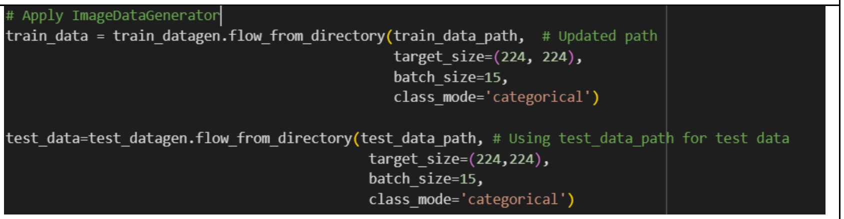
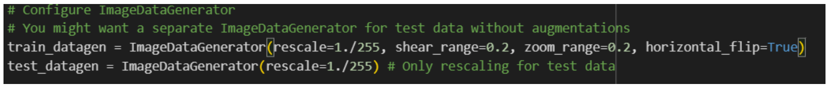
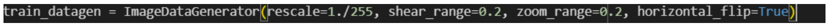
| | |
|---|---|
| Date | 15 March 2024 |
| Team ID | 739877 |
| Project Title | WCE Curated Colon Disease Classification using Deep Learning |
| Maximum Marks | 6 Marks |

## Preprocessing

The images will be preprocessed by resizing, normalizing, augmenting, denoising, adjusting contrast, detecting edges, converting color space, cropping, batch normalizing, and whitening data. These steps will enhance data quality, promote model generalization, and improve convergence during neural network training, ensuring robust and efficient performance across various computer vision tasks.

| **Section** | **Description** |
|---|---|
| Data Overview | The dataset is sourced from **Kaggle**, consisting of colonoscopy images labeled with different colon diseases. |
| Resizing | Images are resized to a target size of **224x224 pixels** |
| Normalization | Pixel values are normalized to the range **[0, 1]** by dividing by **255**. |
| Data Augmentation | Augmentation techniques like **shearing**, **zooming**, and **horizontal flipping** are applied to enhance the dataset. |

## Data Preprocessing Code Screenshots

| | |
|---|---|
| Loading Data | <br>import kagglehub<br>import os<br>from tensorflow.keras.preprocessing.image import ImageDataGenerator<br><br># Download the dataset<br>dataset_path = kagglehub.dataset_download("francismon/curated-colon-dataset-for-deep-learning")<br><br># Determine the actual path to the training and test data<br># This might require inspecting the dataset structure<br># Assuming the training data is in a 'train' folder and test data is in a 'test' folder:<br>train_data_path = os.path.join(dataset_path, 'train')<br>test_data_path = os.path.join(dataset_path, 'test') # Path to the test data |
| Resizing | <br># Apply ImageDataGenerator<br>train_data = train_datagen.flow_from_directory(train_data_path,  # Updated path<br>                                                 target_size=(224, 224),<br>                                                 batch_size=15,<br>                                                 class_mode='categorical')<br><br>test_data=test_datagen.flow_from_directory(test_data_path, # Using test_data_path for test data<br>                                             target_size=(224,224),<br>                                             batch_size=15,<br>                                             class_mode='categorical') |
| Normalization | <br># Configure ImageDataGenerator<br># You might want a separate ImageDataGenerator for test data without augmentations<br>train_datagen = ImageDataGenerator(rescale=1./255, shear_range=0.2, zoom_range=0.2, horizontal_flip=True)<br>test_datagen = ImageDataGenerator(rescale=1./255) # Only rescaling for test data |
| Data Augmentation | <br>train_datagen = ImageDataGenerator(rescale=1./255, shear_range=0.2, zoom_range=0.2, horizontal_flip=True) |