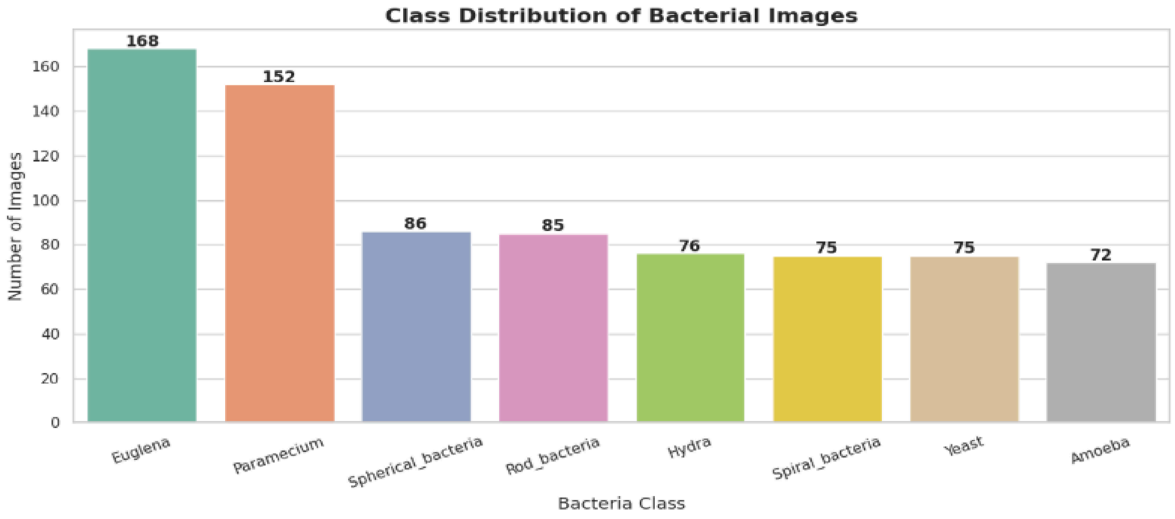
Assignment 2

Name: Abdullah Ahnaf

ID: 1000056118

Course code: CSE713

1. The dataset is collected from kaggle and also uploaded in my github (link: https://www.kaggle.com/datasets/mdwaquarazam/microorganism-image-classification) (Github link: https://github.com/Ahnaf0171/CSE713/tree/main/Dataset/Micro_Organism)
2. For preprocessing the image data set, I used some methods, and also i added my code file in my github (Github link: https://github.com/Ahnaf0171/CSE713/blob/main/codes/assignment2_ahnaf_bacteria_preprocessing.ipynb)
3. **Class count visualization:** the output is



4. **Remove duplicate and blurry images:**
   175 images were removed from the dataset because there were no duplicate or blurry images.
5. **Stratified_train_test_split_and_label_encoding_for_imbalanced_dataset:**
   My dataset is imbalanced because euglena (168) > amoeba (72), which is more than double. I knew it from my class count visualization. That's why I used stratified sampling techniques. For training data, it is 80%, and for testing data, it is 20%.
6. **Resize, normalize, sharpness, and contrast (only for training dataset):**

I got the mean and standard deviation from the training dataset
mean = np.array([0.50151639, 0.51555528, 0.51724466])
std = np.array([0.19444162, 0.19618673, 0.19317427])

7. **Resize and normalization (for test dataset using training data parameter):**
   I used to normalize the test dataset using the training dataset mean and standard deviation to prevent data leakage, and the model learns only from the training data, ensuring it remains unbiased
8. **Smart Data Augmentation (only for training data):**
   I used these methods, such as HorizontalFlip & VerticalFlip, Rotate, GaussianBlur & MotionBlur, RandomGamma, HueSaturationValue, and CLAHE, for more generalization and decreased overfitting.