**Name: Samiksha Sandeep Zokande**

**2448354**

**4MDS – B**

**Exploratory Data Analysis Image Dataset**

*Introduction:-*

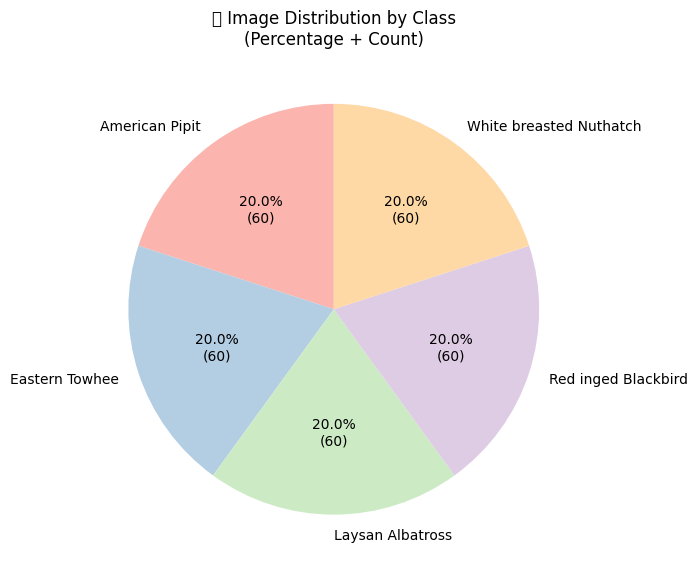
Datasets are raw information which when processed provide insightful information. The trends, patterns, relations etc. in a data are not visible to common eye to overcome this Exploratory Data Analysis plays a huge part. EDA is used mainly by data scientists to build visuals of a dataset which summarize characteristics of a data. Image dataset especially can be used for very through analysis and accurate results.

*Dataset:-*

The **Caltech-UCSD Birds-200-2011** (**CUB-200-2011**) dataset is the most widely-used dataset for fine-grained visual categorization task. It contains 11,788 images of 200 subcategories belonging to birds, 5,994 for training and 5,794 for testing. Each image has detailed annotations: 1 subcategory label, 15 part locations, 312 binary attributes and 1 bounding box. They expand the CUB-200-2011 dataset by collecting fine-grained natural language descriptions. Ten single-sentence descriptions are collected for each image. For the purpose of this study only 5 classes have been considered. The five classes are:



The whole dataset now consists of 300 images. Images in per class are:



Each class contains 60 images i.e. 20% of the whole data. This shows that the data is balanced and ready for pre-processing and further exploration.

*Data pre-processing:-*

The data needs to be prepared to be used in deep learning architectures. A well-structured pre-processing pipeline was applied.

* RGB(Red Green Blue) conversion: The data was transformed using torchvision framework into RGB(Red Green Blue).
* Resize: Images were of varied dimensions so the images were resized to 224\*224 px which aligns with the input size requirement of most of the models.
* Histogram equalization: It spreads out pixel intensities over the whole range to improve contrast.
* The Gaussian blur technique eliminates noise and detail from images while preparing them for subsequent edge detection processing steps.
* The images underwent normalization through ImageNet statistical parameters (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to achieve pre-trained model compatibility and faster training convergence.
* Data augmentation techniques like vflip, hflip, rotate, crop help in model generalization. RandomHosrizontalFlip, RandomRotate, ColorJitter were the data augmentation techniques used.

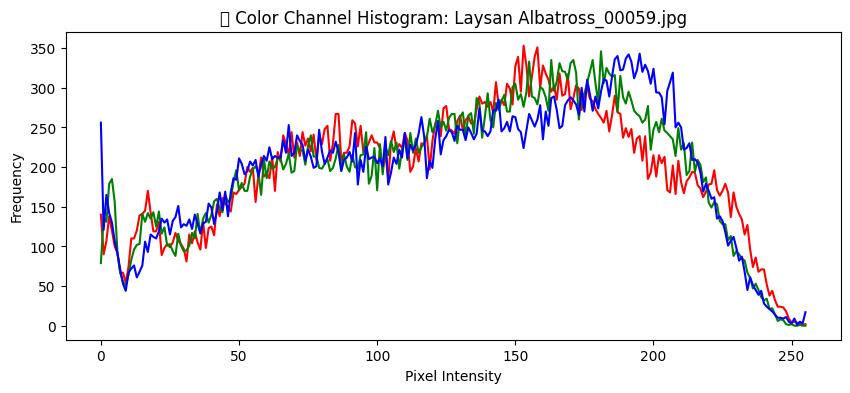
*Graphs and Interpretations:-*

*Image instances:*



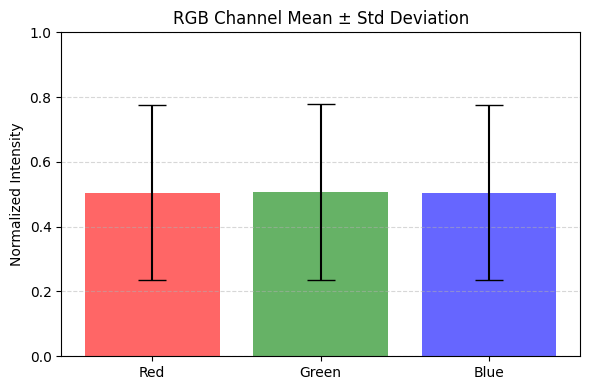
This above image depicts few of the images from the dataset after pre-processing, labelling and data augmentation has been applied. Five distinct birds species can be seen in the above image. Further exploration and studies can be carried out for classification of birds into 5 distinct classes.

*Color channel histogram:*



The color channel histogram of an instance image illustrates the distribution of pixels across red, green and blue. On X-axis the pixel intensity ranges form 0 to 350 and Y-axis represents the frequency. The image indicates that there is a fair and balanced exposure of all three pixels.

*RGB Channel Mean and Standard Deviation:*



The bar chart is titled “RGB Channel Mean +\_ Standard Deviation”. It represents the average intensity and variability of Red Green Blue channels. Each bar represents mean normalized intensity and standard deviation. This is due to the normalization applied on the dataset while pre-processing. This shows consistency across the dataset indicating that it is ready for further analysis.