|  |
| --- |
| **Team ID :** LTVIP2025TMID32402 |
| Project Name : Enchanted Wings: Marvels of Butterfly Species |

**Data Collection and Preparation**

Machine Learning (ML) depends heavily on data. It is the most crucial aspect that makes algorithm training possible. So, this section allows you to download the required dataset and outlines the preparation steps to ensure quality input for model training.

**Key steps involved in data collection and preparation include:**

* Identify or download the dataset (e.g., butterfly species image dataset).
* Inspect the dataset to understand its structure and labels.
* Clean the dataset by removing any corrupt or mislabeled images.
* Perform data preprocessing such as resizing, normalization, and format conversion.
* Apply data augmentation techniques to increase the diversity of training data.
* Split the dataset into training, validation, and test sets.
* **Collect the Dataset**
* There are many popular open sources for collecting data, such as **Kaggle.com**, the **UCI Machine Learning Repository**, and more.
* In this project, we have used **53 classes of playing cards** data. This dataset is downloaded from **Kaggle.com** and can also be accessed through the Kaggle API.  
  Please refer to the link below to download the dataset:
* 🔗 **[Download Dataset - Kaggle LINK]** *(Insert actual URL here)*
* Once the dataset is downloaded, the next step is to **read and understand the data** using visualization and analysis techniques.
* **Note:** There are multiple techniques for exploring and understanding data. Here, we are using some commonly used methods. You are encouraged to explore additional techniques as well.
* **Activity 1.1: Importing the Libraries**
* Before working with the dataset, import all the necessary libraries.  
  Typical libraries include:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import os

import zipfile

**Activity 1.2: Read the Dataset**

Datasets may come in various formats such as .csv, .xlsx, .txt, .json, or .zip. In our case:

* **Step 1:** Unzip the dataset using Python's zipfile module or extract manually.
* **Step 2:** Load the data using pandas.

Example:

python

CopyEdit

with zipfile.ZipFile("playing\_cards.zip", 'r') as zip\_ref:

zip\_ref.extractall("playing\_cards\_data/")

# Load the dataset (example assuming CSV)

df = pd.read\_csv("playing\_cards\_data/cards.csv")

df.head()

**Collect the dataset**

There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc.

In this project, we have used 53 classes of playing cards  data. This data is downloaded from kaggle.com or can be connected by using API. Please refer to the link given below to download the dataset.

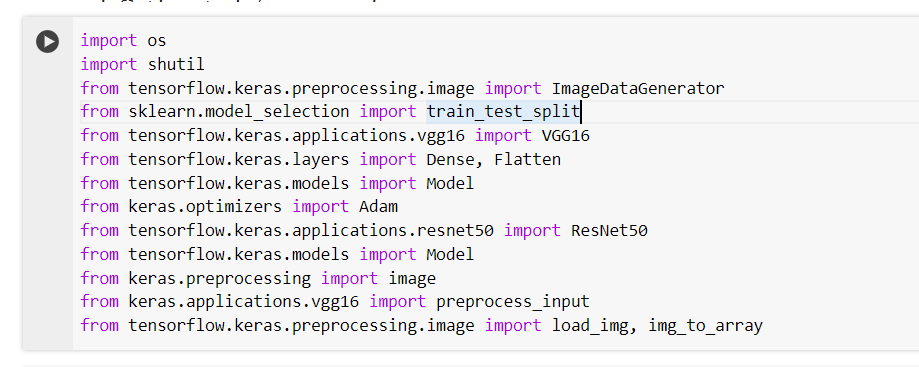
[LINK](https://www.kaggle.com/datasets/phucthaiv02/butterfly-image-classification)

As the dataset is downloaded. Let us read and understand the data properly with the help of some visualization techniques and some analyzing techniques.

Note: There are several techniques for understanding the data. But here we have used some of it. In an additional way, you can use multiple techniques.

Activity 1.1: Importing the libraries:

Import the necessary libraries as shown in the image. 



Activity 1.2: Read the Dataset:

* Our dataset format might be in .csv, excel files, .txt, .json, or zip files, etc. We can read the dataset with the help of pandas.

At first, unzip the data and convert it into a pandas data frame.

