Exploratory Data Analysis (EDA) Report

Date: 29.03.2024

In this report, a brief information will be given about the exploratory data analysis conducted on the dataset in order to show some aspects of data.

Exploratory Data Analysis

The problem I work on in this graduation project is that building a deep learning model to detect and classify the diseases occured in plant leaves of tomato, pepper bell and potato and deploying the trained model into an android mobile application to help farmers, engineers working in this area and even plant enthuasists. The reasons and final thoughts are explained in the 'Related works and competitor analysis' report.

Dataset research conducted on Kaggle and Google Datasets and two appropriate dataset has been found for the problem, PlantDoc and Plantvillage. PlantDoc dataset consits of 2598 images splitted into 27 different categories, mostly field images, taken from various angles, lighting condition and with various backgrounds, some laboratory images. Plantvillage has same aspects and features as PlantDoc dataset but having so much samples around 55000 splitted into 38 different categories. Since Plantvillage dataset has been carrying out same aspects as PlantDoc and there are more images than PlantDoc dataset which actually solves the lack of enough data problem in order to train a more generalized model, Plantvillage dataset has been picked and used throughout the model building process.

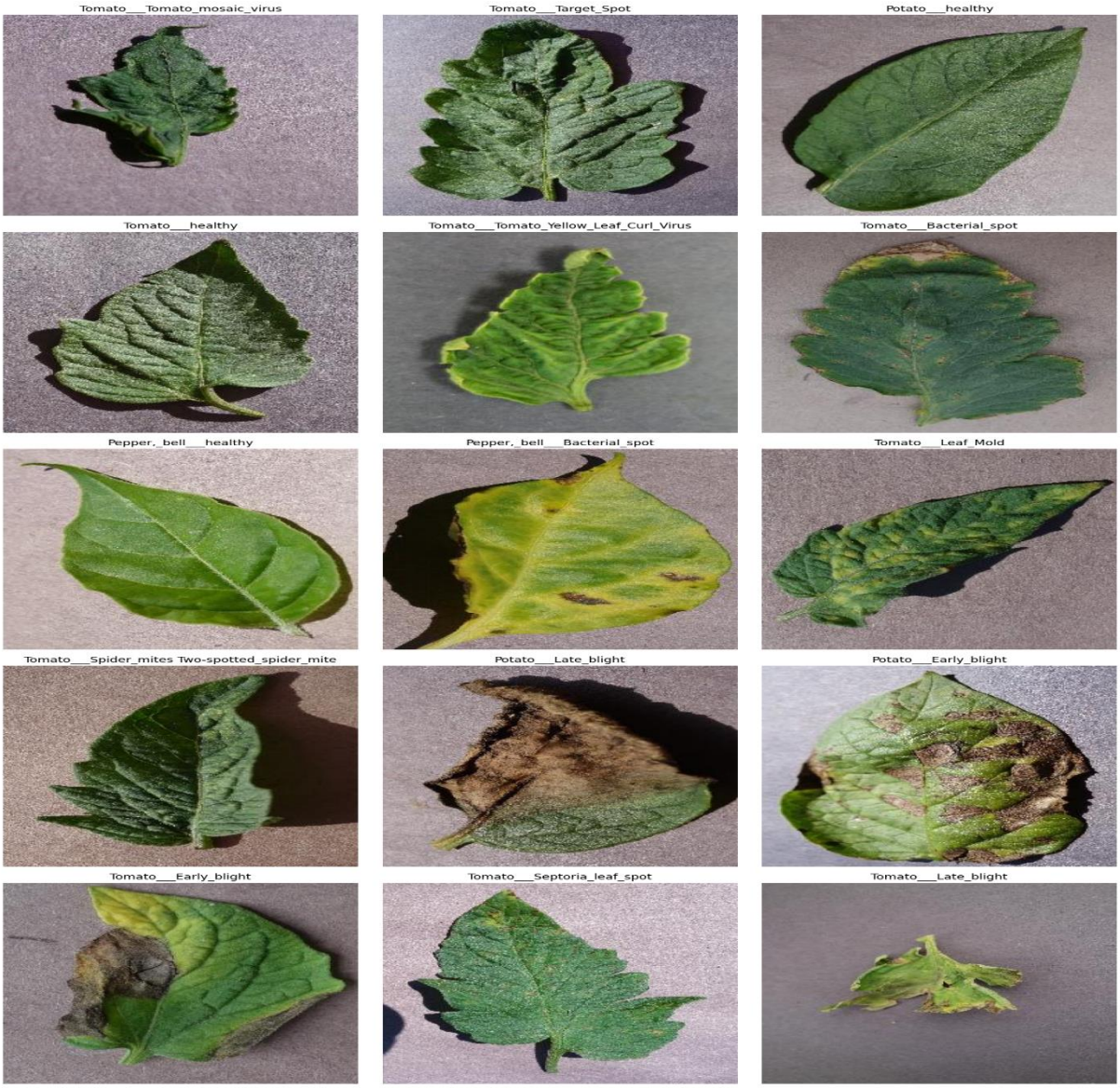
Original Plantvillage dataset which is created and exposed to public use of from a study which is shared here:

https://arxiv.org/ftp/arxiv/papers/1511/1511.08060.pdf

The original dataset itself is not found on the official website of PlantVillage, therefore dataset has been taken from the Mendelev Data website which is redirected from the Tensorflow website:

https://data.mendeley.com/datasets/tywbtsjrjv/1

Since the dataset consists of 38 categories and as I only need to 15 of them to need to classify Tomato, Pepper bell and Potato diseases, rest of them is deleted then uploaded to the Kaggle in order the speed up the download process while using it on Google Colab. Some samples from the dataset in a batch of size 32 with the corresponding labels or classes as shown:



As shown in the picture, some samples have plain background and no background noise, this may lead having some problems while using the model on the fly in the fields which have noiser backgrounds and other bad conditions. Despite the having enough number of data, this is the downside of the PlantVillage dataset and this concern will be evaluated and tested with real life data(image).

The dataset has total 22787 samples and each sample has two features image path and its corresponding label. The dataset consists of 15 files indicating either disease type of plant or it is healthy and each file consists of leaf images. All the information illustrated in jupyter notebook in Google Colab also the class distribution of the dataset, as shown below:

```
df.shape
```

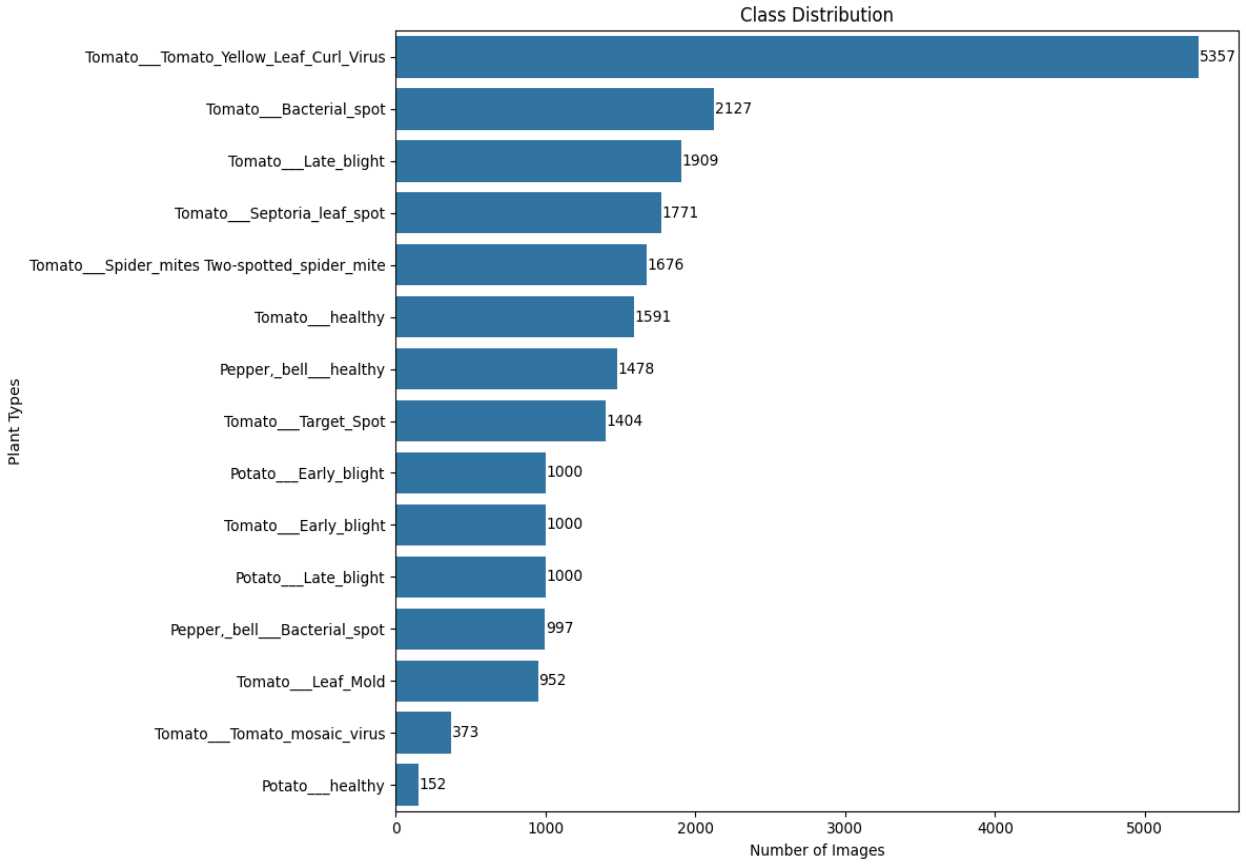
```
(22787, 2)
```

```
df.head()
```

	image_path	label
0	/content/Plant_leave_diseases_dataset_without_...	Tomato___Early_blight
1	/content/Plant_leave_diseases_dataset_without_...	Tomato___Early_blight
2	/content/Plant_leave_diseases_dataset_without_...	Tomato___Early_blight
3	/content/Plant_leave_diseases_dataset_without_...	Tomato___Early_blight
4	/content/Plant_leave_diseases_dataset_without_...	Tomato___Early_blight

```
[ ]  print("The classes:\n", np.unique(df['label']))
```

```
The classes:
 ['Pepper,_bell___Bacterial_spot' 'Pepper,_bell___healthy'
 'Potato___Early_blight' 'Potato___Late_blight' 'Potato___healthy'
 'Tomato___Bacterial_spot' 'Tomato___Early_blight' 'Tomato___Late_blight'
 'Tomato___Leaf_Mold' 'Tomato___Septoria_leaf_spot'
 'Tomato___Spider_mites Two-spotted_spider_mite' 'Tomato___Target_Spot'
 'Tomato___Tomato_Yellow_Leaf_Curl_Virus' 'Tomato___Tomato_mosaic_virus'
 'Tomato___healthy']
```



As shown in the class distribution chart, Tomato has 10 different classes where 9 of them belong to different types of tomato diseases and one of them belongs to healthy. Pepper bell has only 2 classes which are healthy and one type of disease. Potato has 3 classes, 2 of them belong to type of blight disease and one of them is healthy.

As a result, the dataset used in this study is an imbalanced dataset because of each class has different number of samples where majority of samples belongs to one type of plant that is Tomato and its category Yellow Leaf Curl Virus. Potato Healthy class has the least amount of samples. As the initial deduction, the trained model will yield better results while classifying the majority class and other Tomato classes and worse results while classifying other two type of plants. This makes evaluation process biased and only accuracy metric is not enough to understand the result of the model so that other metrics such as precision, recall and f scores (F1) will be used when evaluating the model.

Student Name: Hadican Munis

Student No: 20170702039