Imports

In [1]:

```
import pandas as pd

import os

import numpy as np

import tensorflow as tf

from PIL import Image

import shutil

from keras.preprocessing.image import ImageDataGenerator
from tensorflow.keras.utils import array_to_img, img_to_array, load_img

pip install split-folders

import splitfolders

import matplotlib.pyplot as plt

from zipfile import ZipFile

import random
```

Load and Format Data

In [38]:

```
os.getcwd()
```

Out[38]:

```
'C:\\Users\\musaa\\Documents\\Cassava\\Cassava-Disease-Classification'
```

In [31]:

```
with ZipFile('Data/cassava-leaf-disease-classification.zip', 'r') as zObject:
    zObject.extractall(path="Data")
```

In [33]:

```
labels_df = pd.read_csv("Data/train.csv")
```

In [34]:

```
label_dict = {"0": "Cassava Bacterial Blight (CBB)",
              "1": "Cassava Brown Streak Disease (CBSD)",
              "2": "Cassava Green Mottle (CGM)",
              "3": "Cassava Mosaic Disease (CMD)",
              "4": "Healthy"}
```

In [37]:

```
len(os.listdir("Data/train_images"))
#checking directory size
```

Out[37]:

```
21397
```

In [ ]:

In [ ]:

```

```

In [39]:

```
#Place to move my data
os.mkdir("Data/CBB")
os.mkdir("Data/CBSD")
os.mkdir("Data/CGM")
os.mkdir("Data/CMD")
os.mkdir("Data/Healthy")

os.mkdir("Data/TTVS") # Train Test Validation Split
```

In [40]:

```
labels_df
```

Out[40]:

	image_id	label
0	**1000015157.jpg**	**0**
1	**1000201771.jpg**	**3**
2	**100042118.jpg**	**1**
3	**1000723321.jpg**	**1**
4	**1000812911.jpg**	**3**
...	**...**	**...**
21392	**999068805.jpg**	**3**
21393	**999329392.jpg**	**3**
21394	**999474432.jpg**	**1**
21395	**999616605.jpg**	**4**
21396	**999998473.jpg**	**4**

21397 rows × 2 columns

In [41]:

```
labels_df["label"].value_counts()
#Count of each class of images
```

Out[41]:

```
3    13158
4     2577
2     2386
1     2189
0     1087
Name: label, dtype: int64
```

In [42]:

```
# making lists of all filenames by class
list0 = []
list1 = []
list2 = []
list3 = []
list4 = []
for x in labels_df.index:
    if labels_df.iloc[x,1] == 0:
        list0.append(labels_df.iloc[x,0])
    elif labels_df.iloc[x,1] == 1:
        list1.append(labels_df.iloc[x,0])
    elif labels_df.iloc[x,1] == 2:
        list2.append(labels_df.iloc[x,0])
```

```
    elif labels_df.iloc[x,1] == 3:
        list3.append(labels_df.iloc[x,0])
    elif labels_df.iloc[x,1] == 4:
        list4.append(labels_df.iloc[x,0])
```

In [43]:

```
label_dict #confirming label names for next step
```

Out[43]:

```
{'0': 'Cassava Bacterial Blight (CBB)',
 '1': 'Cassava Brown Streak Disease (CBSD)',
 '2': 'Cassava Green Mottle (CGM)',
 '3': 'Cassava Mosaic Disease (CMD)',
 '4': 'Healthy'}
```

In [44]:

```
#moving images to respective directories
for x in list0:
    shutil.move("Data/train_images/"+x, "Data/CBB")
for x in list1:
    shutil.move("Data/train_images/"+x, "Data/CBSD")
for x in list2:
    shutil.move("Data/train_images/"+x, "Data/CGM")
for x in list3:
    shutil.move("Data/train_images/"+x, "Data/CMD")
for x in list4:
    shutil.move("Data/train_images/"+x, "Data/Healthy")
```

In [46]:

```
#Sanity Check
print("CBB", len(os.listdir("Data/CBB")))
print("CBSD", len(os.listdir("Data/CBSD")))
print("CGM", len(os.listdir("Data/CGM")))
print("CMD", len(os.listdir("Data/CMD")))
print("Healthy", len(os.listdir("Data/Healthy")))
```

```
CBB 1087
CBSD 2189
CGM 2386
CMD 13158
Healthy 2577
```

In [47]:

```
splitfolders.ratio("Data/Sorted", output = "Data/TTVS", seed = 1,
                   ratio = (.6, .2, .2))
```

In [48]:

```
#Sanity Check
for x in os.listdir("Data/TTVS"):
    for i in os.listdir("Data/TTVS/"+x):
        print(x, i, len(os.listdir("Data/TTVS/"+x+"/"+i)))
```

```
test CBB 218
test CBSD 439
test CGM 478
test CMD 2633
test Healthy 516
train CBB 652
train CBSD 1313
train CGM 1431
train CMD 7894
train Healthy 1546
val CBB 217
val CBSD 437
val CGM 477
val CMD 2631
```

```
val Healthy 515
```

Data Augmentation

In [49]:

```
#Generate additional images to rectify imbalnce of data
datagen = ImageDataGenerator(
        width_shift_range=0.5,
        height_shift_range=0.5,
        rotation_range= 20,
        vertical_flip=True,
        horizontal_flip=True,
        brightness_range= (.7,1.3),
        fill_mode="wrap"
        )
```

In [51]:

```
source = 'Data/TTVS/train/CMD'
os.mkdir("Data/ExtraImages")
dest = 'Data/ExtraImages'
files = os.listdir(source)
no_of_files = 2894

for file_name in random.sample(files, no_of_files):
    shutil.move(os.path.join(source, file_name), dest)
```

In [52]:

```
len(os.listdir(source))
```

Out[52]:

```
5000
```

In [53]:

```
def create_aug(directory, n):
    for i in os.listdir(directory):
        if "aug" not in i:
            imagepath = directory+"/"+i
            img = np.array(Image.open(imagepath))
            img = img.reshape((1,) + img.shape)  # this is a rank 4 array
            length = len(os.listdir(directory))
            if length >= n:
                    break
            for batch in datagen.flow(img, batch_size=1,
                                      save_to_dir=directory,
                                      save_prefix = "aug",
                                      save_format= "jpg"):
                length = len(os.listdir(directory))
                if length >= n:
                    break
```

In [54]:

```
create_aug("Data/TTVS/train/CBB", 5000)
create_aug("Data/TTVS/train/CBSD", 5000)
create_aug("Data/TTVS/train/CGM", 5000)
create_aug("Data/TTVS/train/Healthy", 5000)
```

In [ ]:

```
#Sanity Check
for x in os.listdir("Data/TTVS"):
    for i in os.listdir("Data/TTVS/"+x):
        print(x, i, len(os.listdir("Data/TTVS/"+x+"/"+i)))
```