Task Name:

**Data Preprocessing**

Task Leaders:

**Tariq Jamil, Mohamed Soliman, Muhammed Shahroze**

Task Weekly Meeting Time:

**Sunday – 9pm UAE time**

**Task Objectives:**

This section would be responsible

- Collect the image as per required classification

- Consolidate into a single dataset, by removing

\*duplicates

\*sized less than 300x300 images

\*correction for ARGB to RGB

\*saving to JPEG format.

- Perform EDA

**Pipeline**

* *Image Collection through own effort, or web-sources, while classifying into 4 classes (crack, groove, rut, subsidence)*
* *Creation of structured processed datasets out of collected images with appropriate hierarchy and ground-truth annotations.*
* *To perform EDA*

**Completed Tasks:**

* During week1, around 17 datasets were contributed by various team members  
  <https://drive.google.com/drive/folders/10MzMMadzeuKmJCjXq7Eb0DlGE-STxS5G?usp=share_link>
* On the basis of these datasets 3 team members prepared processed and structured datasets (@Muhammad Shahroz, @Chithra Priya Janardhana, @Alaa Sedeeq)  
  <https://drive.google.com/drive/folders/1r056vjgKH4QZrHdv-ZgDauQznLIt-PN_?usp=share_link>

**Challenges**

* Classes 'rut' and 'subsidence' have fewer images.
* We should decide how to achieve a balanced dataset!

**Limitations:**

*NIL*

**EDA Report:**

\*------------------------------------------------------------------\*

number of images | 7753

dtype | uint8

channels | [3]

extensions | ['jpg']

min height | 300

max height | 4608

mean height | 666.7090158648265

median height | 506

min width | 300

max width | 5775

mean width | 892.1849606603895

median width | 720

mean height/width ratio | 0.747276680579028

median height/width ratio | 0.7027777777777777

recommended input size(by mean) | [664 896] (h x w, multiples of 8)

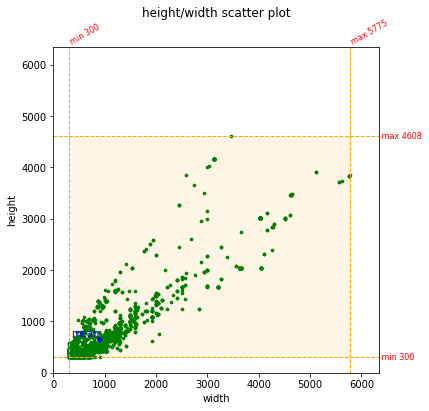
recommended input size(by mean) | [672 896] (h x w, multiples of 16)

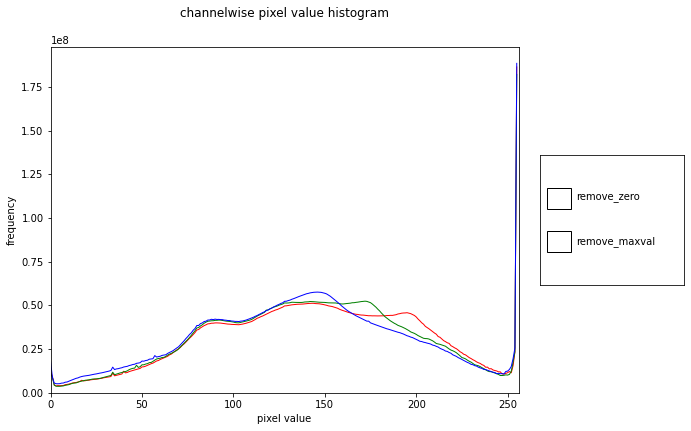
recommended input size(by mean) | [672 896] (h x w, multiples of 32)

channel mean(0~1) | [0.5148, 0.5066, 0.4944]

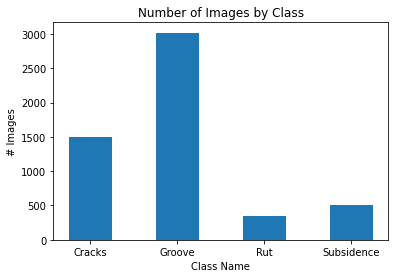
channel std(0~1) | [0.1906, 0.1862, 0.1891]

\*------------------------------------------------------------------\*

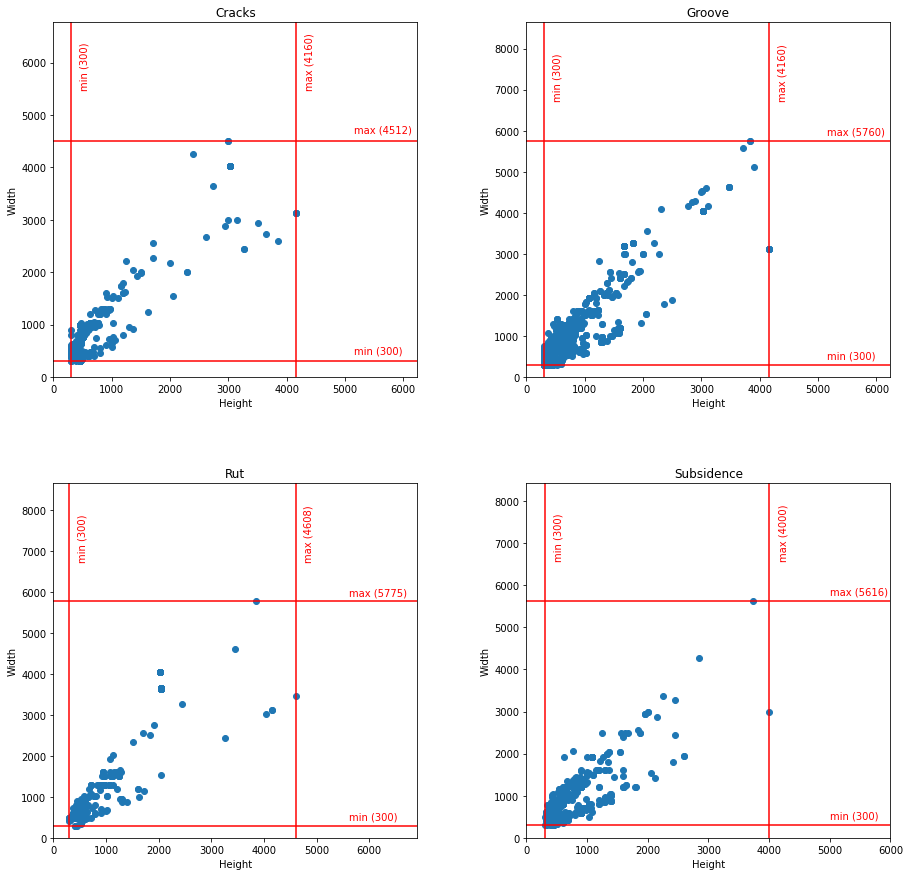
**

**

*Total number of Images in each class:*

**

*Distribution of Each individual Classes Height/Width*

**