Description

The data is engineered from an underlying image dataset and represent as multi-class classification problem. A large number of different features have been extracted to yield this structured data. The data has been rendered more demanding by incorporation of outliers and other challenging data science aspects. The data will be hosted on Kaggle and access link provided to the students shortly. We will hold presentations in the final week.

Rules

- Use of external data set or pre-trained models is strictly not allowed.
- It is an individual project

Dataset

- Train.csv contains the training data. Each row represents a data-point (i.e., an image) in the form of k comma-separated values. In each row, the first k-1 values represent the engineered features and the last value represents the class label.
- The first 200 features in every row are pixel values from randomly selected corresponding positions in the underlying image. The next 144 values represent histogram of oriented gradient features. Then there are 24 features engineered from GLCMs. Subsequently, the next 1024 values are randomly selected from the image with the additional constraint that the selection scheme is different for every image.

- Test.csv contains the test data. Each row has the aforementioned k-1 features. The class labels are of course omitted from the test data points.