Omar Mousa
Melissa Ortiz
Alex Richards
CMPE 188

Project Proposal

Abstract:

Our team is conducting an analysis of multiple data preprocessing pipelines and image classification algorithms. The dataset we've chosen is a collection of high-resolution images of human retinas, many of which are classified as either healthy or one of four degrees of diabetic severity. The tools we're using include Scikit-learn, PySpark, and Tensorflow. Some techniques we are intending to utilize include batch PCA for dimensionality reduction, KNN, and convolutional neural networks (and possibly other types of neural networks). The tools we will be including in our algorithms are different ways to implement image recognition. Furthermore, using these three tools we can compare the methods and determine which one will give us an idea that best predicts the degree of diabetic severity based on the retina images we will be using in our dataset. Ultimately, we are going to expand upon existing solutions for this particular image classification task utilizing newer techniques and benchmarking their performance against the Kaggle solutions.

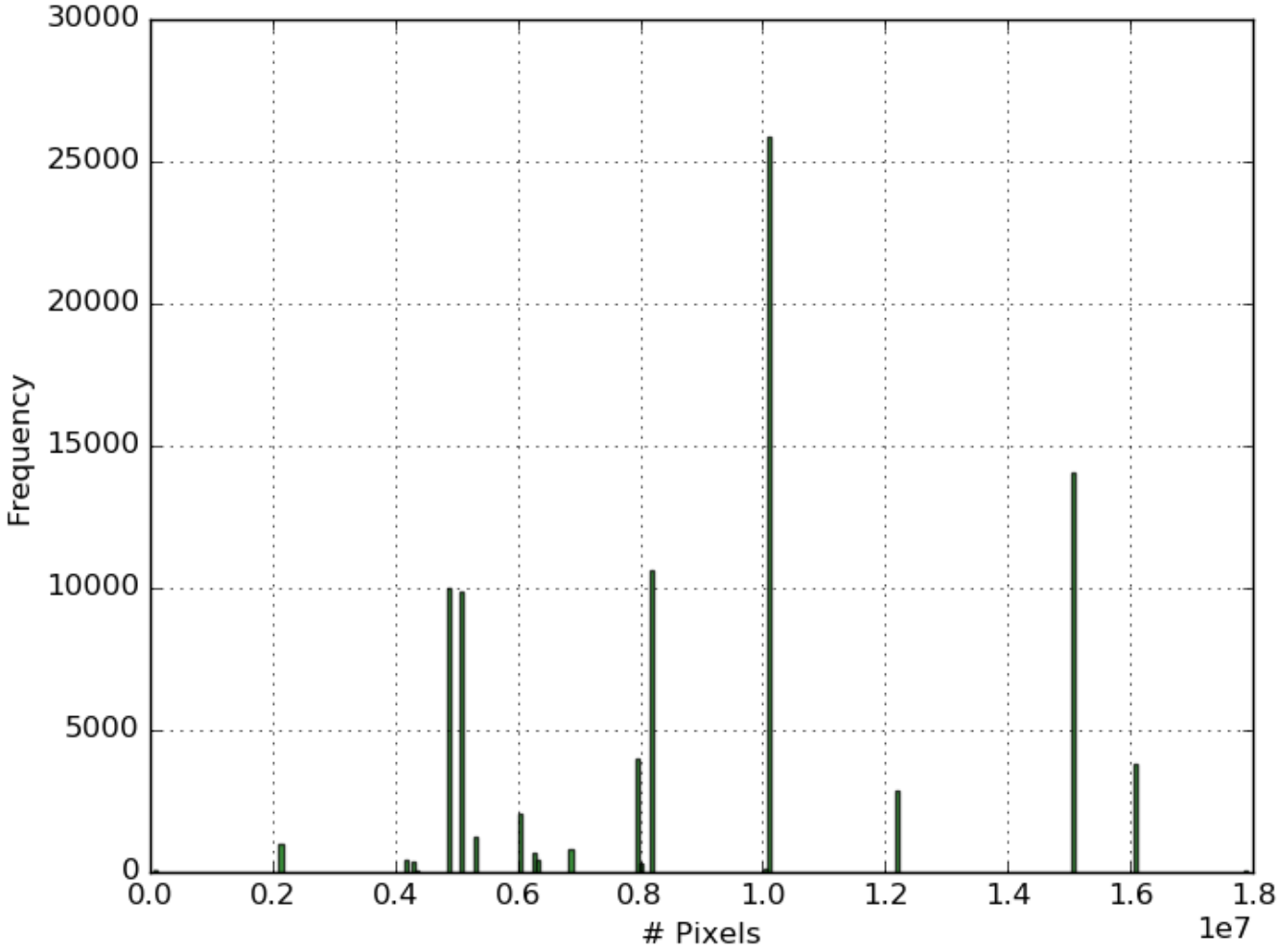
Dataset (https://www.kaggle.com/c/diabetic-retinopathy-detection):
Images are of either the left or right eye -- there are 44,251 individuals represented.

File Sizes / Counts:
Training image folder size: 37.9 GB
Testing image folder size: 57.8 GB
Number of training images: 35,126
Number of testing images: 53,576
Average jpeg image size: 1.0789 MB

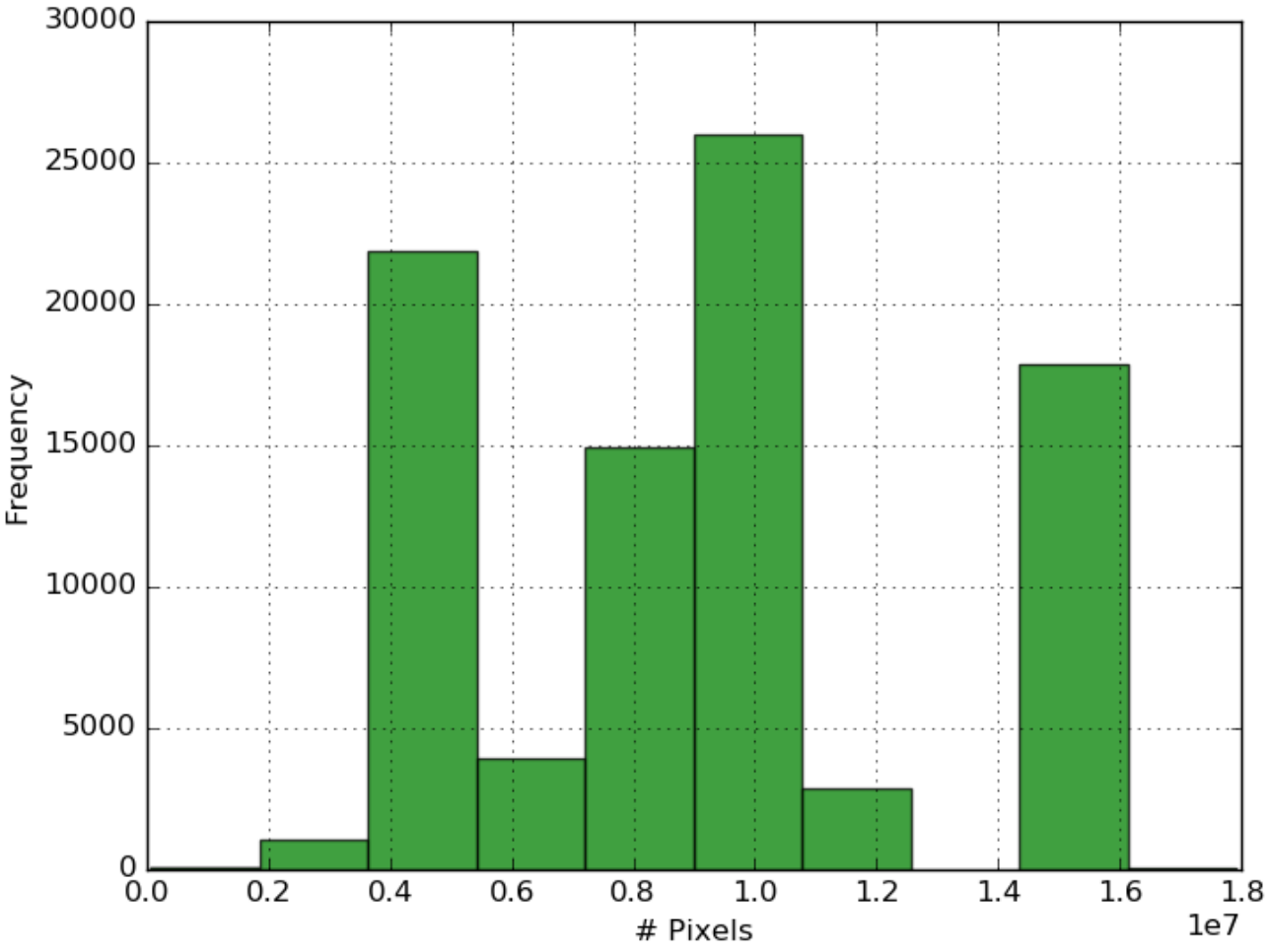
Pixel Statistics:
Number of discrete pixel counts: 31
Minimum: 67,520
Maximum: 17,915,904
Median: 10,077,696
Average: 9,340,450

(Detailed statistical plots of the distribution of data features on the next three pages).

Pixel Dimension Frequencies with $\sqrt{\#\ images}$ Bins



Pixel Dimension Frequencies with 10 Bins



Frequency per Exact Pixel Count