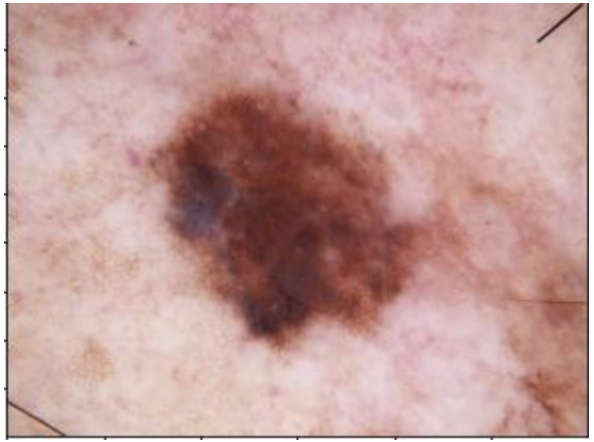
CSC 371 Final Project Data Exploration

Ian Hall and Eric Xu

Overview of Project

- Identifying type of skin cancer from image data
- Typically, identifying the type of skin cancer requires the use of microscopes and general consensus from a group of doctors
- Goal is to create a model that can accurately identify skin cancer type from an image to simplify the problem

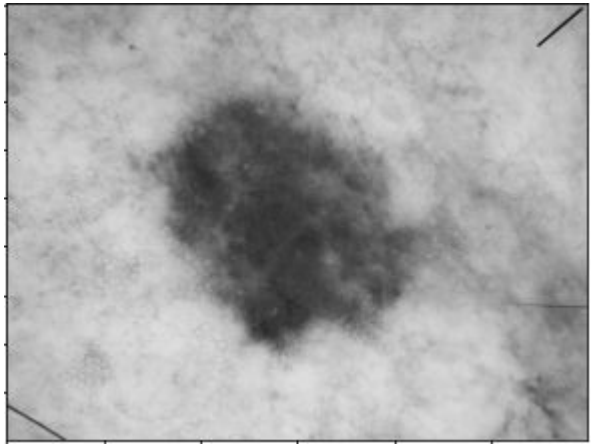


General Information about the Data

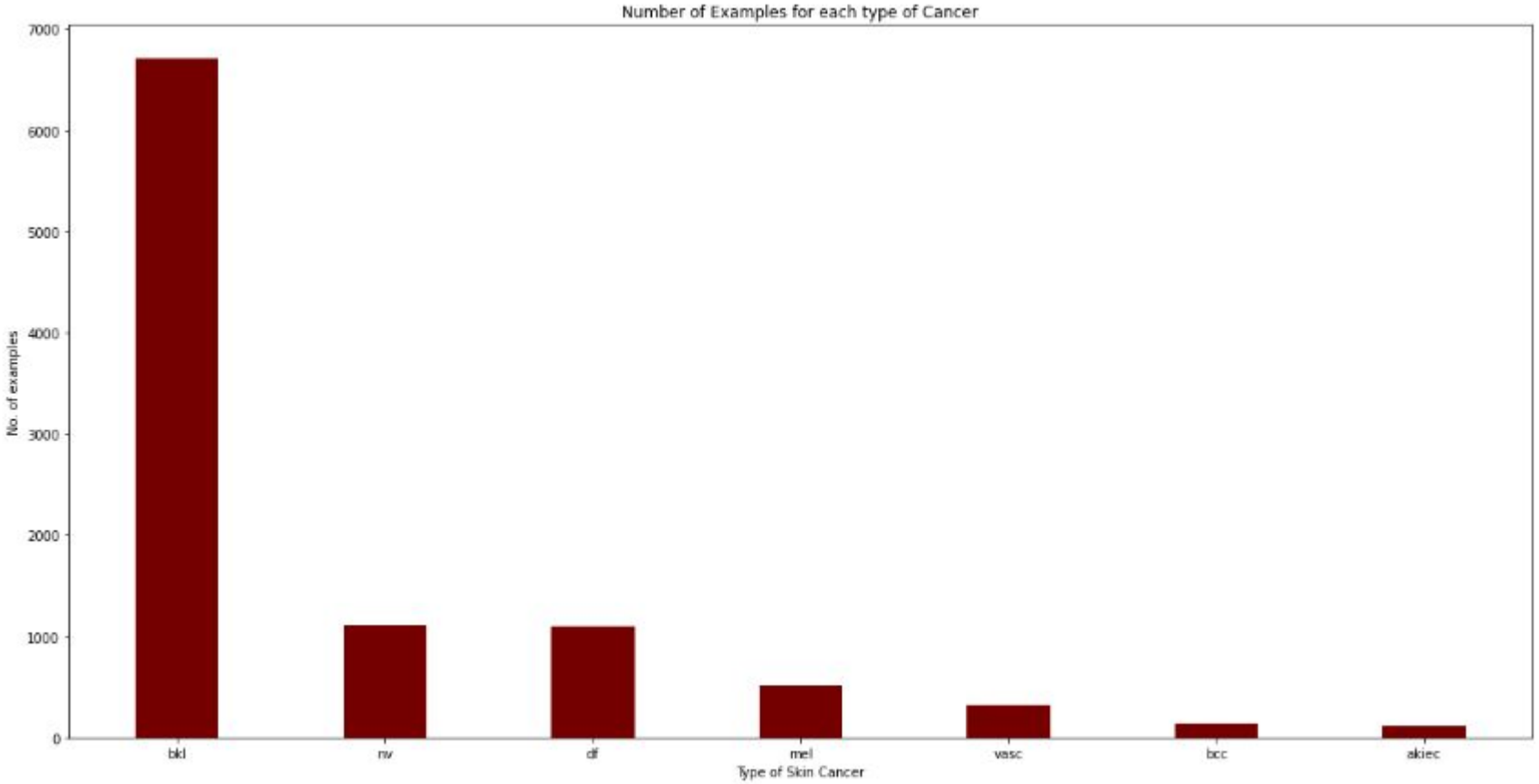
- 10015 unique data points
- Each data point consists of:
  - Patient ID Number
  - Picture of the Skin Cancer
  - Type of Skin Cancer (7 types)
  - How the Skin Cancer was identified
  - Age of Patient
  - Sex of Patient
  - Where the Skin Cancer was located

Image Data

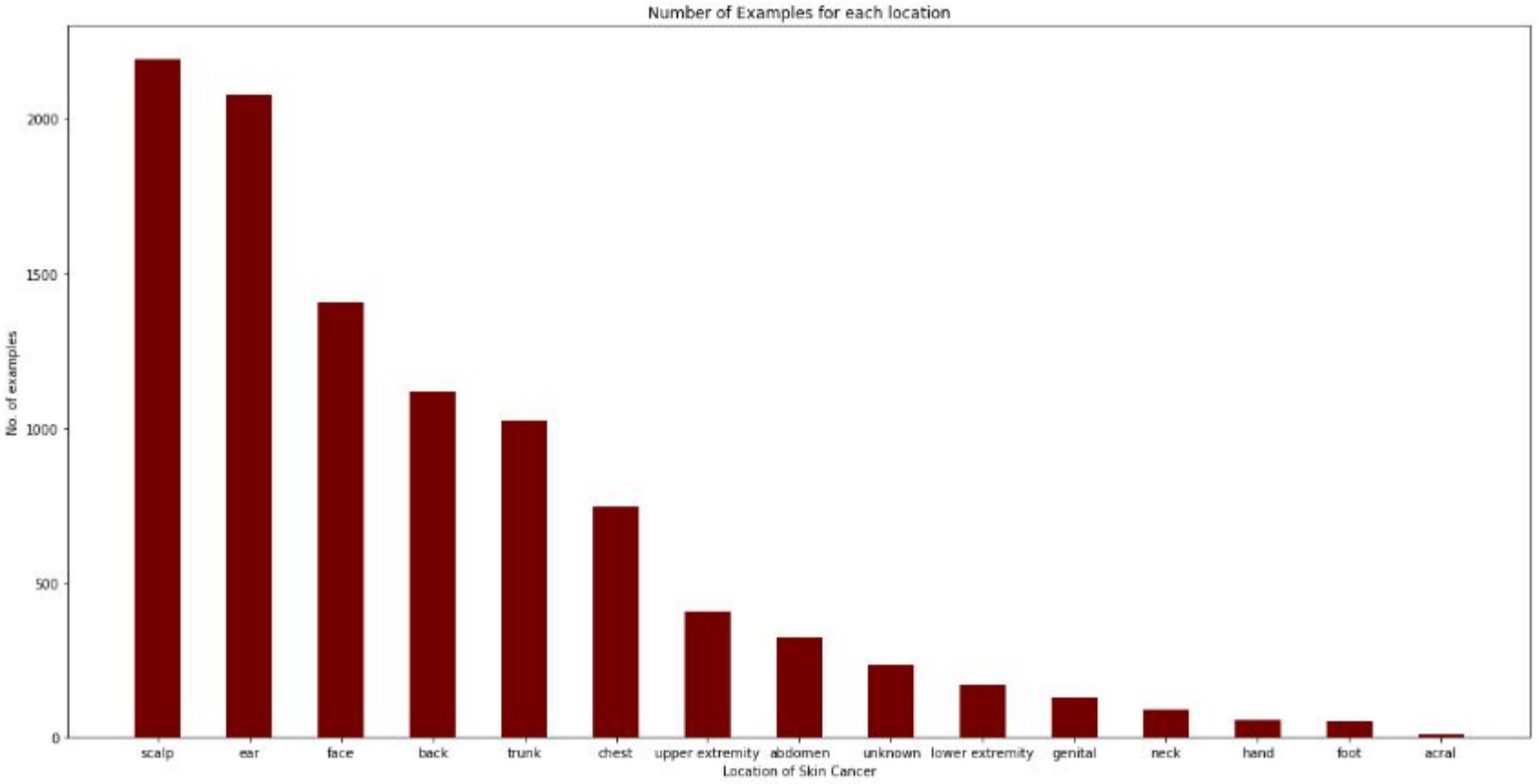
- All of the images are in the format of 450x600 pixels
- Each pixel contains 3 values (R,G,B)
- Considered turning the images into grayscale, but one of the pre-trained models actually uses RGB values
- Pre-trained model says it can take in any input size, but we may adjust size of pixels using max pooling or other techniques



Distribution of Skin Cancer Types



Distribution of Location of Skin Cancer



Normalization

Two ways:

- preprocessing.StandardScaler()
  - Mean = 0, Std = 1
- preprocessing.MinMaxScaler()
  - Normalize data into range of (0,1)

The choice of normalization methods should be made by comparing the performance of the final model.

Possible Models

- Neural Network
- SVM
  - Easy and effective