Christopher Lawton

Capstone 2 proposal

**Benign or Malignant?**

Current literature in cancer epidemiology has shown that higher density in breast tissue increases risk for breast cancer pathogenesis. During my last position at Kaiser Permanente, I worked on a mammographic breast density study in which we were trying to measure risk of contracting cancer based off of breast tissue density. This measurement was done with outdated software and basic techniques.

The cancer imaging archive is an online repository with an amalgam of different medical images surrounding cancer. In the archive is the CBIS-DDSM, a collection of mammograms that were taken from different mammography studies. It contains images that have indicated masses (mass or calcifications) and have a pathological label, benign or malignant as well as other features. The data set is broken up into two different sections, mass and calc, both are growths that can be seen through mammograms. There are four different image types, Craniocaudal (CC), Mediolateral oblique (MLO), a marked image indicating location of growth, and a cropped image of the region of interest (seen below). Each image hovers around 5904x3200 pixels in size excluding the cropped images which are around 214x195.

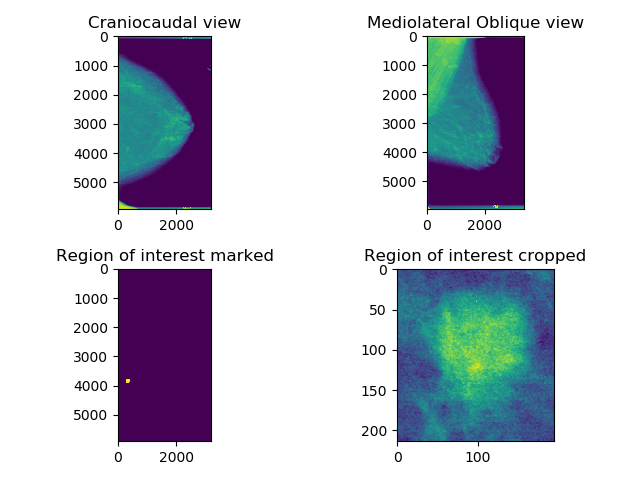
**Features:**

patient id, breast density, left or right breast, image view, abnormality id, abnormality type, mass shape, mass margins, assessment, pathology, subtlety.

**Image counts:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Type | CC | MLO | Total |
| Mass | Benign | 421 | 491 |  |
| Malignant | 363 | 421 |  |
| Total | 784 | 912 | 1696 |
| Calc | Benign | 570 | 629 |  |
| Malignant | 318 | 355 |  |
| Total | 888 | 984 | 1872 |

One concern I have with the data is due to patients having multiple scans of the same image view. The only field that differs between them is the abnormality id. I will need to learn more about the data to see what the differences between the images are and if using all images will be ok. However, there are 892 unique IDs in the mass set and 753 unique IDs in the calc set, so each person will at least have one CC and one MLO image that are unique.



MVP:

1. Create a neural network to classify if an image is Benign or Malignant for the CC view initially.
2. Using other features, test different supervised classifiers (random forest, boosting, etc.) to compare to the neural network.

MVP+:

1. Train a second network on the MLO view
2. Run both networks in parallel to predict in an ensemble manner