# Applied Data Science: Machine Learning Capstone Project Proposal

By Eric Scuccimarra ([skooch@gmail.com](mailto:skooch@gmail.com))

2018-03-08

## Problem

Breast cancer is the second most common cancer in women worldwide. About 1 in 8 U.S. women (about 12.4%) will develop invasive breast cancer over the course of her lifetime. The five year survival rates for stage 0 or stage 1 breast cancer are close to 100%, 93% for stage II, 72% for stage III and 22% for stage IV. This means that early detection can greatly improve the chances of surviving breast cancer.

The ability to automatically detect lesions and predict the probability of their being malignant would be a useful tool for doctors, and would make early detection of breast cancer more likely, and this improve survival rates.

## Dataset

The dataset is the MIAS Mammography data available on Kaggle[1]. The dataset contains images of mammography scans, labels and annotations. The dataset contains 330 mammogram scans, of which 207 are normal with the rest classified into six types of abnormalities. Each scan includes both a left and a right scan for the patient.

The data is annotated with the type of background fatty tissue, the class of the scan, whether the abnormality is benign or malignant, and if abnormal, the center and radius of the abnormality. The high res images are 1024x1024 pixels with 64x64 low-res versions also available.

## Analysis and Methods

Based on the “Standard ML Classifiers” kernel [2] it seems that standard machine learning techniques will not provide very good results on the pixel data.

The first step in approaching this problem will be to apply a convolutional neural network. I foresee some issues with this approach in that the low-res scans may be too small to be useful, and the abnormalities in the high res scans may be difficult to detect due to the large image size. The abnormalities have a range of radii from 3 to 197 pixels, with a mean of 48 pixels.

If the convolutional neural network does not perform the next step would be extracting features from the high-res images with a pretrained neural network. Two networks which can be evaluated for this purpose are OverFeat[3] and VGG[4]. The feature data can be used with other machine learning techniques, including SVM, KNN, decision trees, and non-convolutional neural networks.

## Process

My first step would be to feed the low-res images through a convolutional neural network. The next step would ideally be to feed the high-res images through a ConvNet, although the computation required to analyse 1024x1024 images may make this difficult or impossible.

The next step would be to extract features using pretrained ConvNets and evaluate that data with k-nearest neighbors, SVM, and fully connected neural networks. This would involve finding and evaluating pre-trained models to use in combination with the various machine learning techniques.

If the results from these techniques are not satisfactory the final step would be to review research into this subject to see if there are any techniques which can be applied to this problem.

## References

1. MIAS Mammography Dataset - <https://www.kaggle.com/kmader/mias-mammography>
2. Standard ML Classifiers kernel - <https://www.kaggle.com/kmader/standard-ml-classifiers>
3. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Neural Networks - <https://arxiv.org/abs/1312.6229>
4. Very Deep Convolutional Networks for Large-Scale Image Recognition - <https://arxiv.org/abs/1409.1556>
5. Breast Mass Classification from Mammograms using Deep Convolutional Neural Networks - https://arxiv.org/pdf/1612.00542.pdf