# 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 cancers are close to 100%, but the rates go down dramatically for later stages: 93% for stage II, 72% for stage III and 22% for stage IV. This means that early detection can greatly improve the chances of survival.

Human recall for identifying lesions is estimated to be between 0.75 and 0.92 [5], which means that as many as 25% of abnormalities may go undetected. The ability to automatically detect lesions and predict the probability of their being malignant would be a useful tool for doctors, and could greatly improve survival rates.

As a note, I believe that for this problem recall is more important than accuracy, as a false negative could potentially be life threatening while a false positive would likely be reviewed by a human, and in the worst case would only lead to an unnecessary biopsy.

## 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 centre and radius of the abnormality. Each image is available in a 64x64 low-res version as well as a 1024x1024 high-res version.

## 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. This leaves several approaches to the problem.

The first approach is feeding the images into a ConvNet. However the low-res images may be too small to extract useful information from. The abnormalities have a range of radii from 3 to 197 pixels, with a mean of 48 pixels in the high-res images. Compressing these to 64x64 would make the abnormalities possibly too small to detect with a mean radius of 3 pixels. The high-res images may be difficult to work with due to their size and the computation required to process them.

The second approach is to extract features from the high-res images using a pre-trained neural network. Two networks which can be evaluated for this purpose are OverFeat [3] and VGG [4]. The extracted feature data can be used with other machine learning techniques, including SVM, KNN, decision trees, and fully-connected neural networks.

There are several options as to how to classify the data. The data is labelled into seven classes, one for normal and six types of abnormalities. The abnormalities are further classified into benign and malignant. The first option is to classify as simply normal or abnormal, the second to classify into the type of abnormality, and the third to further classify as benign or malignant. These options will be evaluated to determine if they will affect the results.

## Process

The first step would be to evaluate a ConvNet. We have already established that the low-res images will likely not have the detail to be useful for this purpose. The high resolution images should be suitable for this purpose, however, as previously mentioned, the computation required may make this impractical.

The next step would be to extract features using pre-trained ConvNets and evaluate that data with standard machine learning techniques, including k-nearest neighbours, SVM, random forests, and fully connected neural networks. This will involve evaluating the pre-trained models in combination with the various classifying techniques.

Of course, other possible approaches may come up during the process described above which may yield better results. If any other methods should arise that seem to have potential they would also be explored.

## Communication

The final result will be an overview of the techniques evaluated along with the results, and a detailed description of the technique(s) that provided the best results.

The best accuracy result I have been able to find is 0.929 by GoogLeNet [6] on mammography scans from a different dataset. I do not expect to be able to match this, but I hope to be able to achieve significantly better results than the most frequent baseline accuracy of 0.62.

## 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. Seramet et al., OverFeat: Integrated Recognition, Localization and Detection using Convolutional Neural Networks - <https://arxiv.org/abs/1312.6229>
4. Simonyan et al., Very Deep Convolutional Networks for Large-Scale Image Recognition - <https://arxiv.org/abs/1409.1556>
5. Levy et al., Breast Mass Classification from Mammograms using Deep Convolutional Neural Networks - https://arxiv.org/pdf/1612.00542.pdf