# Applied Data Science: Machine Learning Capstone Project Proposal

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

2018-03-15

## 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.

## Dataset

I began by working with the MIAS Mammography dataset from 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. This dataset proved to be difficult to work with due to its small size.

The Wisconsin Diagnostic Breast Cancer data set [6] contains features extracted from fine needle aspiration biopsies and classifies the masses into benign or malignant. As it comes from biopsies, this provides a much finer degree of detail which can be used to predict the severity of a mass. I will be using this data to gain more insight into the characteristics of abnormalities and attempting to predict the type and severity of abnormality.

The WDBC dataset is only applicable to masses which have been biopsied, so is not usable for detecting abnormalities from scans. Since the MIAS dataset is not suitable for training a ConvNet, I then found the DDSM dataset which was a larger set of high-res scans, but was saved in LJPEG format which is an archaic format which is no longer maintained. The CBIS-DDSM [7] is a curated subset of the DDSM data which is saved as DCM files. The full dataset is about 163 GB in DiCom format, but can be significantly reduced by conversion to JPEGs.

## Analysis and Methods

### Standard Machine Learning Techniques

As the image data has proved difficult to work with I will begin by analysing the WDBC data [6] using standard machine learning classifiers. This data contains three separate datasets:

1. The first classifies masses as benign or malignant based on 9 features extracted from FNA biopsies.
2. The second classifies also masses as benign or malignant using 10 features of cell nuclei extracted from FNA biopsies, with the mean, worst and standard error for each feature for a total of 30 features.
3. The third classifies the prognosis of masses as recurrent or non-recurrent based on the same features as 2 above.

I will perform exploratory data analysis on this data to get ideas about what distinguishes benign and malignant masses and attempt to classify this data using decision trees, SVMs and possibly other methods. The results will be analysed using ROC curves.

### ConvNet

I have already tried two approaches to classifying the images:

1. The first approach was extracting features using a pre-trained ConvNet. I attempted to do this by compressing the high res scans to 299x299 and feeding them into pre-trained VGG 19 and Inception 4 networks. I extracted the features from the flattened outputs and tried to classify them using k-nearest neighbours. The results never exceeded the most frequent baseline. I also attempted extracting features from a variety of other points in the networks, with similar results. Suspecting that ConvNets trained on object recognition might not be suited to extracting the rather subtle features of lesions, I abandoned this tactic.
2. The second approach was to train a ConvNet on the MIAS data, which proved to be too small to use. The training accuracy quickly went up to 1.0 while the validation accuracy never exceeded the most frequent baseline. The small size of the dataset is a likely factor contributing to this.

In order to train a ConvNet I clearly need more data, so will attempt to train a new network on the CBIS-DDSM dataset [7]. This dataset contains around 2,500 images in the training set and 700 images in the test set.

If this does not work, another option to create more data is to cut the high-res images into smaller sections. The DDSM dataset does not contain the location or size of the abnormalities, so this would need to be done with the MIAS data.

## Process

I have already attempted to extract features from the images using a pre-trained ConvNet and attempted to train my own ConvNet on the MIAS data with unsatisfactory results.

The next step is to analyse the UCI data using standard machine learning techniques to discover the characteristics of masses and how to distinguish benign from malignant masses. I will then attempt to create classifiers to predict the severity of a biopsied mass. The same analysis will be performed on the data which classifies masses as recurrent or non-recurrent.

Once this is complete the next step will be to attempt to train a ConvNet on the scans. This will require downloading the very large DDSM dataset, converting the images to JPEG, processing them to be able to label them properly, resizing them to a manageable size, and then attempting to train a ConvNet on the data. I plan to try to resize the images to 512x512, and if that size is not manageable I will try 299x299. The MIAS data may be combined with this data if deemed suitable.

Should this not succeed the backup plan is to create another dataset of 100x100 images, with the positive images centred on the abnormality, and the negative images segmented randomly. The size of 100x100 is big enough to contain the majority of the abnormalities in the MIAS data. As the DDSM data does not include the centre of the masses it cannot be used for that purpose.

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 consist of the results of the analysis of the Wisconsin data, along with a description of the process of creating and training the ConvNet on the image data.

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>
6. Wisconsin Diagnostic Breast Cancer (WDBC) <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>
7. Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi , Daniel Rubin (2016). Curated Breast Imaging Subset of DDSM. The Cancer Imaging Archive. <http://dx.doi.org/10.7937/K9/TCIA.2016.7O02S9CY>
8. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, Journal of Digital Imaging, Volume 26, Number 6, December, 2013, pp 1045-1057.
9. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
10. William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.
11. O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and YuyingLi, editors, SIAM Publications, Philadelphia 1990, pp 22-30.
12. K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).