Automatic Segmentation and Diagnosis of Mass Lesions in Mammograms

Margaret Shen (marshen), William Du (willadu)

Introduction

According to the National Breast Cancer Foundation, one in eight women in the U.S. will be diagnosed with breast cancer and over 40,000 of them will die from the disease every year. Only 5% of women who are diagnosed at the latest stage of cancer will survive past 5 years[1]. Thus, an important step to reducing mortality rates for breast cancer patients is detection of cancer at an early stage. This includes more common and more accurate screening. The most effective method for screening patients for breast cancer is screen film mammography. Identification of lesions in breast mammograms typically requires expert radiologists who undergo years of training. However, interpretation of mammograms is still very difficult and an error-prone task. They have a false negative rate of around 20% and a false positive rate of up to 50%, due to factors such as dense breast tissue[2].

Computer-aided detection (CADe) systems help identify areas in mammograms that may be cancerous that the radiologist may have missed. Previous literature has shown rapid improvement in automatic detection for microcalcifications[3], another symptom of breast cancer, yet progress in detection of mass lesions has remained stagnant. Masses are often more difficult to detect because they are typically indistinguishable from dense breast tissue, can have low image contrast, and may be connected to surrounding breast tissue[4]. In this paper, we will explore the task of segmentation, a crucial part of a CADe system, which provides a high resolution outline of a cancerous region (region of interest, ROI). Using convolved and filtered images, we build a classification model that outputs whether or not a voxel is part of a mass lesion. Using these classifications, we generate a candidate ROI for each image. Computer-aided diagnosis (CADx) systems usually are found in tandem to CADe systems[5]. Computer aided diagnosis systems help classify lesions as benign or malignant. In this paper, we extract textural characteristics and tumor characteristics from ROIs to classify lesions as benign or malignant. We explore CADe and CADx simultaneously using machine learning to produce a tool that radiologists may use to inform their decision process and ultimately provide more accurate diagnoses.

Data & Methods

Data

The dataset provided by Dr. Daniel Rubin (MD), assistant professor of radiology, contains 1992 mammograms from the Digital Database for Screening Mammography (DDSM) Dataset. 996 of these mammograms are from the Cranial-Caudal (CC) view, the view of the breast from above. The remaining 1026 are from the medio-lateral view (MLO) view, a side-angle view. We exclusively use images taken from the CC view to test and train our automatic segmentation methods. Both MLO and CC features are used to test and train the diagnosis model for mass lesions. For all mammograms, we have lesion ROIs annotated by 3rd party state-of-the-art segmentation software. 114 gold standard radiologist ROIs were provided for 58 MLO and 56 CC images.

Computer Aided Segmentation

Each mammogram has an associated detection ROI, annotated by radiologists, which is a general circle placed around a region that contains a mass lesion. We use this circle as a bounding box and crop our image to that general region. In this case the tumor has been detected, thus our associated problem is segmentation of the tumor within this region. We preprocess the image to enhance contrast before extracting features. For this segmentation problem, we exclusively use the subset of CC images that have radiologist gold-standard ROIs available (n=56).

We approach the segmentation problem as a classification problem by treating each voxel of a mammogram independently and associating it with a lesion or non-lesion label. Although we make the assumption that each of the voxels in an image are independent, a high-bias assumption, it significantly expands our sample space from n=56 patients to n=45 million training voxels and lowers variance. Textural, edge, and simple features are generated for each voxel.

Extracting Features for Detection

For each cropped image, we generate 28 features for each voxel in the image by creating 28 different filtered and convolved images from the original. The features are enumerated as follows:

5 basic features represent local data associated with a voxel and/or its neighbors. These features are intensity, local entropy, local standard deviation, local range, and local average, where the latter four are statistics calculated from the 9-by-9 neighborhood around the image that help describe textural features around a voxel.

6 difference-of-Gaussian filters are applied as features that detect edges and remove noise.

16 permutations of Gabor filters are used to detect edges with respect to orientation.

Finally, an additional filter, composed of average intensity over superpixels was calculated. Superpixels are generated by clustering voxels that are similar in intensity and spatially adjacent[6].

These features were selected for their ability to help define edges, characterize local texture indicative of mass lesions, and most importantly, incorporate local information into each observation of a voxel. Thus, even as the classifier is trained on each voxel, information from the neighborhood of each voxel is incorporated as its features.

Training Classifiers and Testing for Detection Models:

After feature extraction, each voxel is associated with 28 features and a label that indicates whether it is part of a mass lesion or not. Using these voxels as individual samples we train classifiers that takes in these 28 features and outputs a label representing if the voxel is part of a lesion. The classifiers we focus on in this paper are decision trees, and Naive Bayes.

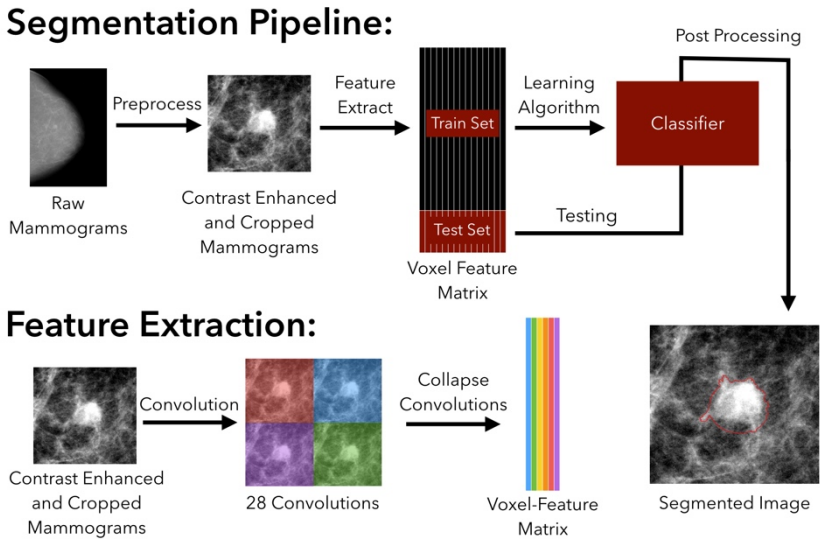

Figure 1. Main methodology for learning pipeline in the supervised segmentation pipeline. Images are preprocessed then cropped. Then features are extracted by using image convolution. A visualization of the feature extraction pipeline follows. The image is then collapsed into a voxel feature matrix and then fed into the learning algorithm.

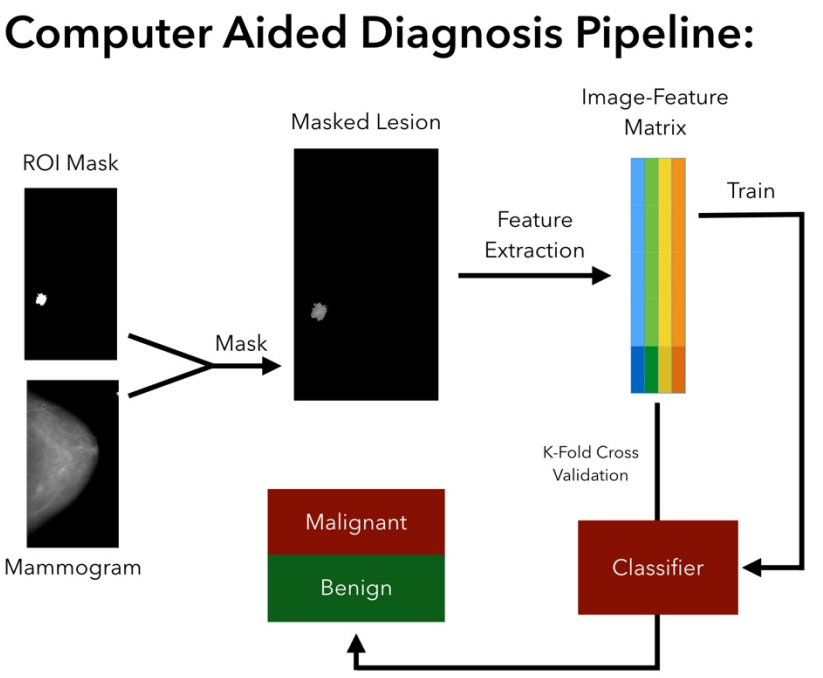

Figure 2. Visualization of methodology for CADx. The ROI is first isolated from the raw mammogram, then image features are combined with annotated features and fed into a learning algorithm for classification. Results are verified with cross-validation.

Logistic regression and linear discriminant analysis were tested but did not produce meaningful segmentations compared to the aforementioned algorithms. Although the dataset itself is large, validating and training a segmentation model is reliant on radiologist ROIS. Only 56 CC mammograms are annotated with radiologist ROIS, thus we choose to perform 3-fold validation on this image set. However, while the number of images that available is relatively small, the classifier is trained on voxels, making the number of samples used as input high. For each fold, approximately 15 million voxels are used as training samples.

Post Processing ROIs:

The classifier produces a 1-dimensional representation of a binary mask that can be rehydrated into a 2 dimensional binary mask. To retrieve the ROI of the tumor from the mask, we consider the largest connected component within the binary mask as the tumor volume, where connections are defined as the 4-connected neighborhood of a voxel. Holes and gaps within the mask are filled. All other connected components are discarded.

Computer Aided Diagnosis

Given a mammogram and the segmented area for the tumor, the goal is to classify the tumor as either malignant or benign. The segmented area was obtained with a third-party segmenting software and not our own implementation so that we were training the classification model on a more reliable image. The total number of samples used for classification was m = 841, which included all samples which had both a CC view and MLO view, as well as corresponding mask ROIs for each. Because the number of samples which were missing the MLO view was less than 10% of the total sample size, we decided to remove those samples instead of imputing the relevant feature values and risk dampening the value of the MLO image features. Each sample was composed of 83 total features: 19 CC view image features, 19 MLO view image features, and 45 binarized radiologist-annotated features.

Image Features:

Image features were chosen based on literature review. These features fall under three main categories:

- **Segment shape properties:** eg. tumor volume, tumor length/width
- **Voxel properties:** eg. max intensity (voxel value), min intensity, average intensity
- **Texture properties:** eg. entropy, contrast, homogeneity

Image texture properties were largely derived from Haralick's oft-cited "Textural Features for Image Classification." This set of features is widely used in any medical imaging application involving supervised learning and classification7. Each value was calculated only for the segmented region, which was isolated by applying a bitmask with the region outline over the original mammogram image. Segment shape properties were normalized by the bounding box area for the ROI.

Annotated Features:

Radiologist-annotated features included the following:

- **Shape of the mass lesion:** 22 categories (e.g. lobulated, irregular, asymmetric)
- **Marginal characteristics of the mass legion:** 20 categories (e.g. circumscribed, microlobulated, spiculated)

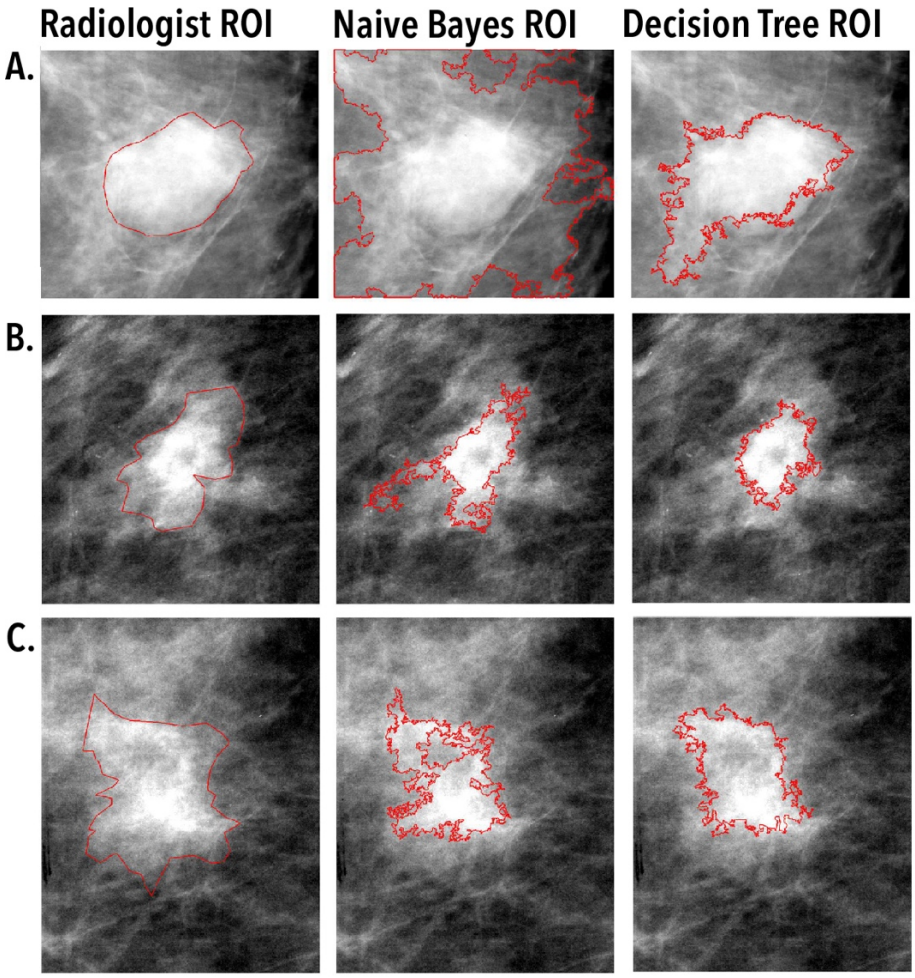

Figure 3. Sample segmentations produced by the models learned on convolutions of mammograms. The first column are the gold standard segmentations produced by radiologists, the second column are segmentations produced by the naive Bayes model, the third are segmentations produced by decision tree. The dice scores are as follows (Naïve Bayes/Decision Tree): A: (0.42/0.72), B: (0.58, 0.54) C: (0.54/0.59).

- **Subjective visibility of the mass**
- **Relative density of the surrounding breast tissue**

Because the first two feature types were qualitative and had no meaningful analog to ordinal rankings, each separate category was treated as its own binarized feature.

A variety of classification methods were tested and tuned using k-fold cross validation, including logistic regression, k-nearest neighbors, decision tree, random forest, and Naive Bayes. All features were normalized on a 0 to 1 scale while testing various classification methods, but original values were used when tuning the final chosen model, random forest, since it is invariant to monotonic transformations of feature values.

For the random forest model, we used the tree bagging algorithm with the random selection of a feature subset at each split in the learning process. We tuned the following hyperparameters: number of trees to use, number of features to samples at each split. Mean decrease impurity was used to determine the most important contributing features to the classifier. In addition, we also trained the model on various subsets of the features to determine which broad categories of features would be most useful to radiologists.

Results

Evaluation of Machine Learning Models for Computer Aided Segmentation

ROIs generated by our approach were evaluated using the Sorensen-Dice Coefficient against existing radiologist gold-standard ROIs.

We tested both models, Naive Bayes and decision tree, using 3-fold cross validation against the radiologist gold standard ROIs. The mean dice score over the 3-fold cross validation was, 0.38 for the Naive Bayes model and 0.50 for the decision tree model. The max dice score for each model was 0.68 and 0.90 respectively. And the minimum score for each model was 0, indicated that there were some images that were completely misannotated. Only three images were completely misannotated out of all 56 samples. Thus, the segmentation algorithm detected at least part of the mass lesion as annotated by the radiologist in 94% of the samples. We can examine a variety of the segmentations in Figure 3.

Evaluation of Machine Learning Models for Computer Aided Diagnosis

Random forest outperformed our other classification algorithms in terms of both training error and test error. The out-of-bag error after training on all m = 841 samples and the full feature set of n = 83 was 0.1980. Below are the average test errors with 10-fold cross-validation using different combinations of features:

- All features: 0.1831
- CC Image Features and Annotated Features: 0.2046
- CC Image Features Only: 0.3817
- MLO Image Features Only: 0.4210

Tuning for the number of features to sample at each split with k-fold validation as well, the optimal value was the square root of the number of features, approximately 9. This yielded the lowest average k-fold test error with a standard t = 200 trees. Figure 5 graphs the results for tuning for the number of trees, using 9 features to sample at each split. The optimum tree number was 100.

	PREDICTED POSITIVE	PREDICTED NEGATIVE
ACTUAL POSITIVE	38	12
ACTUAL NEGATIVE	5	29

FIGURE 4: CONFUSION MATRIX FOR A RANDOM K-FOLD OF RANDOM FOREST FOR TUMOR DIAGNOSIS. 84 SAMPLES TOTAL IN THE FOLD.

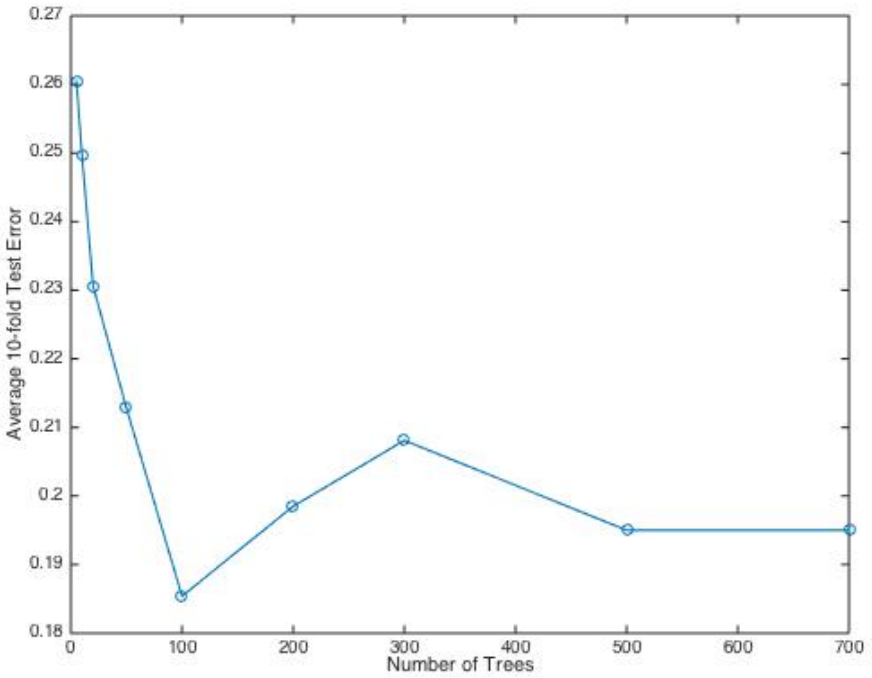

Figure 5. Tuning the number of trees for the random forest model. The lowest point in the graph (lowest error) is at t=100.

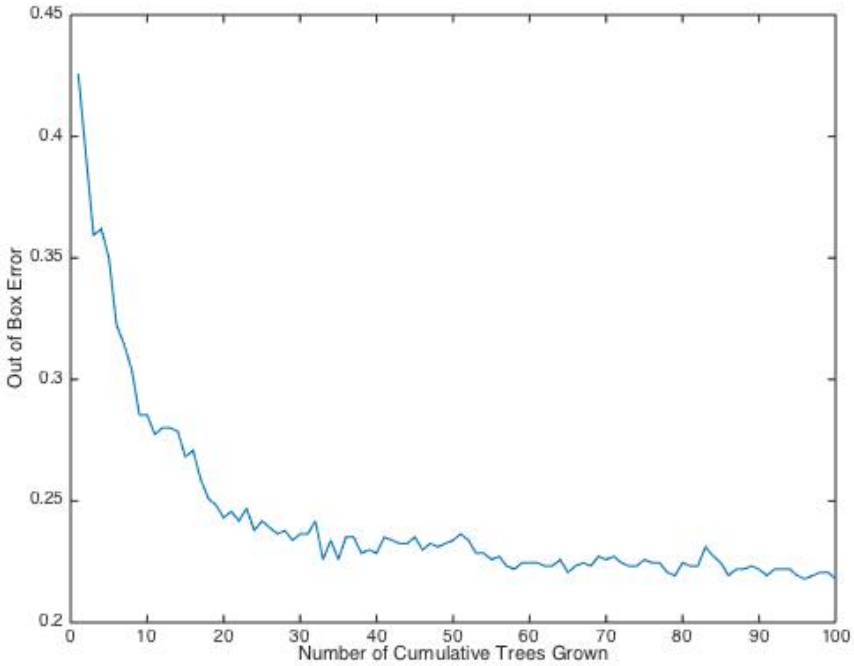

Figure 6. The out-of-bag classification error decreases then plateaus with the number of grown trees. This is for a randomly selected k-fold with the full feature set.

Top 6 determining features based upon mean decrease impurity for a random k-fold:

1. Microlobulated Mass Margin
2. Tubular Mass Shape
3. Irregular Mass Shape
4. Ill-defined Mass Margin
5. Max Intensity in the ROI of the CC view
6. Major axis length of the ellipse that has the same normalized second central moments as the ROI in the CC view

The average errors with k-fold cross validation for other algorithms is as follows (MATLAB implementation in parentheses):

1. Logistic Regression (fitglm with binomial distribution): 0.2361
2. Linear Discriminant Analysis (fitcdiscr): 0.2399
3. Naive Bayes (fitcnb): 0.2780
4. K-nearest Neighbors (fitcknn): 0.4554
5. Decision Tree (fitctree): 0.4631

Discussion and Future Work

Computer Aided Segmentation

One of the main assumptions made in the segmentation process is that voxels are independent, and are only tied by information in their feature matrices. Although this is a high-bias assumption it allows us to significantly expand our training set size from n=56 to n=45 million. By increasing the number of samples, we hope that the decrease in variance overcomes the high-bias assumption.

Evaluating the segmentations, we see that decision trees performed on average better than Naive Bayes evaluated by mean dice scores. Examining specific cases, we note that the mass lesion that are well defined by contrast are exceptionally annotated by the segmentation algorithm. Mass lesions that were adjacent to dense breast tissue, which much of the time, presents itself similar to mass lesions, were more generously segmented in comparison to the radiologist ROI. One common influence on the DICE score revolves around the specificity of the machine learning segmentation. Radiologist are more likely to provide much more general segmentations with regard to edge than to those annotated by the the machine learning segmentation, because of this, an automatic segmentation may appear to segment the correct region, but its DICE score may be lower than expected. The results show promising performance in the mass segmentation algorithm that can help radiologists interpret their mammograms and help computational biologist extract quantitative features as used in the diagnosis framework presented. Future exploration of unsupervised learning provides another avenue to approaching the segmentations of mass lesions in mammograms.

In this work, a region of interest is already identified by a radiologist and thus, the segmentation algorithm is run on images that have been cropped from the complete mammogram. A mass detection algorithm paired with the segmentation algorithm would allow for a more seamless computer aided detection process, a powerful tool that could increase the success of early breast cancer diagnosis.

Computer Aided Diagnosis

The test error of our best-performing model, random forest, was 0.1952; this was a vast improvement from our milestone error which was based on logistic regression. 402/841 samples were malignant, and 439/841 were benign, so our training set was fairly balanced to begin with. It is important to note that the image area which our algorithm inspected is based upon a third-party segmentation software, so there is no guarantee that the excised area represents the true properties of a mass. For example, one of the features is volume of the tumor (represented by a count of the voxels in the segmented region); if the image segment only contained a portion of the true tumor, or contained a large area outside of the tumor, this feature would be less reliable and informative. There was not a sufficient amount of gold-standard radiologist segmentations to use for classification.

Another metric to evaluate these results is the confusion matrix. We can see that for a randomly chosen k-fold, the false negative rate was 14.71% compared to a false positive rate of 24%. This is preferable to the opposite, since it is safer to let a patient know she may have breast cancer when she doesn't instead of letting her know she does not have cancer when she does. Since our overall accuracy is 81.69%, these high false positive/negative rates are expected.

It is interesting to note that classification with decision trees using only MLO view image features performed significantly worse than using only CC view image features. This may imply that mammograms from the MLO view obscure the mass lesion more, making it a poor tool for diagnosis. Using both annotated features and CC view features performed almost as well as using the full 83 features. The importance of the annotated features demonstrates that CADx is perhaps best used as a tool that will supplement radiologists for the time being, but not replace them.

This is also reflected in the top features, which could be useful for radiologists in determining what factors to weight when classifying tumors by hand. One important caveat to the mean decrease impurity method for selecting good features is that correlated features may be down-weighted. Thus there could be features correlated with the top 6 which are just as important, but are lower in the ranking. None of the top 6 most important features in the random forest originated from the MLO image features. A microlobulated mass margin was the most significant feature used in classifying a mass lesion as malignant. One interesting finding was that a tubular mass shape and major axis length were both in the top six features. The former is an annotated feature, and the latter is computed from the image, but a tubular mass shape implies a long major axis. This reinforces our hypothesis that our image feature extraction has the potential to replace human annotations, at least for some characteristics.

In the future, it would be interesting to explore other features that were not limited to the breast tissue. Studies have shown there are strong correlations of age, race and family history with breast cancer[1].

Another future direction would be to explore the use of convolutional neural networks (CNNs), which would directly take as input the raw image ROI. Since there are so many possible textural features and other measurements that can be taken from an image, it is very time-consuming to select these features by hand. The downside would be that the model can be difficult to interpret; it would be impossible for radiologists to learn what characteristics of a mass lesion are most important in diagnosis.

Acknowledgements

Advised by and data acquired through Daniel Rubin, MD, Assistant Professor of Radiology and Medicine. Thank you to Prof. John Duchi and the CS 229 staff for their guidance.

References

[1] Donepudi, M.S., Kondapalli K., "Breast cancer statistics and markers." National Center for Biotechnology Information. U.S. National Library of Medicine, n.d. Web. 06 June 2016.

[2] Johns, Louise E., and Sue M. Moss. "False-Positive Results in the Randomized Controlled Trial of Mammographic Screening from Age 40 ("Age" Trial)." Cancer Epidemiol Biomarkers, Nov. 2010. Web. 06 June 2016.

[3] I. El-Naqa, Yongyi Yang, M. N. Wernick, N. P. Galatsanos and R. Nishikawa, "Support vector machine learning for detection of microcalcifications in mammograms," Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on, 2002, pp. 201-204.

[4] Dheeba, J., N. Albert Singh, and S. Tamil Selvi. "Computer-aided Detection of Breast Cancer on Mammograms: A Swarm Intelligence Optimized Wavelet Neural Network Approach." Journal of Biomedical Informatics 49 (2014): 45-52. Web.

[5] Ravin, Car E. "Computer-aided Diagnosis: Breast Imaging." Carl E Ravin Advanced Imaging Laboratories. N.p., n.d. Web. 06 June 2016.

[6] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels. Technical report, EPFL, 2010.

[7] Haralick, Robert M., K. Shanmugam, and Its'hak Dinstein. "Textural Features for Image Classification." IEEE Transactions on Systems, Man, and Cybernetics IEEE Trans. Syst., Man, Cybern. 3.6 (1973): 610-21.

[8] Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. "Variable Selection Using Random Forests." Pattern Recognition Letters 31.14 (2010): 2225-236. Web.