Combining Boundaries and Ratings from Multiple Observers for Predicting Lung Nodule Characteristics

Ekarin Varutbangkul, Vesna Mitrovic, Daniela Raicu, Jacob Furst
DePaul University, Chicago, IL 60604, USA
{evarutba, vmitrovi}@students.depaul.edu, {draicu, jfurst}@cs.depaul.edu

Abstract

We use the data collected by the Lung Image Database Consortium (LIDC) for modeling the radiologists' nodule interpretations based on image content of the nodule by using decision trees. Up to 4 radiologists delineated nodule boundaries and provided ratings for nine nodule characteristics (lobulation, margin, sphericity, etc). Therefore, there can be up to 4 instances per nodule in our data set. However, to learn a good predictive model, the data set should have only one instance per nodule. In this study, we investigate several approaches to combine delineated boundaries and ratings from multiple observers. From our experimental results, we learned that the thresholded p-map analysis approach with the probability threshold Pr>=0.75 provides the best predictive accuracies for the nodule characteristics. In the long run, we expect that the predictive model will improve radiologists' efficiency and reduce inter-reader variability.

1. Introduction

Several research studies have shown that interpretation performance varies greatly among radiologists. Double reading by two or more trained human observers has been shown to improve the detection of lung cancer by a 3% - 30% increase in sensitivity [1]. Computer-aided diagnosis (CAD) systems can act as a *second reader* and assist radiologists in this task to improve the efficiency of single observer and to reduce variation among multiple observers.

In this paper, we present a framework for learning predictive models for lung nodule interpretation and investigate several ways to combine nodule boundaries and ratings from different radiologists' interpretations. The data we used in this study were collected by the NIH Lung Image Database Consortium (LIDC) [2]. In LIDC's marking process, there were 4 radiologists who delineated nodule boundaries and provided ratings for nine nodule characteristics. Therefore, there can be up to 4 instances (nodule boundaries and ratings) per nodule in our data set. However, to learn an unbiased predictive model, the data set should have only one instance per nodule.

The rest of the paper is organized as follows. We present a literature review relevant to our work in Section 2, the LIDC dataset and our proposed methodology in Section 3, the preliminary results in Section 4, and our conclusions in Section 5.

2. Related work

Our previous work [3, 4] can be considered one of the initial steps in the direction of mapping lung nodule image features to perceptual categories encoding the radiologists' knowledge for lung interpretation. While the derived mappings were significant in terms of their prediction power on the available LIDC data, the absence of a ground truth for the nodules' boundaries determined us to investigate ways to combine the various radiologists' delineations with the final goal of producing a more stable and general set of image features to be used in the mappings learning process.

Meyer et. al. [7] proposed to use probability map (p-map) analysis to measure the variability of radiologists' spatial locations for lung nodules. In their study, there were six radiologists each applying three segmentation methods (one manual method and two semiautomatic methods) to define the spatial extent of 23 different lung nodules from 16 different patients from the LIDC [2] data. The edge maps drawn by radiologists were used to construct binary nodule masks in which voxels inside the edge maps have value 1 and voxels outside the edge maps have value 0. The p-map image was computed by summing up the nodule masks from all radiologist-method

combinations divided by the number of all radiologist-method combinations (18 combinations). Therefore, a voxel in the p-map is the number of votes divided by the number of all radiologists as shown below:

$$V_n(i,j) = \sum_{m=1}^{R} B_{m,n}(i,j)/R \qquad (1)$$

$$B_{m,n}(i,j) = \begin{Bmatrix} 1 & E_{m,n}(i,j) \in D_{m,n} \\ 0 & otherwise \end{Bmatrix} \qquad (2)$$

where

$V_n(i,j)$ is a voxel in the p-map for nodule n at spatial location (i,j); $B_{m,n}(i,j)$ is a voxel in a binary mask of nodule n marked by radiologist m at spatial location (i,j); R is the number of radiologists who provided annotations; $E_{m,n}$ is a voxel at spatial location (i,j) in the edge map of nodule n marked by radiologist m; $D_{m,n}$ is a set of voxels in the edge map of nodule n marked by radiologist m that are inside the contour.

The stages of the voting method are exemplified in Figure 1. In the right most image (the p-map image), the white pixels represent pixels included by all four radiologists within their boundaries, the black pixels represent the ones included by none of the radiologists and the different gray shadows represent the pixels included by one, two or three radiologists.

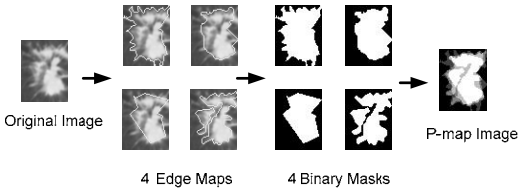

Figure 1. A diagram representing the voting method for the p-map analysis (n = 1, R = 4)

Turner et al. [8] used the voting method and Simultaneous Truth and Performance Level Estimation (STAPLE [9] implemented in the Insight Segmentation and Registration ToolKit, ITK) to extract approximate lung nodule contours (p-map) from the multiple radiologist marks and to characterize reader performance. In their study, they investigated the first 29 LIDC cases released. While the voting method is similar to the one presented in [7] and it is used in this paper as well, the STAPLE method (which is based on the Expectation Maximization (EM) algorithm) is not appropriate for our study. The algorithm considers a collection of segmentations and computes a probabilistic estimate of the true segmentation by estimating an optimal combination of the segmentations and weighting each segmentation based on the estimated performance level parameters (sensitivity and specificity) of the reader. However, we cannot employ this method in our study since we do not have the reader information in the LIDC data (all readers who marked each nodule are anonymous and their ratings are not recorded in the same order across all nodules).

Both the voting and STAPLE p-maps can be thresholded at a particular probability. Normally, the 0.5 level is used in practice. In our study, we employed thresholded p-map analysis with the thresholds at 0.25, 0.5, 0.75, and 1.0 (marked by at least one, two, three or all four radiologists). A voxel in the thresholded p-map and a voxel in the segmented image are computed by equation (3) and equation (4) as shown below.

$$T_{n,Pr}(i,j) = \begin{Bmatrix} 1 & V_n(i,j) \geq Pr \\ 0 & otherwise \end{Bmatrix} \qquad (3)$$

$$S_{n,Pr}(i,j) = \begin{Bmatrix} O_n(i,j) & T_{n,Pr}(i,j) = 1 \\ BG & T_{n,Pr}(i,j) = 0 \end{Bmatrix} \qquad (4)$$

where

$T_{n,Pr}(i,j)$ is a voxel in the thresholded p-map for nodule n with thresholded probability Pr at spatial location (i,j); $V_n(i,j)$ is a voxel in the p-map for nodule n at spatial location (i,j); $S_{n,Pr}(i,j)$ is a voxel in a segmented image of nodule n with thresholded probability Pr at spatial location (i,j); $O_n(i,j)$ is a voxel in the original image of nodule n at spatial location (i,j); and BG is a value that represents the background intensity in a DICOM image (BG = -2000).

An example of thresholded p-maps and segmented nodule images at four different probabilities (Pr) are presented in Figure 2.

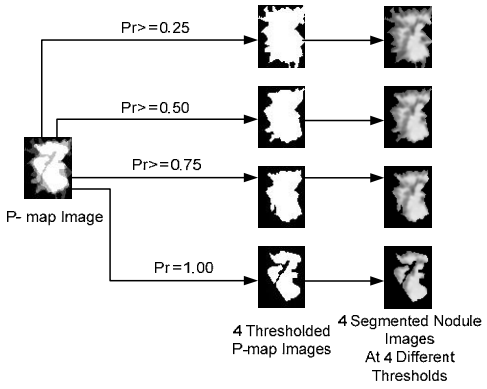

Figure 2. A diagram representing lung nodule segmentation by using thresholded p-maps at 4 different thresholds

To the best of our knowledge, the thresholded p-map analysis described above has been used as a

ground truth for evaluating the performance of segmentation algorithms but not for the evaluation of the computer-aided diagnosis (CAD) systems.

Another study that investigates other approaches for combining the nodule's boundaries is presented by Ferreira et al. in [10]. Ferreira et al. [10] proposed an algorithm that uses several morphological operations to find a "mean contour" of lung from contours manually drawn by six imagiologists to be used as the reference contour in order to evaluate the performance of their pulmonary region segmentation algorithm against the contours detected by experts. While in Ferreira's study the use of the mean contour as a reference contour was a reasonable approach given the small variation in the manually drawn contours of the lung, for our analysis the mean contour approach will not work given the high variability in the nodules' boundaries as marked by the four radiologists. Therefore, we will focus on the p-map analysis as a main approach to combine boundaries and explore different thresholded p-maps to learn which one produces the best mappings from low-level features to high level features (radiologists' characteristics).

3. Methodology

The proposed methodology (Figure 3) consists of two main stages: first, we quantify the lung nodule images using a set of low-level image features automatically extracted from the pixel data; second, we discover the mappings between the image features and radiologists' interpretations using decision trees.

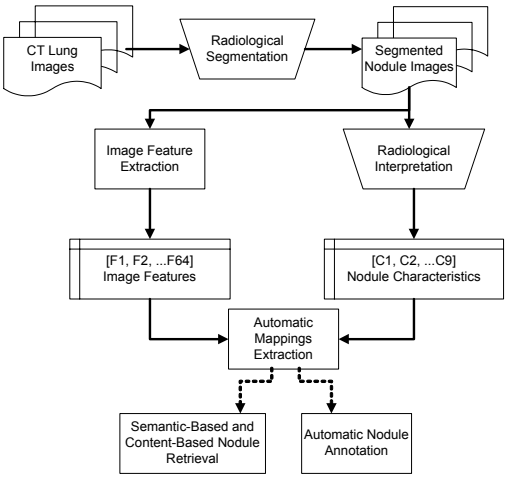

Figure 3. Methodology diagram

3.1. LIDC Dataset

At the time of performing this study, the LIDC dataset contained 85 Computed Tomography (CT) scans with associated XML files. From the available cases, we select all images that contain lesions marked as nodules > 3 mm by LIDC radiologists; for these nodules, boundaries and ratings are available as marked by at least one radiologist.

It is important to notice that the LIDC did not impose a forced consensus; rather, all of the lesions indicated by the radiologists were recorded and are available to users of the database. Therefore, there can be up to 4 different boundaries/images of a nodule marked by up to 4 radiologists on a slice as presented in Figure 4. If a nodule appears on X slices, there can be up to 4*X images for that nodule in the dataset. In this study we select only one slice per nodule (the slice with the largest nodule area) for each radiologist. Therefore, there can be up to 4 images per nodule.

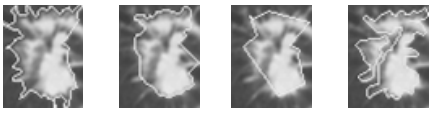

Figure 4. An example of four different delineations of a nodule on a slice marked by 4 different radiologists

From the current 85 cases available, 60 cases had 149 nodules greater than or equal to 3 mm in maximum diameter which generated 379 nodule images. From all nine semantic characteristics, we focused on the relationships between the image content and the radiologists' subjective assessments with respect to seven semantic concepts: *subtlety, lobulation, margin, sphericity, malignancy, texture, and spiculation.* Calcification and internal structure were not considered since most of the ratings for them were dominated by only one rating ('no calcification' appears in the nodule, and the internal composition of the nodule is 'soft tissue'). The description of nine nodule characteristics and their possible ratings is provided in details on our previous paper [3].

3.2. Low-level Image Feature Extraction

We propose to extract 64 image features which include four types of image content (Table 1) that encode shape, size, intensity, and texture information of the region of interest (nodule). The choice of these features was based on a literature review of the most common image features used for pulmonary nodule detection and diagnosis by existent CAD systems [5, 6]. The detailed description of these image features can be found in our previous paper [3].

Table 1. Image features

Shape Features	Size Features	Intensity Features
Circularity Roughness Elongation Compactness Eccentricity Solidity Extent RadialDistanceSD	Area ConvexArea Perimeter ConvexPerimeter EquivDiameter MajorAxisLength MinorAxisLength	MinIntensity MaxIntensity MeanIntensity SDIntensity MinIntensityBG MaxIntensityBG MeanIntensityBG SDIntensityBG IntensityDifference
Texture Features		
11 Haralick features calculated from co-occurrence matrices (Contrast, Correlation, Entropy, Energy, Homogeneity, 3rd Order Moment, Inverse variance, Sum Average, Variance, Cluster Tendency, Maximum Probability)		
24 Gabor features which are mean and standard deviation of 12 different Gabor images (orientation = 0°, 45°, 90°, 135° and frequency = 0.3, 0.4, 0.5)		
5 MRF features which are mean of 4 different response images (orientation = 0°, 45°, 90°, 135°), along with the variance response image		

3.3. Mappings between Image Features and Semantic Interpretations

The technique used for learning the mappings between image features and semantic interpretations in this study is decision tree learning. Decision tree learning [12] is a data mining technique that can be used to map the low-level representation of the data to the high-level representation of the data encoded through class or category labels. The low-level features are sorted based on some criterion that quantifies the discrimination power of the features with respect to the given classes. The tree will be formed by placing the features with the highest discriminative power at the top and the features with lowest discriminative power towards the bottom of the tree. Each internal node in the tree is a test of an attribute and branches from the node correspond to the possible values of the attribute. Therefore, leaf nodes represent classifications and branches represent conjunctions of attributes that lead to those classifications. The leaf nodes can also produce probabilistic classifications by dividing the number of cases for a certain class under the leaf node by the total number of cases grouped under the corresponding node. The complexity of the tree is a tradeoff between high accuracy for the training data and low generalizability for testing or new data.

The decision tree algorithm used in this study is C4.5 pruned tree (J48 in WEKA [13]) with the minimum objects per each leaf node being equal to 2 (best accuracies from all experiments with 2, 3, 4, and 5 objects per node) and the feature selection criterion for growing the tree being the information gain [12].

$$IG(S,A) = Entropy(S) - \sum_{v \in A} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (5)$$

where v is a value of attribute A, $|S_v|$ is the subset of instances of S where A takes the value v, $|S|$ is the number of instances, and the entropy is defined as

$$Entropy(S) = \sum_{i=1}^{C} p_i \log_2 p_i \qquad (6)$$

where p_i is the proportion of instances in the dataset that has the target attribute as i from C categories.

3.4. Experimental Design

In this subsection we present our experimental design (Figure 5) to combine the different boundaries and find the best mappings from image features to radiologists' semantic interpretations.

Although we select only one slice per nodule (the slice with the largest nodule area) for each radiologist, there are up to 4 images per nodule. The design then focuses on two main approaches: individual boundaries and aggregated boundaries.

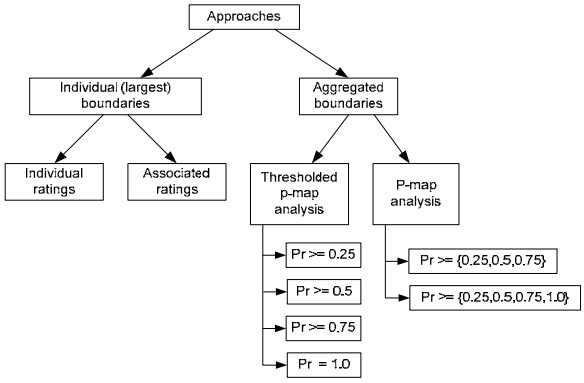

Figure 5. A diagram representing all experiments

For the individual boundaries approach, we selected the *largest nodule image* from up to 4 images of each nodule delineated by up to 4 radiologists. Then, there are two approaches (individual versus associated) to come up with the rating of characteristics for the nodule. For the individual semantic ratings, a set of ratings corresponding to the largest nodule image was selected. For the aggregated rating, we combined the ratings from up to 4 radiologists by using median value. Unlike the work done by Muramatsu et. al. [11] which used average value to combine subjective similarity ratings from several observers, the LIDC ratings are ordinal numbers and each one of them represents a semantic concept. Therefore, combining the ratings by using the average is not the appropriate method for this study.

In cases where the median rating is not an integer, we found that rounding up the median ratings provided us higher hit ratio than rounding down.

For the aggregated boundaries approach, we combined up to 4 nodule boundaries delineations by using two approaches. First, for the thresholded p-map analysis approach, a new nodule boundary was created for each nodule by combining boundaries marked by up to 4 radiologists based on a thresholded probabilities (Pr = 0.25, 0.5, 0.75, or 1.00). For example, when Pr >= 0.50 all pixels in a new created boundary are in the boundaries marked by at least 2 radiologists. In addition to combining boundaries marked by up to 4 radiologists, we also combined the ratings from the radiologists for each characteristic of each nodule by using median rating. Second, for the p-map analysis approach, we created a set of nodule images for each nodule by using the thresholded p-map analysis approach and extracted up to four sets of image features from the nodule thresholded p-map images. Then we calculated the weighted average for each image feature by using the probability thresholds as the weights. The median rating was used for this approach as well.

4. Preliminary results

First, we compared the classification accuracies from the data sets (149 images, 149 nodules, 60 cases) generated by three approaches as presented in Table 2: 1) individual boundaries & individual ratings (IB&IR); 2) individual boundaries & aggregated ratings (IB&AR); and 3) thresholded p-map (Pr >= 0.25).

From Table 2, the thresholded p-map (Pr >= 0.25 or *union boundary*) approach provided us the best accuracies compared to the other two approaches for most characteristics (except sphericity) and the average accuracy.

Table 2. Classification accuracies (hit ratio)

Characteristics	IB&IR	IB&AR	Thresholded p-map (Pr >= 0.25)
Lobulation	25.50%	30.87%	36.24%
Malignancy	42.95%	30.20%	43.62%
Margin	28.19%	31.54%	32.21%
Sphericity	41.61%	31.54%	32.21%
Spiculation	30.20%	29.53%	34.23%
Subtlety	32.21%	26.85%	40.27%
Texture	43.62%	29.53%	46.98%
Average	**34.90%**	**30.01%**	**37.97%**

A reason why the IB&IR approach provided us the best accuracy for sphericity is that the rating and the boundary are correlated with each other in a more direct way (since they are annotated and delineated by the same radiologist and the sphericity of a nodule can be interpreted from its boundary directly) than the median ratings and the union boundary (which might deviate from the concept of sphericity the radiologists had in mind at the time of interpretation).

Furthermore, we experimented with larger probability thresholds and improved even more the prediction accuracy as shown in Table 3.

Table 3. Classification accuracies (hit ratio) from the thresholded p-map analysis approach

Characteristics	Pr>=0.25 (149 images, 149 nodules, 60 cases)	Pr>=0.50 (109 images, 109 nodules, 45 cases)	Pr>=0.75 (77 images, 77 nodules, 42 cases)	Pr=1.00 (40 images, 40 nodules, 28 cases)
Lobulation	36.24%	29.09%	32.47%	2.50%
Malignancy	43.62%	52.73%	50.65%	40.00%
Margin	32.21%	35.45%	50.65%	40.00%
Sphericity	32.21%	46.36%	50.65%	50.00%
Spiculation	34.23%	32.73%	44.16%	50.00%
Subtlety	40.27%	40.00%	50.65%	60.00%
Texture	46.98%	40.91%	64.94%	65.00%
Average	**37.97%**	**39.61%**	**49.17%**	**43.93%**

From Table 3, the overall (average) accuracy improves by almost 12% when we increase the threshold. However, the accuracy drop when the threshold is 1.00, since we lose several nodules and the data set becomes smaller.

In Table 4, we combined several thresholds for p-map analysis by using the probabilities as the weights as explained in Section 3.4. With this approach we do not lose any nodule from increasing the probability threshold like in the above thresholded p-map analysis.

Table 4. Classification accuracies (hit ratio) from decision trees, p-map analysis approach for Pr >={0.25,0.5,0.75} and Pr >={0.25,0.5,0.75,1.00}

Characteristics	Pr >={0.25,0.5,0.75} (149 nodule images, 149 nodules, 60 cases)	Pr >={0.25,0.5,0.75,1.00} (149 nodule images, 149 nodules, 60 cases)
Lobulation	28.86%	24.83%
Malignancy	39.60%	44.30%
Margin	41.61%	30.87%
Sphericity	48.99%	47.65%
Spiculation	29.53%	24.16%
Subtlety	44.30%	42.28%
Texture	55.03%	53.02%
Average	**41.13%**	**38.16%**

From Table 4, the accuracy when we combine 3 thresholds (Pr >= {0.25,0.5,0.75}) is higher than the accuracy when we combine 4 thresholds (Pr >= {0.25,0.5,0.75,1.00}). Compared to the thresholded p-map approach presented in Table 3, the accuracy of this approach is higher than the accuracies for the thresholded p-map approach when the threshold is 0.25 or 0.50, but it is not as good as the accuracy of the threshold 0.75. Although the accuracy of the p-map

analysis approach (Pr >= {0.25,0.5,0.75}) is lower than the accuracy of the thresholded p-map analysis approach (Pr >= 0.75) we do not lose any nodule while we lose about half of nodules for the thresholded p-map analysis approach (Pr >= 0.75).

5. Conclusions

From our preliminary results, we found that the aggregated boundaries approach provided us higher accuracies than the individual boundaries approach. An explanation for this finding is that in the aggregated boundary approach the individual human inexactitudes can be reduced by considering an overall boundary as defined by the p-map approach.

For the thresholded p-map analysis approach we found that the overall (average) accuracy improves when we increase the threshold. An immediate disadvantage that we noticed was the decrease of number of samples (nodules) as we increase the probability value (see Table 3). An alternative to this is the p-map approach which does not decrease the number of samples but has a lower accuracy in our preliminary results.

Future work is needed to investigate if another weighted combination of image features and aggregated ratings could significantly improve the p-map analysis results versus the thresholded p-map results. In the long run, we plan to integrate the active learning approach [14] in our methodology of learning the mappings and investigate how the combined boundaries affect the new active learning approach.

6. References

[1] L.G. Queckel, R. Goei, A.G. Kessels , J.M. van Engelshoven, "Detection of lung cancer on the chest radiograph: impact on previous films, clinical information, double reading, and dual reading", *Journal of Clinical Epidemiology*, vol. 54, pp. 1146-1150, 2001.

[2] S.G. Armato, G. McLennan, M.F. McNitt-Gray, C.R. Meyer, D. Yankelevitz, D.R. Aberle, C.I. Henschke, E.A. Hoffman, E.A. Kazerooni, H. MacMahon, A.P. Reeves, B.Y. Croft, and L.P. Clarke, "Lung Image Database Consortium: Developing a resource for the medical imaging research community", *Radiology*, 232(3), pp. 739–748, 2004.

[3] D.S. Raicu, E. Varutbangkul, J.G. Cisneros, J.D. Furst, D.S. Channin, S.G. Armato III, "Semantics and Image Content Integration for Pulmonary Nodule Interpretation in Thoracic Computed Tomography", *SPIE Medical Imaging Conference*, San Diego, CA, February 2007.

[4] W. Horsthemke, E. Varutbangkul, D.S. Raicu, J.D. Furst, "Predictive Data Mining for Lung Nodule Interpretation", *In the Proceedings of the Seventh IEEE ICDM'07 Workshop on Data Mining in Medicine*, October 2007.

[5] J.M. Goo, J.W. Lee, H.J. Lee, S. Kim, J.H. Kim, J. Im, "Automated Lung Nodule Detection at Low-Dose CT: Preliminary Experience", *Korean J Radiology*, vol. 4, 211-216, 2003.

[6] B. Zhao, G. Gamsu, M.S. Ginsberg, L. Jiang, L.H. Schwartz, "Automatic Detection of Small Lung Nodules on CT Utilizing a Local Density Maximum Algorithm", *Journal of Applied Clinical Medical Physics*, 4(3), pp. 248-260, 2003.

[7] C. Meyer, T. Johnson, G. McLennan, D. Aberle, E. Kazerooni, H. MacMahon, B. Mullan, D. Yankelevitz, E. van Beek, S. Armato III, "Evaluation of Lung MDCT Nodule Annotation Across Radiologists and Methods", Academic Radiology, 13(10): 1254-1265, 2006.

[8] W.D. Turner, T.P. Kelliher, J.C. Ross, and J.V. Miller, "An Analysis of Early Studies Released by the Lung Imaging Database Consortium (LIDC)", in R. Larsen, M. Nielsen, and J. Sporring (Eds.): MICCAI 2006, *LNCS 4191*, pp. 487-494, 2006.

[9] S.K. Warfield, K.H. Zou, W.M. Wells, "Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation", *IEEE Transactions on Medical Im*aging, vol. 23, pp. 903-921, 2004.

[10] C. Ferreira, B.S. Santos, J.S. Silva, A. Silva, "Comparison of a segmentation algorithm to six expert imagiologists (sic) in detecting pulmonary contours on x-ray CT images", *Proceedings of SPIE Medical Imaging 2003*, vol. 5034, pp. 347-358, San Diego, CA, February 2003.

[11] C. Muramatsu, Q. Li, K. Suzuki, R. A. Schmidt, J. Shiraishi, G.M. Newstead, and K. Doi, "Investigation of psychophysical measure for evaluation of similar images for mammographic masses: preliminary results," *Med. Phys.* 32 (7), pp. 2295-2304, July 2005.

[12] T.M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[13] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[14] C. Korner and S. Wrobel, "Multi-class ensemble-based active learning", J. Furnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): ECML 2006, LNAI 4212, Springer-Verlag Berlin Heidelberg 2006, pp. 687-694, 2006.