Exploiting Context in Mammograms: A Hierarchical Neural Network for Detecting Microcalcifications

Paul Sajda[1], Clay D. Spence[1], John C. Pearson[1] and Robert M. Nishikawa[2]

[1]National Information Display Laboratory
David Sarnoff Research Center
Princeton, NJ 08543

[2]Kurt Rossmann Laboratories for Radiologic Image Research
Department of Radiology, The University of Chicago, Chicago IL 60637

ABSTRACT

Microcalcifications are important cues used by radiologists for early detection in breast cancer. Individually, microcalcifications are difficult to detect, and often contextual information (e.g. clustering, location relative to ducts) can be exploited to aid in their detection. We have developed an algorithm for constructing a hierarchical pyramid/neural network (HPNN) architecture to automatically learn context information for detection. To test the HPNN we first examined if the hierarchical architecture improves detection of individual microcalcifications and if context is in fact extracted by the network hierarchy. We compared the performance of our hierarchical architecture versus a single neural network receiving input from all resolutions of a feature pyramid. Receiver operator characteristic (ROC) analysis shows that the hierarchical architecture reduces false positives by a factor of two. We examined hidden units at various levels of the processing hierarchy and found what appears to be representations of ductal location. We next investigated the utility of the HPNN if integrated as part of a complete computer-aided diagnosis (CAD) system for microcalcification detection, such as that being developed at The University of Chicago. Using ROC analysis, we tested the HPNN's ability to eliminate false positive regions of interest generated by the computer, comparing its performance to to the neural network currently used in the Chicago system. The HPNN achieves an area under the ROC curve of Az = .94 and a false positive fraction of FPF=.21 @ TPF=1.0. This is in comparison to the results reported for the Chicago network; Az=.91, FPF=.43 @ TPF=1.0. These differences are statistically significant. We conclude that the HPNN algorithm is able to utilize contextual information for improving microcalcifications detection and potentially reduce the false positive rates in CAD systems.

1 INTRODUCTION

An important problem in image analysis is finding small objects in large images. The problem is challenging because 1) searching a large image is computationally expensive and 2) small objects (on the order of a few pixels in size) have relatively few distinctive features which enable them to be distinguished from non-targets. To overcome these challenges we have developed a hierarchical pyramid/neural network (HPNN) architecture which combines multi-resolution pyramid processing with neural networks. The advantages of the architecture are 1)

both neural network training and testing can be done efficiently, and at a reduced computational cost, through a coarse-to-fine paradigm and 2) such a system is capable of learning low-resolution *contextual information* to facilitate the detection of small objects.

We have previously reported on how this architecture can improve detection for problems in automatic target recognition (ATR).[1] For example, we have shown that for the problem of detecting small buildings in aerial imagery, the hierarchical neural network architecture has higher accuracy than both conventional neural network architectures and standard statistical classification techniques.

Microcalcifications are cues for breast tumors. 30% to 50% of breast carcinomas have microcalcifications visible in mammograms while 60% to 80% of all breast tumors eventually show microcalcifications via histology.[2] Similar to the building/ATR problem, microcalcifications are generally very small point-like objects ($<$ .5mm) which are hard to detect. Algorithms to assist radiologists in detecting these calcifications are being bundled in the form of Computer-Aided Diagnosis (CAD) systems. Some of these CAD systems are currently undergoing clinical evaluation.

In this paper we discuss the application of our context-based hierarchical neural network to the problem of detecting microcalcifications in digital/digitized mammograms. In particular we investigate how the HPNN might be used to reduce false alarms generated by the computer. We begin by first describing the general processing blocks that make up a typical CAD system for microcalcification detection and discuss how a neural network detector/classifier fits within the CAD paradigm. We then conduct a pilot experiment to investigate if the HPNN architecture is useful for extracting context and improving microcalcification detection. Finally we examine how our HPNN algorithms might be used to improve or enhance a CAD system through integration of multi-resolution and contextual information. We implicitly integrate the HPNN into the The University of Chicago's CAD system and test if the network reduces the overall false positive rate of the computer.

2 OVERVIEW OF CAD FOR MICROCALCIFICATION DETECTION

Microcalcification detection is a problem well-suited for computer-assisted search. Radiologists are faced with the difficult task of finding these very small, subtle cues to breast cancer, given that throughput demands they spend only about a minute evaluating a mammogram. Application of image processing techniques for aiding the radiologist have been developed in the form of Computer-Aided Diagnosis (CAD) systems. Several groups are actively pursuing the development of CAD systems for microcalcification detection.[3–6] Many of these CAD systems share a similar overall processing flow, shown in figure 1. First, a pre-processing step is used to segment the breast area and increase the overall signal-to-noise levels in the image. At this early stage, regions of interest (ROIs) are defined, representing local areas of the breast which potentially contain a cluster of calcifications. The next stage typically involves feature extraction and rule-based/heuristic analysis. Thresholds and clustering criteria are applied to the extracted features, given prior knowledge of how calcification clusters typically appear in the breast, in order to prune false positives. The remaining ROIs are processed by a statistical classifier or neural network, which has been trained to discriminate between positive and negative ROIs. The advantage of having a neural network as the last stage of the processing is that a complicated and highly nonlinear discrimination function can be constructed which might otherwise not easily be expressed as a rule-based algorithm.

The University of Chicago has been developing a CAD system for microcalcification detection which is currently undergoing prospective pre-clinical evaluation.[7] Subjective rating studies have been conducted to provide feedback from radiologists regarding the utility of the CAD system. These studies have found that the computer can often have high false positive rates, reducing the acceptability of CAD by radiologists. An important goal has been to establish methods for reducing false positive rates without sacrificing sensitivity.

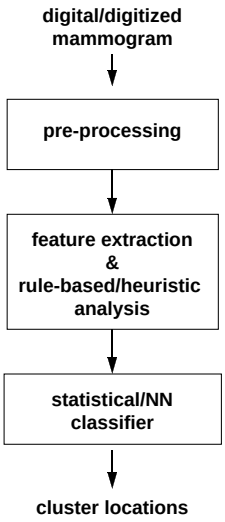


Figure 1: Generalized block diagram for a typical CAD microcalcification detection system.

We set out to test whether our HPNN architecture would be useful for reducing the false positive rate of The University of Chicago CAD system. We first conducted a small pilot experiment to see if the HPNN architecture is useful for detecting individual microcalcifications. Then, working together with The University of Chicago, we implicitly replaced the HPNN classifier with the neural network detector used in the CAD system. Using ROC analysis, we compared the ability of the HPNN for reducing false positive ROIs versus the neural network currently used in the Chicago system.

3 LEARNING CONTEXT FOR MICROCALCIFICATION DETECTION: A PILOT EXPERIMENT

3.1 General Network Architecture

The hierarchical pyramid/neural-network (HPNN) architecture is shown in figure 2. A hierarchy of networks is trained in a coarse-to-fine paradigm on a feature set derived from a pyramid decomposition of the image. Low resolution networks are first trained to detect microcalcifications, however, due to the reduced scale of the mammogram the microcalcifications are absent at this level of the pyramid. For low resolution networks to detect microcalcifications, they must learn something about large scale image structure which is associated with the location of calcifications–they must learn the context in which microcalcifications exist. To integrate context information with the featural information of the calcifications the outputs of hidden units from low resolution networks are propagated hierarchically as inputs to networks operating at higher resolutions.

In this pilot experiment, we compared three neural network architectures (see figure 3) for detecting individual microcalcifications. The input to the networks are features at two different levels of an image pyramid (levels 2 and 3, with level 0 being full-resolution) with the outputs, $p(T)$, representing the probability that a target is present at a given location in the image. Network A consists of a single net processing data from level 3 features (one-eighth of the linear extent of the original image). Network B is a hierarchical (HPNN) architecture,

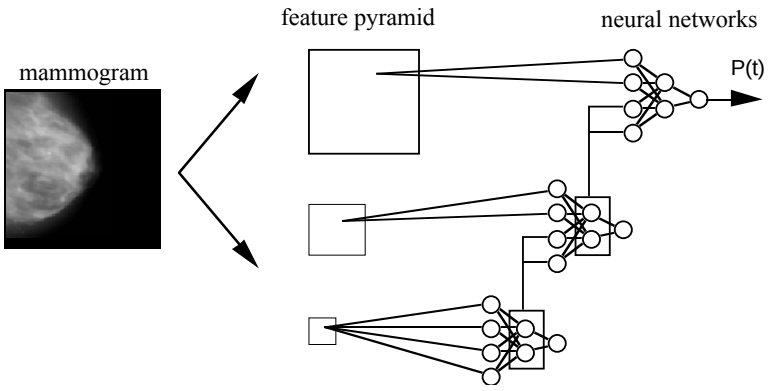


Figure 2: Hierarchical pyramid/neural network architecture for learning context. Context is propagated hierarchically via the hidden units of low resolution networks. The output of the highest resolution network is an estimate of the probability that a target is present.

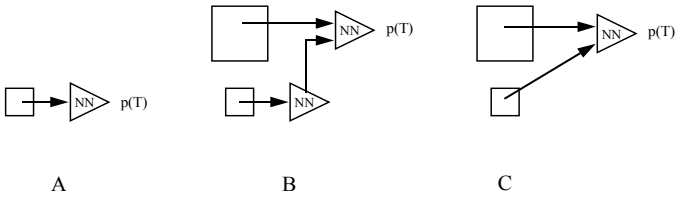


Figure 3: Three architectures that are compared in this paper.

constructed by adding to network A a second net processing level 2 data. In this architecture, information is propagated hierarchically, with the outputs of the hidden units of the level 3 net serving as inputs to the level 2 net. Network C is the "control" architecture. It is a single net that receives input directly from both level 2 and level 3.

3.2 Features

Input to the neural networks come from an integrated feature pyramid (IFP).[8] The IFP contains features constructed at several scales, allowing the network to take advantage of coarse-to-fine search and operate on only a small region of the entire image. The features in the IFP are sorted, oriented "energies" at several image scales (see Figure 4) . These are derived by constructing the Gaussian pyramid of the mammogram (Figure 4A), and applying four oriented high-pass filters to each image of the pyramid (Figure 4B). The pixel values in these images are then squared to get the energies (Figure 4C). This ensures that when the resolution is reduced by low-pass filtering, the resulting image features are present. Orientation-invariant features are constructed by sorting the energy images by their magnitude at each pixel location (Figure 4D). The resulting features are useful because the relative size of the minimum energy compared with the maximum energy indicates the degree to which the local image detail is oriented. Gaussian pyramids of these feature images are then computed (Figure 4E), with a neural network integrating the features across a given level.

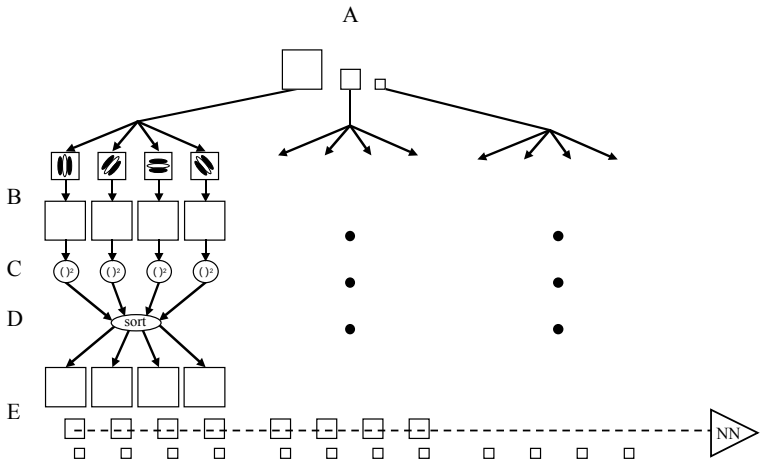


Figure 4: Integrated feature pyramid used in pilot experiment.

3.3 Specifics of the architectures and training protocol

The neural nets used in the three architectures are multi-layer perceptrons, having one hidden layer with four hidden units. All units in a network perform a weighted sum of their inputs, subtracting an offset or threshold from that sum to get the activation;

$$a = \sum_i w_i x_i - \theta \tag{1}$$

This activation is transformed into a unit's output, y, by passing it through the sigmoid function;

$$y = \sigma(a) = \frac{1}{1 + e^{-a}} \tag{2}$$

The networks were trained using a sequential quadratic programming optimization routine from the commercial Numerical Algorithms Group's library. The error function used was the cross-entropy error;

$$E = -\sum_i d_i \log y_i - (1 - d_i) \log(1 - y_i) \tag{3}$$

where $d \in \{0, 1\}$ is the desired output. To obtain the objective function for the optimization routine, we computed the total error on the training examples, adding to it a regularization term;

$$r = \frac{\lambda}{2} \sum_i w_i^2 \tag{4}$$

This type of regularization is commonly referred to as "weight decay", and is used to prevent the network from becoming "over-trained". λ was adjusted to minimize the cross-validation error. Cross-validation error was

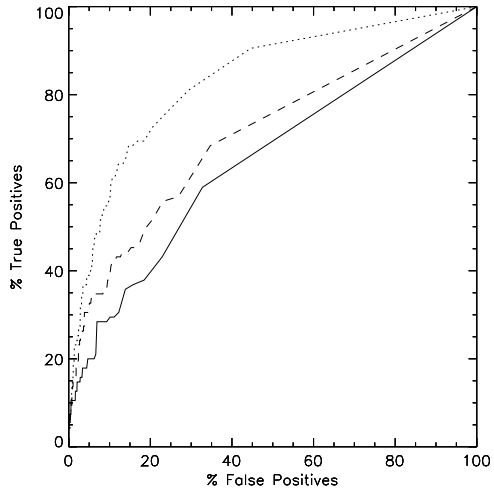


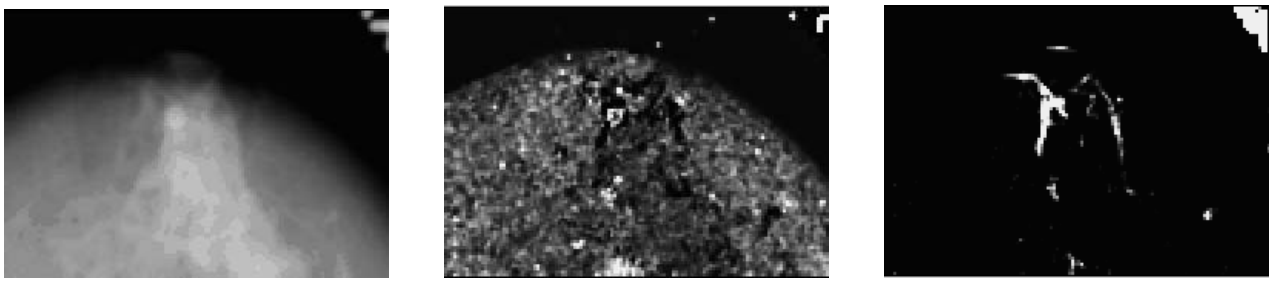
Figure 5: ROC curves for the three networks [plots are of raw true positive/false positive data]. Solid curve is for Network A, the dotted curve is for Network B (the hierarchical network with context inputs) and the dashed curve is for Network C.

computed by dividing the training data into a number of separate disjoint subsets, whose union is the entire set. The network was first trained on all of the training data, and then, starting from this set of weights, the network was retrained on the data with one of the subsets left out. The resulting network was tested on the "holdout" subset. This retraining and testing with a holdout set was repeated for each of the subsets, and the average of the errors on the subsets is the cross-validation error, an unbiased estimate of the average error on new data.

Each network receives as input a single pixel from the same location in each of the feature images at the resolution being searched (see figure 2). Network B also receives hierarchical contextual input (i.e. output of the hidden units of the level 3 net are inputs to the level 2 net). Network C receives inputs from level 2 features and also level 3 features (level 3 features are expanded to level 2 resolution). The output of each of the three networks is an estimate of the probability that a microcalcification is present at a given position, conditioned on its input. We trained the network on five mammograms, each of which had one or two clusters of about twenty microcalcifications, for a total of 97. The results given below were measured on five other mammograms with one cluster each, for a total of 95 microcalcifications.

3.4 Results

Results for the three networks are shown as ROC curves in figure 5. These are the parametric curves of true-positive rate vs. false-positive rate, where the parameter being varied is the threshold on the network's output. Notice the improvement at the higher resolution level, and especially the very large improvement when using context (hierarchical architecture of Network B). We examined whether the system was in fact taking advantage of context information by looking at the representations developed by various hidden units in the network. Figure 6 shows examples of two classes of hidden units which were found. The first class (figure 6 B) appears to be representing point-like structure, similar to the structure of an individual microcalcification. The second class of hidden unit (figure 6 C) seems to be representing something very different. In this case, the unit is selective for long, extended, and oriented structure and appears to be representing ductal location. These results suggest that this hierarchical neural network is able to automatically extract contextual information in imagery and utilize it for detecting calcifications.



A B C

Figure 6: **A** Digitized mammogram **B** One class of hidden unit representing point-like structure. **C** Second class of hidden unit representing elongated structure.

4 REDUCING FALSE POSITIVES IN CAD

Given the positive results we obtained from our pilot experiment, we next investigated the utility of the HPNN when integrated as part of The University of Chicago's CAD system for microcalcification detection. We compared the HPNN to the Chicago neural network (described in Zhang et al.[9]) in terms of reducing false positive regions of interest (ROIs). We obtained from The University of Chicago the same ROI dataset used to evaluate their network. This data set consisted of 137 ROIs, all of which their CAD system (minus the Chicago neural network) classified as positive (e.g. having a cluster of microcalcifications). The ROIs were 99x99 pixels (1 pixel = $100\mu m$ resolution), taken from mammograms scanned at 10 bits/pixel. Of these 137 ROIs, 50 were true positives and 87 were false positives. The HPNN and Chicago networks were compared in their ability to reduce false positives without sacrificing sensitivity.

4.1 Features: Steerable filters

The HPNN architecture and algorithms were slightly different from those used in the pilot study. A four level hierarchy was used (levels 0 to 3) compared to the two levels used in the pilot study. In addition, the feature set was slightly different. Before constructing the integrated feature pyramid, a background trend correction technique was applied to all the ROIs (the same procedure used for testing the Chicago network[9]). To construct the IFP, we used steerable filters[10] to compute local orientation energy. The steering properties of these filters enables the direct computation of the orientation having maximum energy. We constructed features which represent, at each pixel location, the maximum energy (energy at θ_{max}), the energy at the orientation perpendicular to θ_{max} ($\theta_{max} - 90^o$), and the energy at the diagonal (energy at $\theta_{max} - 45^o$).* Pyramids for these features were constructed and input into the network hierarchy as shown in figure 7.

4.2 Specifics of the architectures and training protocol

Each network in the HPNN hierarchy receives $3(L+1)$ inputs (L≡ level of network) from the integrated feature pyramid and 4 hidden unit inputs from the $L-1$ network, with the exception of the level 3 network, which has no hidden unit inputs.

*We found that the energy at the two diagonals was nearly identical so for compactness we use only one of the diagonals as a feature.

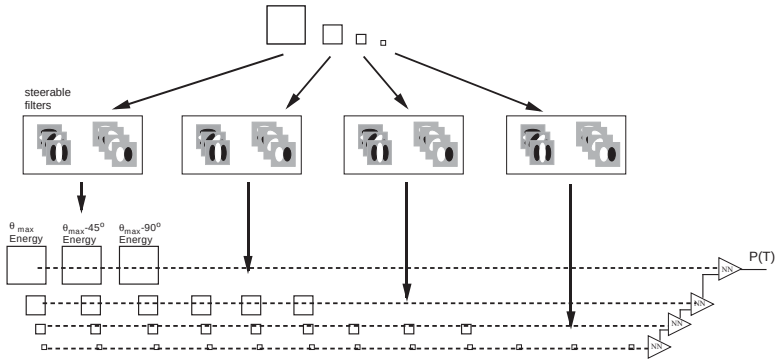


Figure 7: Integrated feature pyramid used in CAD experiment.

Network training was slightly modified from the training used in the pilot experiment. The objective function in the pilot experiment used a cross-entropy error term. In examining the truth data for the ROI data set, we found that radiologists often make small errors in localizing the individual calcifications. These errors appear to be within ±2 pixels of the correct position. To take this uncertainty in position into account, we used the following error term in our objective function;

$$E_{UOP} = -\sum_{p \in Pos} \log\left(1 - \prod_{x \in p}(1 - y(x))\right) - \sum_{x \in Neg} \log(1 - y(x)) \tag{5}$$

The first term of equation 5 is related to the probability of detecting at least one pixel in a positive region while the second term relates to no detections in a negative region.[11] The complete objective function is the sum of this error term and the regularization term in equation 4.

As in the pilot experiment, the HPNN was applied to every pixel in the input, in raster scan, and a probability map was constructed from the output of the Level 0 network.[†] This map represents the network's estimate of the probability (continuous between 0.0 and 1.0) that a microcalcification is at a given pixel location. Training and testing was done using as jackknife protocol,[12] whereby one half of the data (25 TPs and 43 FPs) was used for training and the other half for testing. Five different combinations were trained and tested.

4.3 Results

In the pilot experiment, networks were evaluated via ROC analysis for detecting individual microcalcifications. In this experiment we were instead interested in evaluating the network's ability for reducing the number of false positive ROIs. We applied the same ROC technique used by Chicago for evaluating their network.[9] For a given ROI, the probability map produced by the network is thresholded at a given value (between 0.0 and 1.0) to produce a binary detection map. Region growing is used to count the number of distinct regions. If the number of regions is greater than or equal to a certain cluster criterion, then the ROI is classified as a positive, else it is classified a negative. True positives and false positives are established and ROC curves fit using the LABROC software developed at The University of Chicago.[12]

Table 1 compares ROC results for the HPNN and Chicago network using five different cluster criterion (cc). Reported are the area under the ROC curve (A_z), the standard deviation of A_z across the subsets of the jackknife

[†]Note that for the pilot study the inputs were entire mammograms while for this study the ROI data set was the input.

cc	HPNN				Chicago NN			
	A_z	σ_{A_z}	$FPF_{\text{TPF=1.0}}$	σ_{FPF}	A_z	σ_{A_z}	$FPF_{\text{TPF=1.0}}$	σ_{FPF}
1	.93	.03	.24	.11	.88	.04	.50	.11
2	.94	.02	.21	.11	.91	.02	.43	.10
3	.94	.03	.39	.19	.91	.03	.48	.19
4	.93	.03	.48	.15	.90	.05	.56	.21
5	.93	.03	.51	.06	.88	.05	.68	.21

Table 1: Comparison of HPNN and Chicago networks.

(σ_{A_z}), the false positive fraction at a true positive fraction of 1.0 ($FPF@TPF = 1.0$) and the standard deviation of the FPF across the subsets of the jackknife (σ_{FPF}). A_z and $FPF@TPF = 1.0$ represent the averages of the subsets of the jackknife.[‡]. Note that both networks operate best when the cluster criterion is set to two. For this case the HPNN has a higher A_z than the Chicago network while also halving the false positive rate. This difference, between the two networks' A_z and FPF values, is statistically significant (z-test; $p_{A_z} = .0018$, $p_{FPF} = .00001$).

A second set of data was also tested. 150 ROIs classified as positive by the full Chicago CAD system (including the Chicago neural network) were used to test the HPNN. Though the Chicago CAD system classified all 150 ROIs as positive, only 47 were in fact positive while 103 were negatives. We applied a pre-trained HPNN (trained on the jackknife data set) to this new set of ROIs. The HPNN was able to reclassify 47/103 negatives as negative, without loss in sensitivity (no false negatives were introduced). This 46% reduction in false positive rate further corroborates the hypothesis that the HPNN algorithm has potential for significantly improving the performance of The University of Chicago's CAD system.

5 SUMMARY

In this paper we have described the application of our hierarchical pyramid/neural network architecture to the problem of finding microcalcifications in mammograms. An advantage of this network architecture is that it is designed to exploit contextual and multi-resolution information for improving detection.

We began by first conducting a pilot experiment comparing the HPNN with two other network architectures for the problem of finding individual microcalcifications in digitized mammograms. Using ROC analysis, we found that the HPNN is a better detector than its non-hierarchical counterparts, presumably due to the passing of contextual information from coarse-to-fine levels of processing. This learning of contextual information is supported by our findings that certain hidden units appear to represent information about the location of ducts, implying that the network utilizes context to increase microcalcification detection accuracy.

We then tested the HPNN by implicitly integrating it into The University of Chicago's CAD system for microcalcification detection. We compared the HPNN versus the Chicago neural network in terms of their ability for reducing the number of false positive ROIs generated by the computer. Following the same experimental

[‡]Note that we found that one of the subsets had a single positive ROI in which it was difficult to detect more than one microcalcification. This single outlier made the $FPF = 1.0$ for nearly all cluster criteria, thereby skewing the average. This was apparent in the large standard deviation we obtained for the FPF when this subset was added (for cluster criterion of 2, $\sigma_{FPF} = .33$ vs. $\sigma_{FPF} = .11$ if subset left out). Given that the Chicago network had a $\sigma_{FPF} = .10$, we decided that the subset was an outlier and therefore eliminated it from our jackknife results reported in table 1. Had we included the subset, the ROC results for cluster criterion of 2 would be $A_z = .94$, $FPF = .35$, $\sigma_{FPF} = .33$.

protocol and data set used by Chicago, we trained and tested our network and found that we could increase both the sensitivity and specificity of the computer detection scheme, potentially reducing the false positive rate by a factor of two. Finally, a second data set of 150 ROIs was tested, with findings showing the HPNN reducing the false positive rate of the complete Chicago CAD system by 46%, without loss in sensitivity.

6 ACKNOWLEDGEMENTS

This research was supported by a contract from the National Information Display Laboratory, ARPA N00014-93-C-0202, a grant from the Murray Foundation, and NIH grant CA60187.

7 REFERENCES

[1] C. Spence, P. Sajda, S. Hsu, and J. Pearson. Neural network/pyramid architectures that learn target context. In *ARPA Image Understanding Workshop*, pages 853–862, Monterey, CA, 1994.

[2] R. Mills, R. Davis, and A. Stacey. The detection and significance of calcification in the breast: A radiological and pathological study. *British Journal of Radiology*, 49:12–26, 1976.

[3] B. Fam, S. Olson, P. Winter, and F. Scholz. Algorithm for the detection of fine clustered calcifications on film mammograms. *Radiology*, 169:333, 1988.

[4] D. Davies and D. Dance. Automatic computer detection of clustered microcalcifications in digital mammograms. *Phys. Med. Biol.*, 35:1111, 1990.

[5] R. Nishikawa, M. Giger, and K. Doi. Computer-aided detection of clustered microcalcifications: An improved method for grouping detected signals. *Medical Physics*, 20:1660, 1993.

[6] H. Chan, N. Petrick, B. Sahiner, D. Wei, M. Helvie, and D. Adler. Computer-aided diagnosis in mammography. In *Radiological Society of North America*, page 461, Chicago, IL, November 1995.

[7] R. Nishikawa, R. Haldemann, M. Giger, D. Wolverton, R. Schmidt, and K. Doi. Performance of a computerized detection scheme for clustered microcalcifications on a clinical mammography workstation for computer-aided diagnosis. In *Radiological Society of North America*, page 425, Chicago, IL, November 1995.

[8] P. Burt. Smart sensing with a pyramid vision machine. *Proceedings of the IEEE*, 76:1006–1015, 1988.

[9] W. Zhang, K. Doi, M. Giger, Y. Wu, R. Nishikawa, and R. Schmidt. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. *Medical Physics*, 21(4):517–524, April 1994.

[10] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Trans. on PAMI*, 13(9):891–906, September 1991.

[11] C. Spence. Supervised learning of detection and classification tasks with uncertain training data. In *ARPA Image Understanding Workshop*, Palm Springs, CA, February 1996.

[12] C. Metz. Current problems in ROC analysis. In *Proceedings of the Chest Imaging Conference*, pages 315–336, Madison, WI, November 1988.