

Eliminating false-positive microcalcification clusters in a mammography CAD scheme using a Bayesian neural network

Darrin C. Edwards^{*a}, John Papaioannou^a, Yulei Jiang^a, Matthew A. Kupinski^b,
Robert M. Nishikawa^a

^aKurt Rossmann Laboratories for Radiologic Image Research,
Department of Radiology, The University of Chicago, Chicago IL 60637

^bDepartment of Radiology, University of Arizona, Tucson, AZ 85724

ABSTRACT

We have applied a Bayesian neural network (BNN) to the task of distinguishing between true-positive (TP) and false-positive (FP) detected clusters in a computer-aided diagnosis (CAD) scheme for detecting clustered microcalcifications in mammograms. Because BNNs can approximate ideal observer decision functions given sufficient training data, this approach should have better performance than our previous FP cluster elimination methods. Eight cluster-based features were extracted from the TP and FP clusters detected by the scheme in a training dataset of 39 mammograms. This set of features was used to train a BNN with eight input nodes, five hidden nodes, and one output node. The trained BNN was tested on the TP and FP clusters detected by our scheme in an independent testing set of 50 mammograms. The BNN output was analyzed using ROC and FROC analysis. The detection scheme with the BNN for FP cluster elimination had substantially better cluster sensitivity at low FP rates (below 0.8 FP clusters per image) than the original detection scheme without the BNN. Our preliminary research shows that a BNN can improve the performance of our scheme for detecting clusters of microcalcifications.

Keywords: CAD, mammography, clustered microcalcifications, Bayesian artificial neural networks, ideal observer approximation

1. INTRODUCTION

We have developed a computerized scheme for the detection of clustered microcalcifications that is used in an “intelligent” mammography workstation for computer-aided diagnosis (CAD).¹ The primary goal of this research is to provide radiologists additional information to assist them in making diagnoses, and to this end one of the ways we are attempting to improve the computerized scheme is by reducing its false-positive (FP) rate without greatly reducing its true-positive (TP) rate of detecting lesions. This is important because one would expect that the fewer the FP detections produced by the computerized scheme, the less the likelihood of increasing the radiologist’s callback rate without finding more cancers. Also, reducing the FP rate of the detection scheme should reduce the reading time required by a radiologist using the CAD scheme to evaluate and reject the scheme’s FP detections.

Our computerized scheme consists of five steps.¹ In the first step, the breast region is segmented from a digitized mammogram. Candidate signal locations are then determined using a difference filter and thresholding techniques. Five feature values are calculated at each candidate signal location: the first moment of the signal power spectrum, mean signal pixel value, signal edge gradient, signal area, and signal contrast. A series of rules are applied to the feature values to eliminate FP detected signals. The remaining signals are then grouped into clusters, and in the last step a shift-invariant artificial neural network (SIANN) is used to eliminate FP clusters.² A flowchart of the scheme is shown in Figure 1.

Recently we were able to improve the performance of our computerized scheme by replacing the rule-based signal feature analysis stage with a Bayesian neural network (BNN),³ as shown in the flowchart in Figure 2. A BNN uses a prior distribution model for its parameters (the neural network weights), in a Bayesian sense, to regularize the training procedure.⁴ Effectively, the training procedure maximizes the quantity

$$f_{\vec{w}}(\vec{w}|X, T) = \frac{f_{X,T}(X, T|\vec{w})f_{\vec{w}}(\vec{w})}{\int f_{X,T}(X, T|\vec{w})f_{\vec{w}}(\vec{w}) d\vec{w}}, \quad (1)$$

^{*}Correspondence: E-mail: d-edwards@uchicago.edu; Telephone: 773 834 5094; Fax: 773 702 0371.

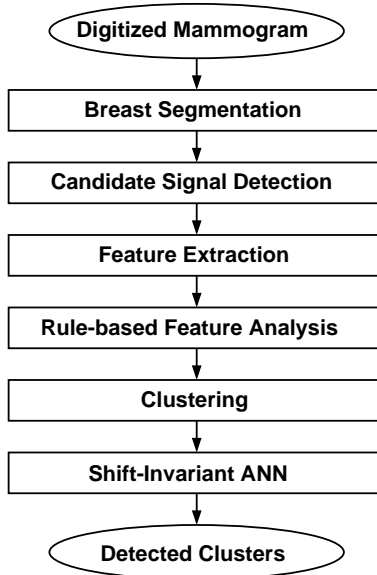


Figure 1. Flowchart of the original scheme for detecting clustered microcalcifications developed in our laboratories.

where \vec{w} are the neural network weights, X are the data observations, T are the corresponding truth states (TP or FP) of each observation, $f_{X,T}(X, T|\vec{w})$ is the probability density function of the data observations given a particular set of weight values, and the regularization term $f_{\vec{w}}(\vec{w})$ is the probability density function of the weights.^{4,5} The negative of the logarithm of the left-hand side of Eqn. 1 is the cost function which is minimized during training. The goal of regularization is to reduce the likelihood that the neural network will become “overtrained”; that is, that performance on the training set will not be representative of performance on independent testing datasets. We have found that the regularization term makes the BNN’s performance less sensitive to the number of hidden units in the neural network than a conventional ANN;⁵ furthermore, it can be shown that, given sufficient training data, a BNN can approximate the ideal observer decision function of the data population.⁵

In the particular case of distinguishing between TP and FP individual microcalcification signals, at a false-positive fraction (FPF) of detected signals of 0.30, the original scheme had a sensitivity, or true-positive fraction (TPF) of detected signals of 0.77, while the scheme with the signal BNN had a TPF of 0.89. The improvement in performance in terms of detected clusters was much less dramatic, which we argued was due to the complicated and nonlinear clustering and SIANN stages of the scheme performed after the signal BNN is applied.³

In this work we extend the previous methodology by training a BNN on cluster-based features, which were originally developed to distinguish between benign and malignant lesions.⁶ The cluster BNN is then combined with the modified detection scheme, as shown in the flowchart in Figure 3. Our training database and methodology are described in more detail in section 2. In section 3, we present the results of this method when applied to an independent testing database. We discuss these results and present our conclusions in sections 4 and 5.

2. MATERIALS AND METHOD

Our training set consisted of 39 mammograms digitized on a Fuji drum scanner at a resolution of $0.1\text{mm} \times 0.1\text{mm}$ quantized to 1024 gray levels. Our testing set consisted of 50 mammograms digitized in the same manner. At a signal BNN threshold of 0.9995, 37 out of 41 true clusters, in addition to 28 FP clusters, were detected by the modified scheme described in Figure 2. This threshold corresponded to the highest TPF achievable by the modified scheme on the testing dataset (excluding operating points with FP/image rates greater than 5), with the lowest FP/image rate at that TPF.

We extracted eight cluster-based features from each of the 37 TP clusters and 28 FP clusters detected in the training dataset. The features used were (1) the number of microcalcifications in the cluster; (2) mean microcalcification area; (3) mean effective microcalcification volume; (4) relative standard deviation of effective microcalcification

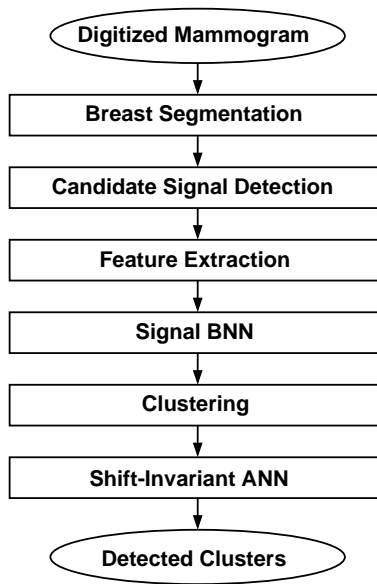


Figure 2. Flowchart of the modified scheme for detecting clustered microcalcifications, incorporating a signal BNN.

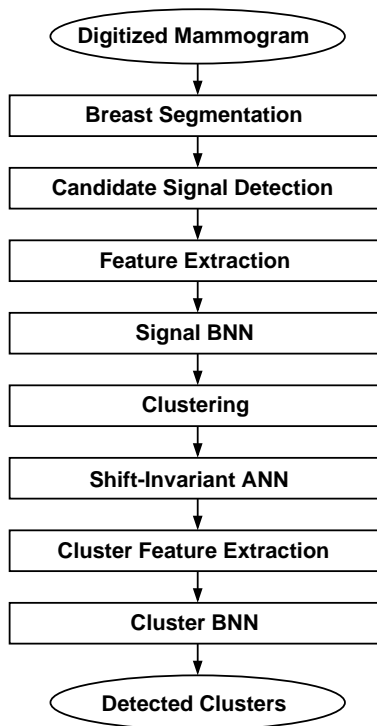


Figure 3. Flowchart of the proposed scheme for detecting clustered microcalcifications, incorporating a BNN to distinguish TP and FP clusters (in addition to the previously implemented signal BNN).

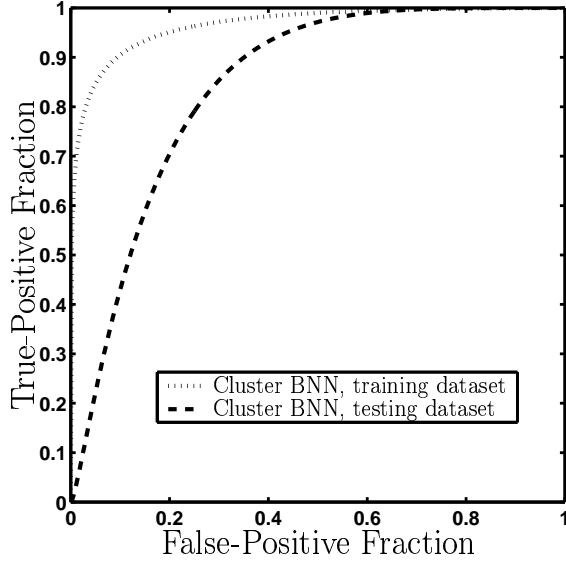


Figure 4. ROC curves of the cluster BNN, applied to the clusters initially detected in the training and testing datasets at a signal BNN threshold of 0.9995. The training dataset ROC has $A_z = 0.97$, and the testing dataset ROC has $A_z = 0.84$.

thickness; (5) relative standard deviation of effective microcalcification volume; (6) second highest microcalcification shape irregularity measure; (7) cluster area; and (8) cluster volume. These features are described in detail elsewhere.⁶

We used the extracted features to train a BNN with eight input units, five hidden units, and one output unit. We applied the trained BNN to the 36 out of 41 TP clusters, and 124 FP clusters, detected by the modified scheme in the testing dataset at the same signal BNN threshold of 0.9995. The BNN output for the testing dataset was fit to an ROC curve using the Metz LABROC4 algorithm.⁷ We then scaled the fitted ROC curve by the actual number of TP clusters and number of images in the testing dataset to evaluate the BNN output using FROC analysis.

3. RESULTS

Figure 4 shows the ROC curves produced by the the cluster BNN, when applied to the clusters initially detected in the training and testing datasets at a signal BNN threshold of 0.9995. The area under the ROC curve (A_z) for the training dataset was 0.97, and that for the testing set was 0.84.

Figure 5 shows the FROC curve produced by the modified scheme, employing only the signal BNN, for the testing dataset. This curve is generated by varying the signal BNN threshold from 0 to 1, with clustering and cluster filtering performed as in the original scheme. Also shown is the FROC curve produced by the scheme with both the signal BNN, at a threshold of 0.9995, and the cluster BNN, applied to the same testing dataset. For comparison, the FROC curve produced by the same scheme on the training dataset is also shown in this figure.

4. DISCUSSION

Figure 5 shows that below 0.8 FP clusters per image, the sensitivity of the detection scheme with both signal and cluster BNNs is strictly better than that of the detection scheme with only the signal BNN. Above 0.8 FP clusters per image, the two schemes have similar performance.

Our use of BNNs is motivated by several observations.⁵ We have found that a BNN is less prone to overtraining than a conventional artificial neural network (ANN) due to the regularization term in the BNN training cost function, making the BNN more robust to training set properties due to sampling rather than characteristics of the data population. The performance of the BNN also tends to be less sensitive to the number of hidden units chosen. We have also been able to show that, given sufficient training data, a BNN can accurately model the ideal observer decision function for the data population.

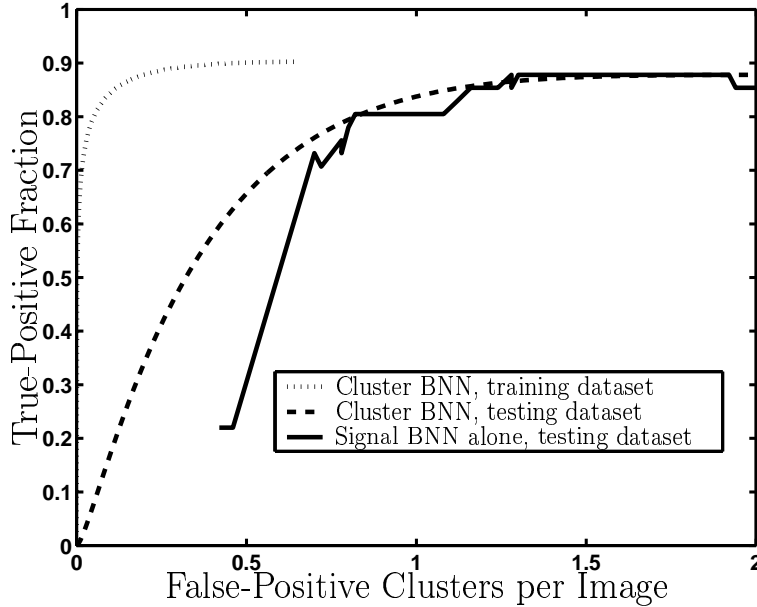


Figure 5. FROC curves of the detection scheme with only the signal BNN, and with both the signal BNN (at a threshold of 0.9995) and the cluster BNN, applied to the testing dataset of 50 images. Also shown is the FROC curve for the detection scheme with both signal and cluster BNNs, applied to the training dataset of 39 images.

Thus, although it is reassuring that the performance of the detection scheme with both signal and cluster BNNs is not significantly worse than that of the scheme with only the signal BNN in the high sensitivity regime, it would of course be preferable if the performance of the scheme with both signal and cluster BNNs were always better. We believe that this was not achieved in this case due to the small size of the training dataset.

This particular database of 39 training and 50 testing mammograms has proven adequate in the past for selecting the parameters for signal-based decision rules⁸ and the pixel-based cluster filter,² and for training the signal BNN³ as reported previously. This is because there are nearly 200 TP, and over 1000 FP, initially detected signals available for analysis in the training dataset, and similar numbers of signals in the testing dataset. Figure 6 shows the ROC curves for the signal BNN alone in the task of distinguishing TP and FP microcalcification signals on the same training and testing databases; notice that the curves are closer together (training $A_z = 0.96$, testing $A_z = 0.90$) than the cluster ROC curves shown in Figure 4 (training $A_z = 0.97$, testing $A_z = 0.84$).

As stated above, however, the number of initially detected clusters available for cluster-based feature analysis is much smaller, at 37 TP and 28 FP clusters in the training dataset, and 36 TP clusters (representing a maximum achievable testing sensitivity of 0.88) and 124 FP clusters in the testing set. We expect that training and testing the cluster BNN on a larger dataset, which we are currently performing in our laboratory, will yield results more consistent with the success attained in discriminating between TP and FP microcalcification signals. This expectation is based largely on the strong relationship other researchers have found between training dataset size and the difference between classifier performance measures in training and testing datasets.^{9,10}

It should also be noted that the maximum sensitivity achievable by the cluster BNN, as we are currently employing it, is limited by the initial detection of clusters, performed in our scheme by clustering detected microcalcification signals and then filtering out FP clusters with the SIANN. It may be possible eventually to optimize these stages; as currently implemented, this would require a search in a very large parameter space, and we have had to choose the operating parameters of these stages of our detection scheme subjectively.

5. CONCLUSIONS

We have found BNNs to have many theoretical and practical advantages over previous discrimination methods such as rule-based thresholding and conventionally trained ANNs, and we are therefore making increasing use of them in our CAD scheme development. Previously we showed that our detection scheme's performance could be improved

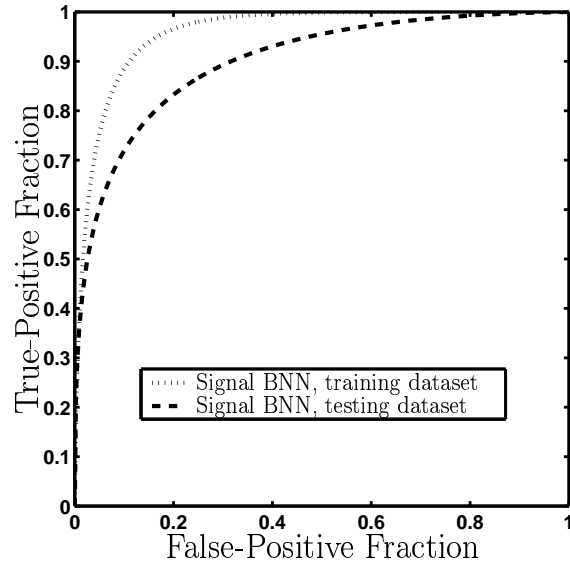


Figure 6. ROC curves of the signal BNN when applied to the initially detected microcalcification signals in the training and testing datasets. The training dataset ROC has $A_z = 0.96$, and the testing dataset ROC has $A_z = 0.90$.

by using a BNN to distinguish between TP and FP microcalcification signals. In the present work, we found that the detection scheme’s sensitivity was always better at low FP rates when the cluster BNN was added.

At higher sensitivities, the difference in performance between the schemes with and without the cluster BNN was much smaller. We believe this is due to the small effective sizes of the training and testing sets used, rather than to any theoretical deficiencies in BNNs or in our implementation.

ACKNOWLEDGMENTS

The authors thank Charles E. Metz for the use of his LABROC4 program for ROC analysis. This work was supported in part by a grant from the NIH (CA60187). Robert M. Nishikawa and John Papaioannou are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interests which would reasonably appear to be directly and significantly affected by the research activities.

REFERENCES

1. R. M. Nishikawa, Y. Jiang, M. L. Giger, R. A. Schmidt, C. J. Vyborny, W. Zhang, J. Papaioannou, U. Bick, R. Nagel, and K. Doi, “Performance of automated CAD schemes for the detection and classification of clustered microcalcifications,” in *Digital Mammography*, A. G. Gale *et al.*, ed., pp. 13–20, Elsevier Science, 1994.
2. W. Zhang, K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, “An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms,” *Med. Phys.* **23**, pp. 595–601, 1996.
3. D. C. Edwards, M. A. Kupinski, R. H. Nagel, R. M. Nishikawa, and J. Papaioannou, “Using a Bayesian neural network to optimally eliminate false-positive microcalcification detections in a CAD scheme,” in *Digital Mammography*, Medical Physics Publishing, 2000. (in press).
4. D. MacKay, *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, Pasadena, California, 1992.
5. M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, “Ideal observer approximation using Bayesian classification neural networks,” *IEEE Trans. Med. Imag.*, 2000. (in review).
6. Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, “Malignant and benign clustered microcalcifications: Automated feature analysis and classification,” *Radiology* **198**, pp. 671–678, 1996.

7. C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Statist. Med.* **17**, pp. 1033–1053, 1998.
8. R. H. Nagel, R. M. Nishikawa, J. Papaioannou, and K. Doi, "Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms," *Med. Phys.* **25**, pp. 1502–1506, 1998.
9. R. F. Wagner, J. T. Mossoba, H.-P. Chan, B. Sahiner, and N. Petrick, "Finite-sample dependence of classifier assessment in Computer-Aided Diagnosis," in *Computer-Aided Diagnosis in Medical Imaging*, K. Doi, H. MacMahon, M. L. Giger, and K. R. Hoffmann, eds., pp. 555–560, Elsevier Science, 1999.
10. H.-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," *Med. Phys.* **26**, pp. 2654–2668, 1999.