

# Using a Bayesian Neural Network to Optimally Eliminate False-Positive Microcalcification Detections in a CAD Scheme

D.C. Edwards, M.A. Kupinski, R. Nagel, R.M. Nishikawa, J. Papaioannou

Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology,

The University of Chicago, 5841 S. Maryland Ave., MC 2026, Chicago, IL 60637, USA

We compared the performance of a Bayesian neural network (BNN) for feature classification with a rule-based classifier and a conventional artificial neural network (ANN) in a computer-aided diagnosis (CAD) scheme for the detection of clustered microcalcifications. Five features were extracted from the images at each signal location. A BNN, which can approximate the behavior of the ideal observer, was trained on a database of 39 mammograms containing clustered microcalcifications. The performance of the trained BNN on an independent database of 50 mammograms was compared to the performance of a combined rule-based and conventional-ANN method. For both methods, detected signals were clustered to yield detected cluster FROC curves. At a true-positive fraction of detected clusters of 0.83, the number of false-positive clusters per image was 0.8 for the combined method, and 1.16 for the BNN. The BNN does not require subjective selection of thresholds as in the rule-based and combined methods, its performance is robust to the properties of the testing dataset, and it is able in theory to approximate the performance of the ideal observer.

## 1. Introduction

We are developing a computerized scheme for the detection of clustered microcalcifications to be used in an “intelligent” mammography workstation for computer-aided diagnosis (CAD) (Nishikawa et al. 1994). 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.

Our computerized scheme consists of five steps (Nishikawa et al., 1994). 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. Feature values are calculated at each candidate signal location, and a series of rules are applied to the features to eliminate FP detected signals. The remaining signals are then grouped into clusters, and in the last step a shift-invariant neural network (SIANN) is used to eliminate FP clusters (Zhang et al., 1996).

The use of rule-based methods to combine feature values for retaining or eliminating candidate signals has two drawbacks. First, the rules are determined subjectively; in our laboratory this is accomplished by plotting the feature values of signals in a training set, and setting thresholds on those feature values which eliminate as many FP detected signals as possible without eliminating more than 2% of the TP signals (Nagel et al., 1998). Second, the relatively large number of parameters involved (generally two or three for each rule) makes optimization of the scheme's performance difficult. Recent work from our group therefore explored the use of an artificial neural network (ANN) either alone or in combination with a set of rules to eliminate candidate signals. At a sensitivity of 83%, the combined method yielded 0.8 FPs per image, significantly better than the 1.9 FPs per image of the original rule-based method (Nagel et al., 1998).

The combined method, however, still uses subjectively determined rules prior to processing the feature values with the ANN. Also, conventionally trained ANNs such as the one used by the combined method can be “overtrained” when trained on small datasets. That is, the ANN mapping function will perform very well on the training set, but this performance will not be robust for independent testing datasets because the training set’s properties do not adequately reflect those of the data population. To address this, many researchers are investigating means of regularizing ANN training to reduce the possibility of overtraining. We have investigated the use of Bayesian neural networks for feature classification (Kupinski et al., 2000). A BNN uses a prior distribution model for its parameters (the neural network weights), in a Bayesian sense, to regularize the training procedure (McKay, 1992). The regularization procedure has the added benefit in practice of making the BNN’s performance less sensitive to the number of hidden units in the neural net (Kupinski et al., 2000); for conventionally trained ANNs, this is a parameter which must be selected with some care (Nagel et al., 1998). Furthermore, it can be shown that a BNN mapping function can approximate the ideal observer decision function for a given population of feature data (Kupinski et al., 2000); although the same is true for a conventional ANN in the limit of infinite data (Ruck et al., 1990), the BNN can be expected to yield a more accurate (less overtrained) approximation to the ideal observer decision function for finite datasets. We trained and tested a BNN on the same datasets used previously, and compared the performance of the BNN with the combined rule-based and ANN method reported in (Nagel et al., 1998).

## **2. Materials and Method**

Our training set consisted of 39 mammograms digitized on a Fuji drum scanner at a resolution of 0.1mm x 0.1mm quantized to 1024 gray levels. Our testing set consisted of 50 mammograms digitized in the same manner. For each image in the training set, a set of five calculated feature

values at each candidate signal location was available, as was the TP or FP state of that signal. The five features used were signal area, signal contrast, first moment of the signal power spectrum, mean pixel value, and edge gradient.

A BNN with five inputs corresponding to these five features, ten hidden nodes, and one output node was trained on the signals in the 39 training images. The trained BNN was then applied to the signals in the independent set of 50 images. To compare this result with previous work, the signals were grouped into two sets: the “obvious” signals (those eliminated by the rule-based method) and the “overlap” signals (those not eliminated by the rule-based method, which must be classified by the conventional ANN in the combined method) (Nagel et al., 1998). The performance of the BNN on each of the two groups was measured separately for comparison with the corresponding stage of the combined method.

To evaluate the performance of the BNN in terms of detected clusters of microcalcifications, the BNN was incorporated into our computerized CAD scheme. The parameters of the scheme not related to feature analysis, namely the initial thresholding, clustering, and SIANN parameters, were set to the values used to obtain the signal candidates for training the conventional networks, and for measuring the FROC curves reported in (Nagel et al., 1998).

### **3. Results**

The ROC curve showing the performance of the BNN on the signals in the independent testing dataset of 50 images is shown in figure 1. The area under this ROC curve is 0.90. Also shown in that figure is the operating point corresponding to the rule-based component of the combined method.

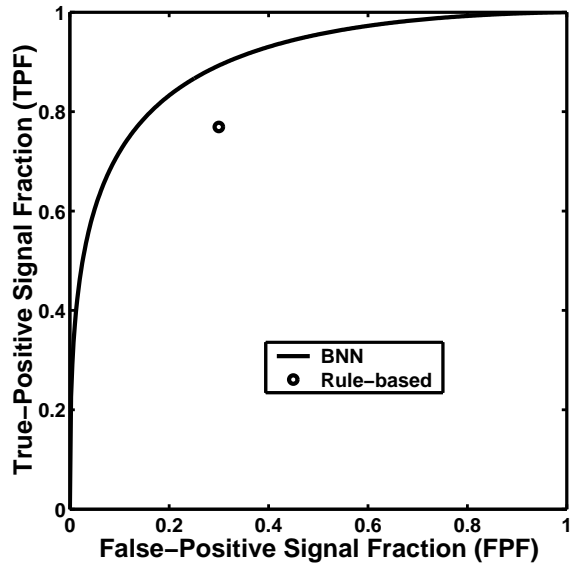


Figure 1: ROC curve for the BNN on all signals in the independent testing dataset (area 0.90), and the operating point of the rule-based component of the combined method for the same dataset.

Figure 2 shows the performance of the BNN on the “overlap” subset of signals in the testing database, those signals not eliminated by the rule-based component of the combined method. The area under this ROC curve is 0.88. For comparison, we also show the ROC curve for the conventional ANN of the combined method, trained only on the “overlap” signals of the training dataset (area 0.85), and the conventional ANN trained on all signals in the training dataset (area 0.64) (Nagel et al., 1998).

Figure 3 shows the performance of the BNN on only the “obvious” subset of signals in the testing database, those signals which would be eliminated by the rule-based component of the combined method. The area under this ROC curve is 0.89. We emphasize that the rule-based method would eliminate all these signals, yielding an effective operating point of (0,0).

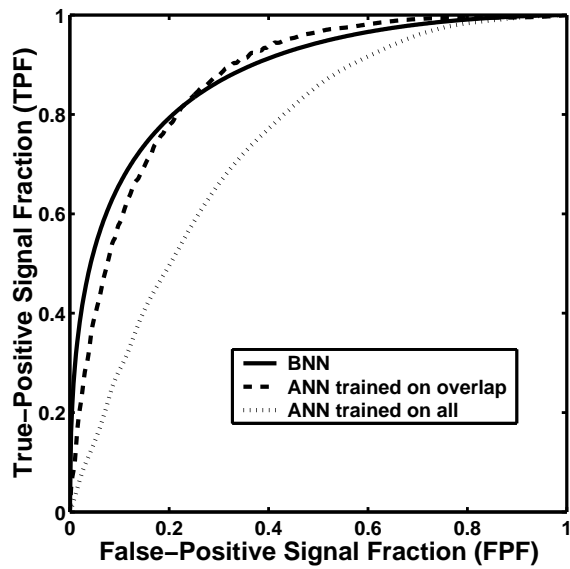


Figure 2: ROC curves for the conventional ANN component of the combined method trained only on “overlap” signals (area 0.85); a conventional ANN trained on all signals (area 0.64); and the BNN trained on all signals (area 0.88). All three curves represent performance on only the “overlap” signals in the independent testing dataset (those not eliminated by the rule-based component of the combined method).

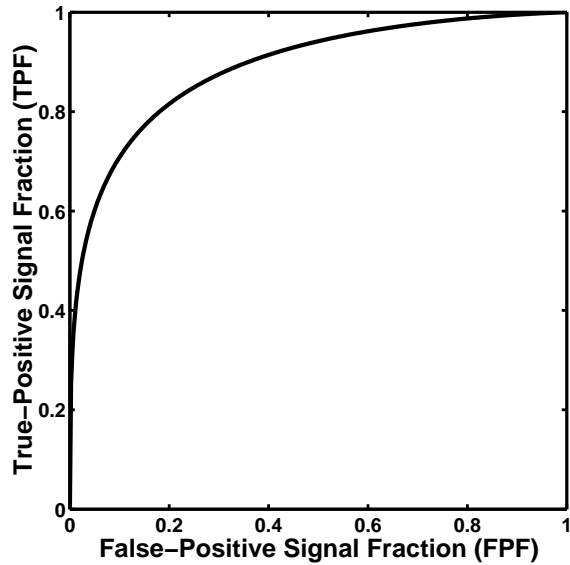


Figure 3: ROC curve (area 0.89) for the BNN on the “obvious” signals in the independent testing dataset (those eliminated by the rule-based component of the combined method).

The preceding results are all given in terms of detected signals, whereas our computerized scheme is designed to detect clusters of microcalcifications. The FROC curve of figure 4 shows the performance of the BNN and the combined method in terms of detected clusters.

#### 4. Discussion

Figures 1 through 3 show that the performance of the BNN is fairly robust to the nature of the testing dataset. The BNN ROC curves are all qualitatively similar even though they represent performance on data subsets which are very different in nature (those signals which are readily classified as clinically significant or not based on human-designed rules, and those signals which are more difficult to so classify). In contrast, the conventional ANN used in the combined method, trained only on the “overlap” signals from the training dataset, was found to have very different performance from a conventional ANN trained on all of the signals in the training dataset; the areas

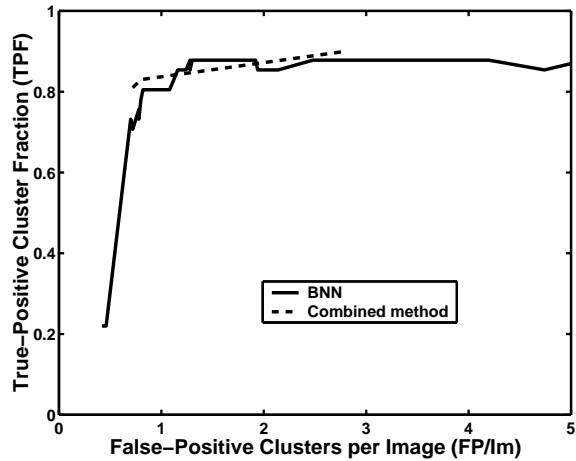


Figure 4: FROC curves for two versions of the computerized scheme for detecting clustering microcalcifications, one incorporating the BNN, and the other incorporating the combined rule-based and ANN method.

under the conventional ANN ROC curves were 0.85 and 0.64, respectively (Nagel et al., 1998).

Figure 1 implies that there is a range of BNN operating thresholds such that the BNN's performance on signals will be unambiguously better than that of the rule-based component of the combined method, because within that range the BNN will have both greater sensitivity and specificity than the rule-based component of the combined method. Figure 2 shows that the BNN's ROC curve is generally above that of the combined method's conventional ANN, at least for that range in which the BNN performance is unambiguously better than the rule-based component of the combined method. If the ROC curve for the BNN was always above that of the combined method's ANN, it would always be possible to choose an operating threshold for the BNN such that its performance on signals was better than that of the combined method. Because the ROC curves in figure 2 cross,

however, it is difficult to make such a claim unequivocally.

Figure 4 shows that for at least some values of the number of false-positive clusters per image, the sensitivity of the BNN version of the detection scheme is better than that of the combined-method detection scheme. While the gain in performance for clusters is much less impressive than that for signals, the complexity and nonlinearity of the detection scheme stages not affected by either method, namely clustering and the SIANN, leave much room for investigation of this issue.

## **5. Conclusions**

The BNN was found to have generally better performance than other methods for the task of detecting individual microcalcification signals. This is consistent with the theoretical observation that BNNs have been shown to be able to estimate the ideal observer given sufficient training data. Also, the BNN has clear advantages over subjective methods such as the subjective selection of rule-based parameters by humans. The performance of the BNN tends to be robust to the nature of the testing dataset, reducing the possibility of overtraining and the need for partitioning the training dataset.

When further processing is performed on signals detected by the BNN, such as clustering and cluster filtering, the overall performance in terms of cluster detection may not necessarily be optimal. A more careful analysis of these nonlinear cluster-related methods is required in order to make progress in this area.

## **References**

Kupinski, M.A., D.C. Edwards, M.L. Giger, and C.E. Metz. 2000. Ideal Observer Approximation Using Bayesian Classification Neural Networks. *IEEE Transactions on Medical Imaging* (in

review).

MacKay, D. 1992. Bayesian Methods for Adaptive Models. PhD thesis, California Institute of Technology, Pasadena, California.

Nagel, R.H., R.M. Nishikawa, J. Papaioannou, and K. Doi. 1998. Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms. *Med Phys* 25:1502-1506.

Nishikawa, R.M., Y. Jiang, M.L. Giger, R.A. Schmidt, C.J. Vyborny, W. Zhang, J. Papaioannou, U. Bick, R. Nagel, and K. Doi. 1994. Performance of Automated CAD Schemes for the Detection and Classification of Clustered Microcalcifications. In: Gale, A.G., et al. (eds.), *Digital Mammography '94*. Amsterdam: Elsevier Science.

Ruck, D.W., S.K. Rogers, M. Kabrisky, M.E. Oxley, and B.W. Suter. 1990. The Multilayer Perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Transactions on Neural Networks* 4:296-298.

Zhang, W., K. Doi, M.L. Giger, R.M. Nishikawa, and R.A. Schmidt. 1996. An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. *Med Phys* 23:595-601.