

**Estimating three-class ideal observer decision variables for
computerized detection and classification of mammographic mass
lesions**

Darrin C. Edwards, Li Lan, Charles E. Metz, Maryellen L. Giger, and Robert M. Nishikawa

*Department of Radiology,
The University of Chicago,
Chicago, Illinois 60637*

(Received 24 April 2003; revised 3 September 2003; accepted
for publication 17 October 2003; published 16 December 2003)

Abstract

We are using Bayesian artificial neural networks (BANNs) to classify mammographic masses in schemes for computer-aided diagnosis, and we are extending this methodology to a three-class classification task. We investigated whether a BANN can estimate ideal observer decision variables to distinguish malignant, benign, and false-positive computer detections. Five features were calculated for 63 malignant and 29 benign computer-detected mass lesions, and for 1049 false-positive computer detections, in 440 mammograms randomly divided into a training and testing set. A BANN was trained on the training set features and applied to the testing set features. We then used a known relation between three-class ideal observer decision variables and that used by a two-class ideal observer when two of three classes are grouped into one class, giving one decision variable for distinguishing malignant from non-malignant detections, and a second for distinguishing true-positive from false-positive computer detections. For comparison, we grouped the training data into two classes in the same two ways and trained two-class BANNs for these two tasks. The three-class BANN decision variables were essentially identical in performance to the specifically trained two-class BANNs, with the average difference in area under the ROC curves being less than 0.0035 and no differences in area being statistically significant. Thus, the BANN outputs obey the same theoretical relationship as do the three-class and two-class ideal observer decision variables, which is consistent with the claim that the three-class BANN output can provide good estimates of the decision variables used by a three-class ideal observer. © 2004 American Association of Physicists in Medicine. [DOI: 10.1118/1.1631912]

Keywords: Bayesian artificial neural networks, ideal observer estimation, three-class classification, computer-aided diagnosis, mammography

I. INTRODUCTION

In the past, two computerized methods for the detection of mammographic mass lesions have been investigated at the University of Chicago.¹⁻⁵ These methods involve the initial segmentation of the breast region,¹ followed by the identification of candidate lesions.²⁻⁵ These mass candidates are segmented from the surrounding breast tissue using histogram analysis² or a radial gradient index measure applied to a set of morphologically constrained candidate borders.⁶ Features including lesion area,² distance from the breast border,² circularity,³ contrast,³ perimeter irregularity,⁵ margin spiculation determined by the radial edge gradient distribution,⁷ and margin sharpness determined by the average magnitude of the gradient along the margin⁷ are calculated. The features are then subjected to rule-based analysis³ or a Bayesian artificial neural network (BANN)⁵ to retain or reject the candidate mass.

Once a radiologist has detected a structure consistent with a mass lesion, either through unaided examination of the mammogram or with the assistance of an automated scheme such as those just described, a decision of whether to perform a biopsy on the patient must be made. To aid radiologists in the task of classifying a detected mammographic mass lesion as benign or malignant, an automated scheme was developed in our laboratory.⁷⁻¹¹ The mass is first segmented from the surrounding tissue using histogram analysis of a series of grown regions.¹² Five features⁷ are then calculated for the mass: margin spiculation, determined by the radial edge gradient distribution; margin sharpness, determined by the average magnitude of the gradient along the margin; average gray level within the lesion; a normalized radial gradient measure;⁵ and a texture measure, the standard deviation of the average gradient within the mass. In the original version of this classification scheme, these features are subjected to a hybrid analysis method,⁷ in which a rule-based threshold is first applied to the most salient feature (the spiculation measure), and then a conventional artificial neural network (ANN) is applied to the remaining features. The output of the hybrid analysis method is an estimate of the likelihood of malignancy of the mass.⁷ In the version of the classification scheme currently under investigation in our laboratory, a BANN is applied to the features to estimate the likelihood of malignancy.¹³

It should be noted that the above-described schemes for detecting and classifying masses are not inherently compatible. The classification scheme is trained to distinguish between lesions which are known to be either benign or malignant, but in either case corresponding

to actual structures of clinical interest within the breast. The detection scheme on the other hand will report a set of detected lesions, some of which are true-positive (TP) detections and some of which are false-positive (FP) detections; the latter category will include detections corresponding to normal tissue structures within the breast or to various artifacts due to the imaging and digitization processes. (It is important to emphasize that these artifactual FP detections will occur regardless of whether the detection scheme is optimized so that “FP” means “not a *malignant* lesion” or “not a *clinically significant* lesion” — *i. e.*, an image structure which is neither a benign nor a malignant lesion.) The output of the detection scheme will contain objects which were by design not included in the training sample of the classification scheme, and which are perhaps not even members of the data population (benign and malignant mass breast lesions) for which the classification scheme was created. It is clear then that the detection scheme’s output cannot be used unmodified as the input to the classification scheme. This poses a problem for the development of a fully automated classification scheme for computer-aided diagnosis (CAD).

Recently, we have explored the use of BANNs for feature analysis in our automated scheme for detecting clustered microcalcifications in digitized mammograms,^{14,15} as well as for the mass lesion classification scheme as described above.¹³ A trained BANN is simply a feed-forward neural network identical in structure to a conventional ANN. The BANN differs from a conventional ANN in that the error function used in training includes a regularization term, equivalent to a Bayesian prior probability on the neural network weight values, to penalize solutions which are more complicated than the training data justify.^{16,17} The purpose of the regularization term is to reduce the likelihood that the neural network will *overtrain*, *i. e.*, that it will estimate too closely the Bayes optimal discriminant function for the finite training sample, which is nearly always a poor estimate of the population Bayes optimal discriminant function (the ideal observer). The Bayesian regularization term also obviates the need for *ad hoc* early stopping techniques based on, *e. g.*, iteration number.¹⁷ Our motivation for investigating BANNs is based, first, on our theoretical observation that, in the limit of infinite training data, a BANN can yield an ideal observer decision function for that data population (an extension of a prior result¹⁸ derived for conventional ANNs); and second, on empirical observations that even given a finite sample of training data, a BANN can estimate an ideal observer decision function reasonably well.¹⁶ Furthermore, in practical situations where the data probability density functions (PDFs) are unknown or

difficult to determine, the two-class BANN classifier has been found empirically to have better performance than other classifiers.^{14,15} We have also performed simulation studies showing that BANNs can accurately estimate ideal observer decision variables in a three-class classification task as well.¹⁹ Recent work by other researchers²⁰ used simulation studies to compare the performance of linear and conventional ANN classifiers for two-class and three-class classification tasks.

In the work presented here, we extend our BANN results to the practical case of estimating decision variables for classifying computer-detected malignant and benign mass lesions and FP computer detections (*i. e.*, the output of the automated mass lesion detection scheme) as malignant, benign, or FP computer detections. The features used for analysis were those of the existing mass classification scheme.^{5,7} We recently presented preliminary results of this methodology using a “simulated” mass data set, in which the computer-detected mass lesions were supplemented by radiologist-identified mass lesions.²¹ The goal of that work was to demonstrate the applicability of our methodology in a “proof-of-concept” study. The present work is different from that previous study and practically significant in that the mass lesions used here were all computer-identified; thus, the present work represents a further step in the development of a fully automated classification scheme for CAD.

We currently lack a fully general equivalent of receiver operating characteristic (ROC) analysis for the three-class classification task. As a preliminary step toward such an equivalent, we have developed a method for assessing the three-class BANN’s decision *variables* (as distinct from the decision *rules* which would be used to classify observations based on the values of the decision variables). This preliminary method compares the three-class BANN decision variables to two-class BANN decision variables using conventional (two-class) ROC analysis. The purpose of this comparison was to show that the BANN can generate decision variables which could be used by a classifier in a three-class classification task. Such a classifier would ultimately allow the fully automated combination of our mass lesion detection and classification schemes, and in principle, other such schemes for CAD.

II. THEORY

Given the success of applying BANNs to two-class problems,^{14–16} we are interested in determining whether the BANN methodology can be extended to a situation in which the

observational data are drawn from three classes. That is, the data are assumed to come from a distribution of the form

$$p(\vec{x}) = \sum_{i=1}^3 p(\vec{x}|\pi_i)P(\pi_i), \quad (1)$$

where $p(\vec{x}|\pi_i)$ is the conditional probability density of an observation \vec{x} being drawn from the i th class, and $P(\pi_i)$ is the *a priori* class probability of the i th class; we use boldface type to denote statistically variable quantities. The ideal observer forms a pair of likelihood ratio decision variables;²² without loss of generality, we can take these to be the two ratios

$$\text{LR}_i(\vec{x}) \equiv \frac{p(\vec{x}|\pi_i)}{p(\vec{x}|\pi_3)} \quad \{i : 1 \leq i \leq 2\}. \quad (2)$$

Decisions are then made by partitioning the likelihood ratio decision variable space into three regions in an extension of the two-class method,²² and in principle observer performance could be measured using methodology analogous to ROC analysis in the two-class case. Considerations of partitioning the decision variable space and evaluating the classifier's performance are beyond the scope of the present work, however. We are interested here only in whether a BANN implementation can in principle produce output which is related to the ideal observer decision variables, and in the quality of the BANN output as an estimate of the ideal observer decision variables in the practical case of finite sized data sets. In this section, we derive a theoretical relationship between the decision variables used by a three-class ideal observer and that used by a two-class ideal observer when two of the three classes are grouped into a single class. If the actual decision variables produced by a three-class and a two-class BANN obey this same relationship, this will be consistent with the claim that the BANN outputs represent good estimates of the ideal observer decision variables. This claim currently is supported by simulation studies¹⁶ and experiments on real data¹³⁻¹⁵ in the two-class case, but only by simulation studies¹⁹ in the three-class case.

In the two-class case it was shown¹⁶ that the output of a BANN is an estimate of the *a posteriori* class probability $P(\pi_1|\vec{x})$, which is related to the likelihood ratio decision variable *via* the monotonic transformation

$$P(\pi_1|\vec{x}) = \frac{\text{LR}(\vec{x})k}{1 + \text{LR}(\vec{x})k}, \quad (3)$$

where k is the ratio of the *a priori* class probabilities $P(\pi_1)/P(\pi_2)$. Since in the two-class case any monotonic transformation of the likelihood ratio is still an ideal observer decision

variable, the output of the two-class BANN is in this sense an estimate of the ideal observer decision variable. In the case of more than two classes, the generalization of this monotonicity property is nontrivial; however, note that we can rewrite Eq. (2) to obtain

$$\begin{aligned} \text{LR}_i(\vec{x}) &= \frac{P(\pi_i|\vec{x})p(\vec{x})/P(\pi_i)}{P(\pi_3|\vec{x})p(\vec{x})/P(\pi_3)} \quad \{i : 1 \leq i \leq 2\} \\ &= \frac{P(\pi_i|\vec{x})}{P(\pi_3|\vec{x})k_i}, \\ P(\pi_i|\vec{x}) &= \text{LR}_i(\vec{x})P(\pi_3|\vec{x})k_i, \end{aligned} \quad (4)$$

where k_i is now the i th *a priori* class probability ratio $P(\pi_i)/P(\pi_3)$.

Let \vec{y} be the output of a trained BANN for the input \vec{x} , given a set of weight values \vec{w} . Since the output of the BANN is intended to be an estimate of the two decision variables $P(\pi_i|\vec{x})$, we will also write

$$P(\pi_i|\vec{x}, \vec{w}) \equiv y_i \quad \{i : 1 \leq i \leq 2\}. \quad (5)$$

We treat \vec{w} as a random variable in the Bayesian sense, and so we have written the above expression using the standard notation for conditional probability. We have shown previously that the above-stated results for two-class BANNs also hold for three-class BANNs.¹⁹ That is, given infinite training data, the outputs y_i of a BANN of sufficiently complex architecture are the *a posteriori* probabilities $P(\pi_i|\vec{x})$; and, according to simulation studies, the outputs of the BANN provide reasonable estimates of these *a posteriori* probabilities, which are directly related to the likelihood ratio decision variables by Eq. (4).

In the present work, we have applied the BANN methodology to feature values obtained from actual mass lesions and FP computer detections, that is, real data whose PDFs are unknown. Furthermore, we currently lack the equivalent of ROC analysis for three-class classification tasks, which would allow us to compare the BANN output to a known ideal observer model as is done in the two-class case.²³ We wish to compare the output of the three-class BANN with an existing classifier whose performance is fairly well understood. The obvious candidate for such a classifier is the two-class BANN. To be concrete, we will take the three classes under consideration to be π_{mal} , the set of detections corresponding to actually malignant lesions; π_{ben} , the set of detections corresponding to actually benign lesions; and π_{FP} , the set of FP computer detections corresponding to “normal” (neither malignant nor benign) regions of the image. The detections here are assumed to be produced by a computerized scheme for CAD. Since we are explicitly restricting our attention to a set

of computer detections, the question immediately arises as to the influence of false negative (FN) lesions (*i. e.*, actually present lesions that are not identified as candidates by the computerized detection scheme), whether malignant or benign, on the overall CAD scheme under consideration, as well as the influence of the FP per image rate on such a scheme. In the appendix, we attempt to justify this focus of our attention on the classification stage of such a scheme, effectively ignoring the effects of these two factors (the FN and FP per image rates of the scheme’s initial detection stage).

As we have shown previously,¹⁹ the three-class BANN outputs, which we denote \mathbf{y}_1 and \mathbf{y}_2 , are estimates of the *a posteriori* probabilities $P(\pi_{\text{mal}}|\vec{\mathbf{x}})$ and $P(\pi_{\text{ben}}|\vec{\mathbf{x}})$, *i. e.*, the probabilities that a given observation $\vec{\mathbf{x}}$ belongs to the malignant class π_{mal} or the benign class π_{ben} , respectively. (The quantity $1 - \mathbf{y}_1 - \mathbf{y}_2$ is then an estimate of $P(\pi_{\text{FP}}|\vec{\mathbf{x}})$.)

Consider now an observer acting on the same data as defined in Eq. (1), but performing the simplified task of distinguishing the malignant (π_{mal}) detections from the nonmalignant (either benign or FP computer detections, which we will denote $\pi_{\text{ben}} \oplus \pi_{\text{FP}}$). The quantity $P(\pi_{\text{mal}}|\vec{\mathbf{x}})$ is known¹⁶ to be a monotonic transformation of the likelihood ratio for this task, and is therefore an ideal observer decision variable²² for this task. That is, the first output \mathbf{y}_1 of the three-class BANN is an estimate of the ideal observer decision variable for the simplified task of distinguishing malignant from nonmalignant detections. Note, however, that a two-class BANN can, given sufficient training data, provide a good estimate of the same ideal observer decision variable.¹⁶

Now let us consider a second simplified task in which the observer must distinguish TP detections (either malignant or benign, which we will denote $\pi_{\text{mal}} \oplus \pi_{\text{ben}}$) from FP computer detections (π_{FP}). Since the classes are by definition mutually exclusive, it immediately follows²⁴ that

$$P(\pi_{\text{mal}} \oplus \pi_{\text{ben}}|\vec{\mathbf{x}}) = P(\pi_{\text{mal}}|\vec{\mathbf{x}}) + P(\pi_{\text{ben}}|\vec{\mathbf{x}}). \quad (6)$$

Reasoning as above, this quantity is an ideal observer decision variable for the simplified task of distinguishing TP from FP computer detections. It is estimated by the sum $\mathbf{y}_1 + \mathbf{y}_2$ of the three-class BANN outputs, or by the output of a two-class BANN trained specifically for this task.

To summarize, we wish to evaluate the quality of the three-class BANN decision variables — in the sense of their ability to estimate three-class ideal observer decision variables — in the absence of a fully general equivalent of ROC analysis for the three-class classification task,

and in the absence of the “true” PDFs from which the observational data are drawn. Our method of evaluation relies on the theoretical relationship between the above-described three-class and two-class ideal observer decision variables, and on previous work demonstrating the superior performance of two-class BANNs on real data^{13–15} (consistent with the claim that the two-class BANN output provides a good estimate of the two-class ideal observer decision variable). If the BANN output for the observational data were to obey the same theoretical relationship, this would of course not constitute proof that the three-class BANN output is indeed the three-class ideal observer decision variables; such proof is, in principle, impossible in the absence of the PDFs from which the observational data are drawn. Nevertheless, such obedience would be at least consistent with the claim that the three-class BANN output represents good estimators of the three-class ideal observer decision variables.

III. MATERIALS AND METHOD

Our data set consisted of 63 actually malignant and 29 actually benign computer-detected mass lesions in a set of 440 mammograms from 153 patients, and 1049 FP computer detections in those images. Lesions were confirmed to be malignant or benign by biopsy, and FP computer detections were confirmed as such by the original radiologists’ reports. In a preliminary study,¹⁹ we applied the methodology described here to a “simulated” data set consisting of both computer-detected mass lesions and radiologist-identified mass lesions. However, the mass lesions used in the present work were all computer-identified; this more closely approximates the target of a fully automated classification scheme.

We randomly divided the detections by patient into a training and testing set, ensuring that images from a given patient did not appear in both the training and testing set. Five feature values⁷ for each detection were obtained from our mass classification scheme: margin spiculation, margin sharpness, average gray level within the lesion, normalized radial gradient index,⁵ and a texture measure. The feature values in the training set were used to train a set of BANNs with five inputs, five to ten hidden units, and two outputs. As explained in Sec. II, the outputs are $\mathbf{y}_1 = P(\pi_{\text{mal}}|\vec{\mathbf{x}}, \vec{\mathbf{w}})$ and $\mathbf{y}_2 = P(\pi_{\text{ben}}|\vec{\mathbf{x}}, \vec{\mathbf{w}})$; the quantity $1 - \mathbf{y}_1 - \mathbf{y}_2$ then provides the remaining estimator $P(\pi_{\text{FP}}|\vec{\mathbf{x}}, \vec{\mathbf{w}})$.

We applied PROPROC²³ to the first component of the three-class BANN output \mathbf{y}_1 obtained from the feature value observations in the testing set. For the purposes of this analysis,

an observation was considered actually positive if it belonged to the class π_{mal} , and actually negative if it belonged to the class $\pi_{\text{ben}} \oplus \pi_{\text{FP}}$ (*i. e.*, if it belonged to either of the classes π_{ben} or π_{FP}). For comparison, we then trained a two-class BANN on the feature values in the training set, grouped into the two classes π_{mal} and $\pi_{\text{ben}} \oplus \pi_{\text{FP}}$, and applied PROPROC to the output of this two-class BANN on the similarly grouped testing set observations.

Next, we applied PROPROC to the sum of the two components of the three-class BANN output $\mathbf{y}_1 + \mathbf{y}_2$ obtained from the feature value observations in the testing set. For the purposes of this analysis, an observation was considered actually positive if it belonged to the class $\pi_{\text{mal}} \oplus \pi_{\text{ben}}$ (*i. e.*, if it belonged to either of the classes π_{mal} or π_{ben}), and actually negative if it belonged to the class π_{FP} . For comparison, we then trained a two-class BANN on the feature values in the training set, grouped into the two classes $\pi_{\text{mal}} \oplus \pi_{\text{ben}}$ and π_{FP} , and applied PROPROC to the output of this two-class BANN on the similarly grouped testing set observations.

The above-described procedure was repeated for 200 different random jackknifings of the data set into training and testing sets. The number of observations in each set for a given class was not fixed across the jackknifings, but was approximately half of the total number of observations in that class. The areas under the curves (AUCs) produced by PROPROC for a given (three- or two-class) BANN and task were averaged, and the standard error in each mean was also computed. To test the (null) hypothesis that the AUCs for the three-class and two-class BANNs for a given task are drawn from the same distribution, we calculated the 0.975 percentile of the Student- t distribution with 199 degrees of freedom, which we denote $t_{\text{df}=199;1-\alpha=0.975} = 1.972$. We then computed a confidence interval $[\overline{\Delta\text{AUC}} - t_{\text{df}=199;1-\alpha=0.975} \widehat{\text{SE}}_{\Delta\text{AUC}}, \overline{\Delta\text{AUC}} + t_{\text{df}=199;1-\alpha=0.975} \widehat{\text{SE}}_{\Delta\text{AUC}}]$, which under the null hypothesis should have a 95% probability of including 0.²⁵ Here $\overline{\Delta\text{AUC}}$ is the sample mean of the difference in AUC values between the three-class and two-class BANNs for a given architecture and task, and $\widehat{\text{SE}}_{\Delta\text{AUC}}$ denotes the standard error in this mean. It could be argued that the measured areas contributing to $\overline{\Delta\text{AUC}}$ are not independent across jackknifings, which would seem to invalidate the use of the Student- t test for determining statistical significance here. Note, though, that the independence assumption primarily affects the calculation of $\widehat{\text{SE}}_{\Delta\text{AUC}}$, and that for positive correlations such as would be expected in this case, the value of $\widehat{\text{SE}}_{\Delta\text{AUC}}^2$ becomes an underestimate of the true variance. This means that the “true” confidence intervals would, in fact, be even larger than those calculated here.

Similarly, we would then expect the actual distribution of $\overline{\Delta\text{AUC}}/\widehat{\text{SE}}_{\Delta\text{AUC}}$ to be even wider than a Student- t distribution with the assumed number of degrees of freedom, and so the use of the boundary value $t_{\text{df}=199;1-\alpha=0.975}$ should further serve to underestimate the true size of the confidence interval. Thus, our results are “conservative” in the sense that the actual confidence intervals involved will be, at worst, even larger than those calculated, and so our conclusions in terms of statistical significance will be unchanged.

In order to determine whether 200 jackknifings are adequate to estimate the means in question, we found the mean AUCs and standard errors in the means for BANNs with seven hidden units for 50, 75, 100, 125, 150, 200, 250, 300, and 600 jackknifings of the data set. Under the assumption that these quantities are adequately estimated, the mean values should remain constant, and the square of the standard error should vary as the inverse of the number of jackknifings.

IV. RESULTS

Figure 1 shows the mean AUCs across 200 jackknifings of the data set, plus and minus two standard errors in the means, of the ROC curves produced by PROPROC for the first output \mathbf{y}_1 of the three-class BANNs, as well as for the two-class BANNs trained specifically for the task of distinguishing malignant from nonmalignant detections. Values are shown for BANN architectures with five to ten hidden units.

Similarly, Fig. 2 shows the mean AUCs across 200 jackknifings of the data set, plus and minus two standard errors in the means, of the ROC curves produced by PROPROC for the sum of the outputs $\mathbf{y}_1 + \mathbf{y}_2$ of the three-class BANNs, as well as for the two-class BANNs trained specifically for the task of distinguishing TP from FP computer detections. Values are again shown for BANN architectures with five to ten hidden units.

Table I shows the computed confidence intervals $[\overline{\Delta\text{AUC}} - t_{\text{df}=199;1-\alpha=0.975}\widehat{\text{SE}}_{\Delta\text{AUC}}, \overline{\Delta\text{AUC}} + t_{\text{df}=199;1-\alpha=0.975}\widehat{\text{SE}}_{\Delta\text{AUC}}]$ ($t_{\text{df}=199;1-\alpha=0.975} = 1.972$) for each of the two classification tasks (malignant *vs.* nonmalignant, and TP *vs.* FP computer detections), and for the six BANN architectures (numbers of hidden units) described in Sec. III. The inclusion of 0 in any given confidence interval implies that the difference in mean AUC between the three-class and two-class BANNs for that architecture and task was not statistically significant at $p = 0.05$.²⁵ In this sense, no statistically

significant differences were found between the three-class and two-class BANNs for any given task and architecture.

Because AUC is not a complete description of a ROC curve, ROC curves for the three-class and two-class BANNs are shown in Fig. 3 for the task of distinguishing malignant from nonmalignant detections. Similarly, Fig. 4 shows ROC curves for the three-class and two-class BANNs for the task of distinguishing TP from FP computer detections. The curves in these two figures are averages across the 200 jackknifings.

To evaluate the adequacy of 200 jackknifings for estimating the means in question, we show the mean AUC in Fig. 5, and the square of the standard error in the mean in Fig. 6, *versus* the inverse of the number of jackknifings, for the three-class and two-class BANNs with seven hidden units for the task of distinguishing malignant from nonmalignant detections. Similarly, we show the mean AUC in Fig. 7, and the square of the standard error in the mean in Fig. 8, *versus* the inverse of the number of jackknifings, for the three-class and two-class BANNs with seven hidden units for the task of distinguishing TP from FP computer detections. Note that the error bars in Figs. 5 and 7 are again plus and minus two standard errors in the means, while the error bars in the squares of the standard errors in Figs. 6 and 8 are

$$\left[\frac{n-1}{\chi_{0.975}^2(n-1)} \widehat{\mathbf{SE}}_{\text{AUC}}^2, \frac{n-1}{\chi_{0.025}^2(n-1)} \widehat{\mathbf{SE}}_{\text{AUC}}^2 \right], \quad (7)$$

where n is the number of jackknifings. (This is the 95% confidence interval for the variance in the sample mean of the AUC for a given task and BANN architecture, which is estimated by the square of the standard error in that mean.)

V. DISCUSSION

The mean AUC values shown in Figs. 1 and 2 do not show much variation across BANN architecture (*i. e.*, the number of hidden units present). This is consistent with previous observations made on BANNs,¹⁶ namely that, given sufficient training data, the presence of additional hidden units beyond those needed to model the complexity of the classification problem does not lead to overtraining as it would for a conventional ANN.

The agreement between the three-class BANN output applied to a specific two-class classification task as explained in Sec. II, and the corresponding specifically trained two-class

BANN, is excellent. This relationship was derived under the explicit assumption that the decision variables involved are ideal observer decision variables. We have shown in previous simulation studies that two-class¹⁶ and three-class¹⁹ BANNs can accurately estimate ideal observer decision variables given sufficient training data.

The variation in mean values with number of jackknifings of the data set shown in Figs. 5 and 7 is fairly small at 200 or more jackknifings, and within the error bars shown in those figures. Similarly, the dependence in the square of the standard error in these means on the inverse of the number of jackknifings shown in Figs. 6 and 8 is essentially linear at 200 or more jackknifings. These observations suggest that 200 jackknifings are adequate for estimating the means and standard errors in Figs. 1 and 2, and for determining the statistical significance of the differences in these means as summarized in Table I.

In the present study, as in previous studies we have made using two-class BANNs,^{5,13-15} our data set consists of real observational data in the form of feature values calculated from objects detected in actual mammographic images. We therefore do not have access to the “true” underlying PDFs of this observational data. It is conceivable that the relationship between the three-class and two-class decision variables, derived in Sec. II for ideal observer decision variables, holds fortuitously for the decision variables actually obtained from our observational data *via* the trained BANNs. Nevertheless, we consider the results presented here to be strong circumstantial evidence that the BANN can yield decision variables that will be appropriate and useful for three-class classification tasks, just as it has for conventional two-class classification tasks.

VI. CONCLUSIONS

We have found that a BANN can be trained to produce decision variables for three-class classification of actual (nonsimulated) observational data. Although interpretation of the decision variable data is currently difficult due to the lack of a fully general method for evaluating three-class classifiers, we have conducted a simplified analysis which interprets the three-class decision variables in terms of two two-class decision tasks. No statistically significant differences in performance were found between the three-class BANN outputs when interpreted for such two-class tasks, and the corresponding two-class BANNs trained specifically for these tasks. This result is consistent with the theoretical relationship between

three-class and two-class ideal observer decision variables, and strongly suggests that the BANN will prove to be as valid a choice for three-class classification tasks as it has for two-class classification tasks.

We currently lack a method for performing a fully general three-class evaluation of this method — *i. e.*, a three-class extension of conventional two-class ROC analysis. Although further work is clearly needed in this area, we are confident that the BANN will prove to be of utility for three-class classification tasks, particularly medical decision tasks. The success of the method in the present application to real (nonsimulated) image feature data is particularly encouraging. Such a tool could be of great use in automatically combining CAD schemes for detecting and classifying lesions in medical images, like the mammographic mass detection and classification schemes described here.

ACKNOWLEDGMENTS

Thanks to Kwang-taeg Oh for obtaining classification scheme feature values for detection scheme false positives in the images used, and to Zhimin Huo, Ph.D., for very helpful conversations regarding the features used in her mass classification scheme. This work was supported in part by Grant No. R01-CA60187 from the National Cancer Institute (R.M. Nishikawa, principal investigator), Grant No. R01-GM57622 from the National Institutes of Health (C.E. Metz, principal investigator), and Grant No. R01-CA89452 from the National Cancer Institute (M.L. Giger, principal investigator). Li Lan, Charles E. Metz, Maryellen L. Giger, and Robert M. Nishikawa 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.

APPENDIX: A THREE-CLASS EXTENSION OF OUR INITIAL-DETECTION-AND-CANDIDATE-ANALYSIS FROC MODEL

In this work we investigated the capabilities of BANN classifiers for the three-class classification of computer-identified mammographic mass lesions and FP computer detections. In a two-class classification task, analysis of the performance of such a scheme would require a

methodology such as free-response operating characteristic (FROC) analysis,²⁶ or one of its variants,^{27,28} which can take into account the location information included in the computer detections. It is reasonable to ask whether a three-class extension of FROC analysis might be required for the analysis of a computerized scheme such as that under consideration. Such an extension is conceptually daunting, however, in light of the considerable difficulties involved in developing a three-class extension to ROC analysis (which is arguably better understood than FROC analysis); indeed, many if not all of these difficulties remain unresolved to date.

Our investigation was carried out under the assumption that classification of detected lesions can be considered effectively independent of the computer scheme’s initial detection of candidates for classification. That is, we model the computer scheme as consisting of two independent stages: the initial detection (*i. e.*, location) of a large set of candidates of interest, followed by the classification of those candidates. This allows us to sidestep the issue of FROC analysis of the overall performance of such a scheme, as we are concerned primarily with the performance, in a ROC sense, of the (three-class) classification stage of the scheme. This assumption of independence is justified primarily by analogy with our recently proposed model for fitting FROC curves to (two-class) observational data,²⁹ which we call the “initial-detection-and-candidate-analysis” (IDCA) FROC model. Although the IDCA model may be of limited generality due to assumptions it imposes on the distributions of the data even beyond those made by existing FROC methods,²⁸ we have had success in using it to fit curves to data produced by computerized schemes.²⁹

The IDCA FROC model has two components: an expression for the observer’s overall performance in terms of the performance of the two stages (detection and classification) considered separately; and a likelihood function for the observer’s performance which can be factored into separate terms for the two stages. In this appendix, we briefly describe extensions of these two components to the case of N classes; we will not review here the case of two classes, but refer the interested reader to our previous paper.²⁹

Suppose an observer must locate and identify $N - 1$ different types of “signal,” denoted by the class labels π_1 through π_{N-1} , which may appear in a set of images. We define an N th “noise” class π_N to describe those image regions not included in any of the other $N - 1$ classes. For simplicity, one might choose to evaluate the overall performance of this observer in terms of its ability to discriminate any type of signal ($\pi_S \equiv \pi_1 \oplus \cdots \oplus \pi_{N-1}$)

from noise. Bunch *et al.* showed^{30,31} that if the observer uses a constant critical value of a latent decision variable t to make decisions for all detections, the observer's performance can be characterized by two quantities: the signal detection fraction (SDF), defined as the probability that an actually positive signal is detected; and the expectation value of the number of false-positive detections per image (FPpI). The Bunch FROC model makes the assumption that the observer's detections are all mutually independent; we adopt this assumption in the IDCA FROC model.

Let the operating point of the observer's candidate detection stage be $(\text{FPpI}_c, \text{SDF}_c)$, here assumed fixed, and let the performance of the observer's analysis stage for a corresponding population of selected candidates be given by a point on a ROC hypersurface whose coordinates are probabilities of the form $P(\text{"}\pi_i\text{"}|\pi_j; t)$, where the quotation marks indicate the observer's decision, π_j is the true class of the observation, and the nonrandom parameter t indicates the observer's decision threshold. For a given decision threshold t , the observer's SDF is just the conditional probability that a TP decision is made for a given candidate detection, multiplied by the probability that the candidate was detected in the first place:

$$\begin{aligned}
\text{SDF}(t) &\equiv P(\text{signal detected as candidate and decided "}\pi_S\text{"}) \\
&= P(\pi_S \text{ cand. decided "}\pi_S\text{"} | \pi_S \text{ detected as cand.}) \times P(\pi_S \text{ detected as cand.}) \\
&= \left[\frac{\sum_1^{N-1} P(\text{"}\pi_S\text{"} | \pi_i; t) P(\pi_i)}{\sum_1^{N-1} P(\pi_i)} \right] \times \text{SDF}_c. \tag{A.1}
\end{aligned}$$

Note that the fraction in brackets depends only upon the classification stage (ROC) performance parameters, and not directly upon the detection stage performance parameter SDF_c . (The SDF_c value may influence the spectrum of difficulty of the candidates for classification, however, which in turn can influence the classification stage ROC performance.)

Similarly, the false-positive rate $\text{FPpI}(t)$ can be evaluated in terms of ROC parameters and FPpI_c :

$$\begin{aligned}
\text{FPpI}(t) &\equiv \langle \text{No. of } \pi_N \text{ detections called "}\pi_S\text{" in 1 image} \rangle \\
&= \sum_f f P(f \pi_N \text{ candidates detected and decided "}\pi_S\text{"}) \\
&= \sum_f f \sum_{g \geq f} P(f \text{ out of } g \pi_N \text{ cands. decided "}\pi_S\text{"} | g \pi_N \text{ cands.}) \times P(g \pi_N \text{ cands.}) \\
&= \sum_g P(g \pi_N \text{ cands.}) \sum_{f \leq g} f P(f \text{ out of } g \pi_N \text{ cands. decided "}\pi_S\text{"} | g \pi_N \text{ cands.}) \\
&= \sum_g P(g \pi_N \text{ cands.}) g [1 - P(\text{"}\pi_N\text{"} | \pi_N; t)]
\end{aligned}$$

$$= [1 - P(\text{“}\pi_N\text{”}|\pi_N; t)] \times \text{FPpI}_c. \quad (\text{A.2})$$

(Here f and g are particular numbers of candidate detections, and the angular brackets indicate that the expectation value of the enclosed expression is being taken.) Thus the IDCA model implies that the overall FROC performance is a linearly scaled version of a curve determined solely by the analysis stage ROC hypersurface coordinates, where the scaling factors are given by the candidate detection FROC operating point coordinates. It should also be clear that, had we chosen to define a nonsimplified N -class FROC methodology, its coordinates could still be expressed as conditional probabilities factorable into two terms, one depending only on the classification stage ROC operating point coordinates, and the other depending only on the detection stage FROC coordinates ($\text{FPpI}_c, \text{SDF}_c$).

Let

$$\mathbf{L}_{\text{FROC}} \equiv P(\vec{\mathbf{r}}, \vec{\mathbf{s}}_1, \dots, \vec{\mathbf{s}}_{N-1}) \quad (\text{A.3})$$

represent the likelihood of obtaining a set of category responses $\vec{\mathbf{r}}$ for the actually negative candidate detections (π_N) and, jointly, the sets of category responses $\vec{\mathbf{s}}_i$ for the candidate detections actually from class π_i ($i < N$), respectively. \mathbf{L}_{FROC} is implicitly a function of the model parameters to be estimated. Note that this expression can be expanded using Bayes’s theorem²⁴ to obtain

$$\mathbf{L}_{\text{FROC}} = P(\vec{\mathbf{r}}, \vec{\mathbf{s}}_1, \dots, \vec{\mathbf{s}}_{N-1} | \mathbf{B}, \mathbf{C})P(\mathbf{B}, \mathbf{C}), \quad (\text{A.4})$$

where \mathbf{B} is the total number of actually negative candidate detections (class π_N) and \mathbf{C} is the total number of actually positive candidate detections (class π_S) in the data set. That is, \mathbf{B} and \mathbf{C} are the counts of the candidate detections initially detected in the set of I images, while $\vec{\mathbf{r}}$ and $\vec{\mathbf{s}}_i$ contain the category responses for the candidates (each element corresponds to a different category, and the sums of the category counts are the number of actually negative and actually positive candidates, respectively). This requires that $\mathbf{B} = \sum_j \mathbf{r}_j$ and $\mathbf{C} = \sum_{i,j} \mathbf{s}_{ij}$.

As in the two-class IDCA FROC model, the first term on the righthand side of Eq. (A.4) is dependent only on the parameters governing the classification stage probabilities $P(\text{“}\pi_i\text{”}|\pi_j; t)$, and can be interpreted as an N -class ROC likelihood function. The second term on the righthand side of Eq. (A.4) depends only on the detection stage parameters FPpI_c and SDF_c . Based on this hypothesized separability, we have focused our attention in the present work on the classification stage of our CAD scheme, and have not

attempted to model its performance in a three-class FROC sense.

-
- ¹ U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, “Automated segmentation of digitized mammograms,” *Acad. Radiol.* **2**, 1–9 (1995).
 - ² F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, “Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images,” *Med. Phys.* **18**, 955–963 (1991).
 - ³ F.-F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, “Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses,” *Invest. Radiol.* **28**, 473–481 (1993).
 - ⁴ F.-F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, “Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique,” *Med. Phys.* **21**, 445–452 (1994).
 - ⁵ M. A. Kupinski, *Computerized Pattern Classification in Medical Imaging*, Ph.D. thesis, The University of Chicago, Chicago, Illinois (2000).
 - ⁶ M. A. Kupinski and M. L. Giger, “Automated seeded lesion segmentation on digital mammograms,” *IEEE Trans. Med. Imag.* **17**, 510–517 (1998).
 - ⁷ Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, “Automated computerized classification of malignant and benign masses on digitized mammograms,” *Acad. Radiol.* **5**, 155–168 (1998).
 - ⁸ Z. Huo, M. L. Giger, and C. E. Metz, “Effect of dominant features on neural network performance in the classification of mammographic lesions,” *Phys. Med. Biol.* **44**, 2579–2595 (1999).
 - ⁹ Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz, “Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness,” *Acad. Radiol.* **7**, 1077–1084 (2000).
 - ¹⁰ Z. Huo, M. L. Giger, and C. J. Vyborny, “Computerized analysis of multiple-mammographic views: Potential usefulness of special view mammograms in computer-aided diagnosis,” *IEEE Trans. Med. Imag.* **20**, 1285–1292 (2001).
 - ¹¹ Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, “Breast cancer: Effectiveness of computer-aided diagnosis — Observer study with independent database of mammograms,” *Radiology* **224**, 560–568 (2002).

- ¹² Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, “Analysis of spiculation in the computerized classification of mammographic masses,” *Med. Phys.* **22**, 1569–1579 (1995).
- ¹³ Z. Huo and M. L. Giger, “Effect of case mix on feature selection in the computerized classification of mammographic lesions,” in *Medical Imaging 2002: Image Processing*, edited by Milan Sonka and J. Michael Fitzpatrick, Proc. SPIE **4684**, pp. 762–767 (2002).
- ¹⁴ 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 *IWDM 2000: 5th International Workshop on Digital Mammography*, edited by M. J. Yaffe (Medical Physics Publishing, Madison, Wisconsin, 2001), Proceedings of the Workshop, pp. 168–173.
- ¹⁵ D. C. Edwards, J. Papaioannou, Y. Jiang, M. A. Kupinski, and R. M. Nishikawa, “Eliminating false-positive microcalcification clusters in a mammography CAD scheme using a Bayesian neural network,” in *Medical Imaging 2001: Image Processing*, edited by Milan Sonka and Kenneth Hanson, Proc. SPIE **4322**, pp. 1954–1960 (2001).
- ¹⁶ 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.* **20**, 886–899 (2001).
- ¹⁷ D. J. S. MacKay, *Bayesian Methods for Adaptive Models*, Ph.D. thesis, California Institute of Technology, Pasadena, California (1992).
- ¹⁸ D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, “The multilayer perceptron as an approximation to a Bayes optimal discriminant function,” *IEEE Trans. Neural Net.* **1**, 296–298 (1990).
- ¹⁹ D. C. Edwards, C. E. Metz, and R. M. Nishikawa, “Estimation of three-class ideal observer decision functions with a Bayesian artificial neural network,” in *Medical Imaging 2002: Image Perception, Observer Performance, and Technology Assessment*, edited by Dev P. Chakraborty and Elizabeth A. Krupinski, Proc. SPIE **4686**, pp. 1–12 (2002).
- ²⁰ H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, “Design of three-class classifiers in computer-aided diagnosis: Monte carlo simulation study,” in *Medical Imaging 2003: Image Processing*, edited by Milan Sonka and J. Michael Fitzpatrick, Proc. SPIE **5032**, pp. 567–578 (2003).
- ²¹ D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, “Bayesian ANN

- estimates of three-class ideal observer decision variables for classification of mammographic masses,” in *Medical Imaging 2003: Image Perception, Observer Performance, and Technology Assessment*, edited by Dev P. Chakraborty and Elizabeth A. Krupinski, Proc. SPIE **5034**, pp. 474–482 (2003).
- ²² H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I* (John Wiley & Sons, New York, 1968).
- ²³ C. E. Metz and X. Pan, “‘Proper’ binormal ROC curves: Theory and maximum-likelihood estimation,” *J. Math. Psychol.* **43**, 1–33 (1999).
- ²⁴ A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, Inc., New York, 1991).
- ²⁵ C. E. Metz, “Quantification of failure to demonstrate statistical significance: The usefulness of confidence intervals,” *Invest. Radiol.* **28**, 59–63 (1993).
- ²⁶ D. P. Chakraborty, “Maximum likelihood analysis of free-response operating characteristic (FROC) data,” *Med. Phys.* **16**, 561–568 (1989).
- ²⁷ R. G. Swensson, “Unified measurement of observer performance in detecting and localizing target objects on images,” *Med. Phys.* **23**, 1709–1725 (1996).
- ²⁸ D. P. Chakraborty, “Chapter 16: The FROC, AFROC, and DROC variants of the ROC analysis,” in *Handbook of Medical Imaging, Vol. 1: Physics and Psychophysics*, edited by J. Beutel, H. L. Kundel, and R. L. Van Metter (SPIE, Bellingham, WA, 2000), pp. 771–796.
- ²⁹ D. C. Edwards, M. A. Kupinski, C. E. Metz, and R. M. Nishikawa, “Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model,” *Med. Phys.* **29**, 2861–2870 (2002).
- ³⁰ P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, “A free response approach to the measurement and characterization of radiographic observer performance,” in *Application of Optical Instrumentation in Medicine VI*, edited by J. E. Gray and W. R. Hendee, Proc. SPIE **127**, pp. 124–135 (1977).
- ³¹ P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, “A free response approach to the measurement and characterization of radiographic-observer performance,” *J. Appl. Photogr. Eng.* **4**, 166–172 (1978).

TABLE I: Confidence intervals $[\Delta\text{AUC} - t_{\text{df}=199;1-\alpha=0.975}\widehat{\text{SE}}_{\Delta\text{AUC}}, \Delta\text{AUC} + t_{\text{df}=199;1-\alpha=0.975}\widehat{\text{SE}}_{\Delta\text{AUC}}]$ ($t_{\text{df}=199;1-\alpha=0.975} = 1.972$) for each of the two classification tasks (malignant *vs.* nonmalignant, and TP *vs.* FP computer detections), and for the six BANN architectures (numbers of hidden units) described in Sec. III. (The inclusion of 0 in any given confidence interval implies that the difference in mean AUC between the three-class and two-class BANNs for that architecture and task was not statistically significant at $p = 0.05$.)

Task	$H = 5$	6	7	8	9	10
Malignant <i>vs.</i> non-malignant	[-0.004, 0.009]	[-0.004, 0.008]	[-0.007, 0.007]	[-0.006, 0.006]	[-0.003, 0.010]	[-0.007, 0.006]
TP <i>vs.</i> FP computer detection	[-0.003, 0.007]	[-0.002, 0.008]	[-0.004, 0.007]	[-0.005, 0.005]	[-0.003, 0.007]	[-0.004, 0.006]

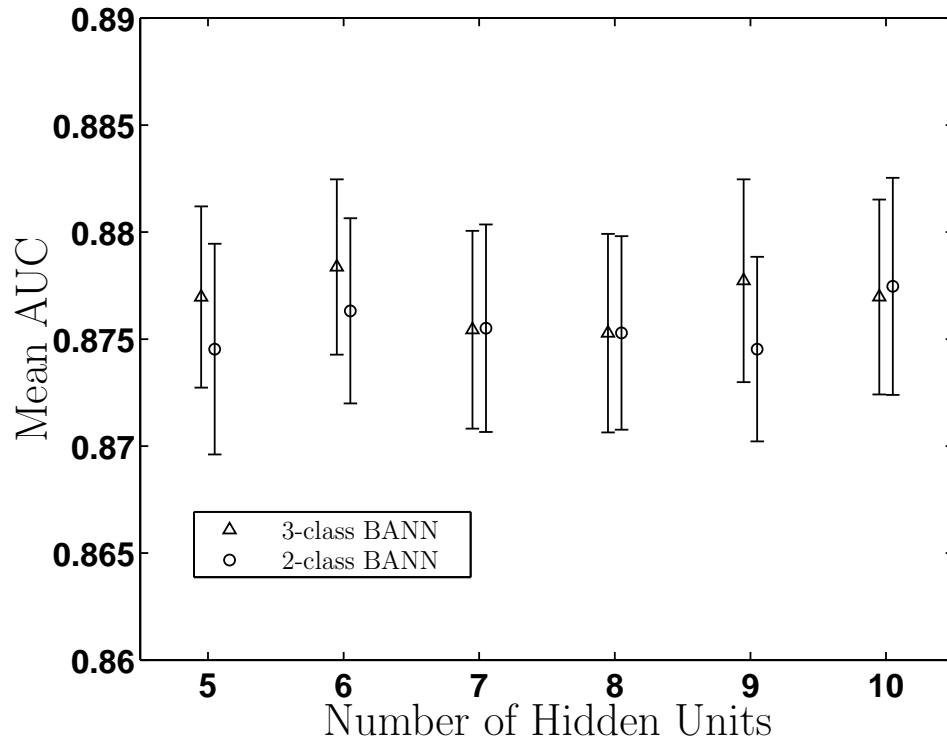


FIG. 1: Mean AUCs across 200 jackknifings of the data set, plus and minus two standard errors in the means, of ROC curves produced by PROPROC for the first output of the three-class BANN, and for a two-class BANN trained specifically to distinguish malignant from nonmalignant detections. Values are shown for BANN architectures with five to ten hidden units.

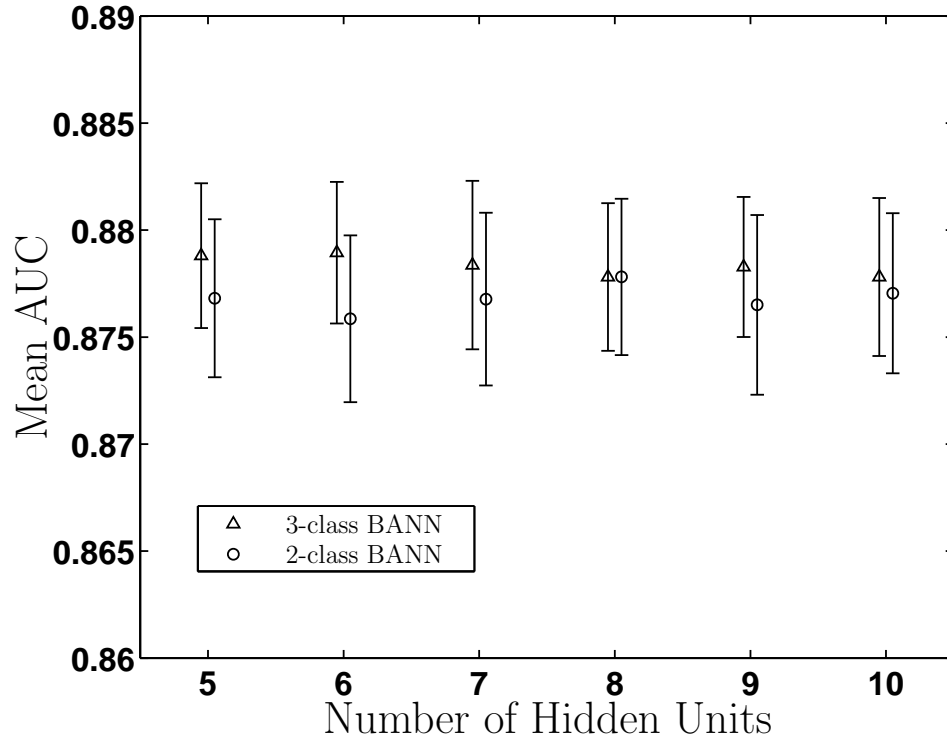


FIG. 2: Mean AUCs across 200 jackknifings of the data set, plus and minus two standard errors in the means, of ROC curves produced by PROPROC for the sum of the outputs of the three-class BANN, and for a two-class BANN trained specifically to distinguish TP from FP computer detections. Values are shown for BANN architectures with five to ten hidden units.

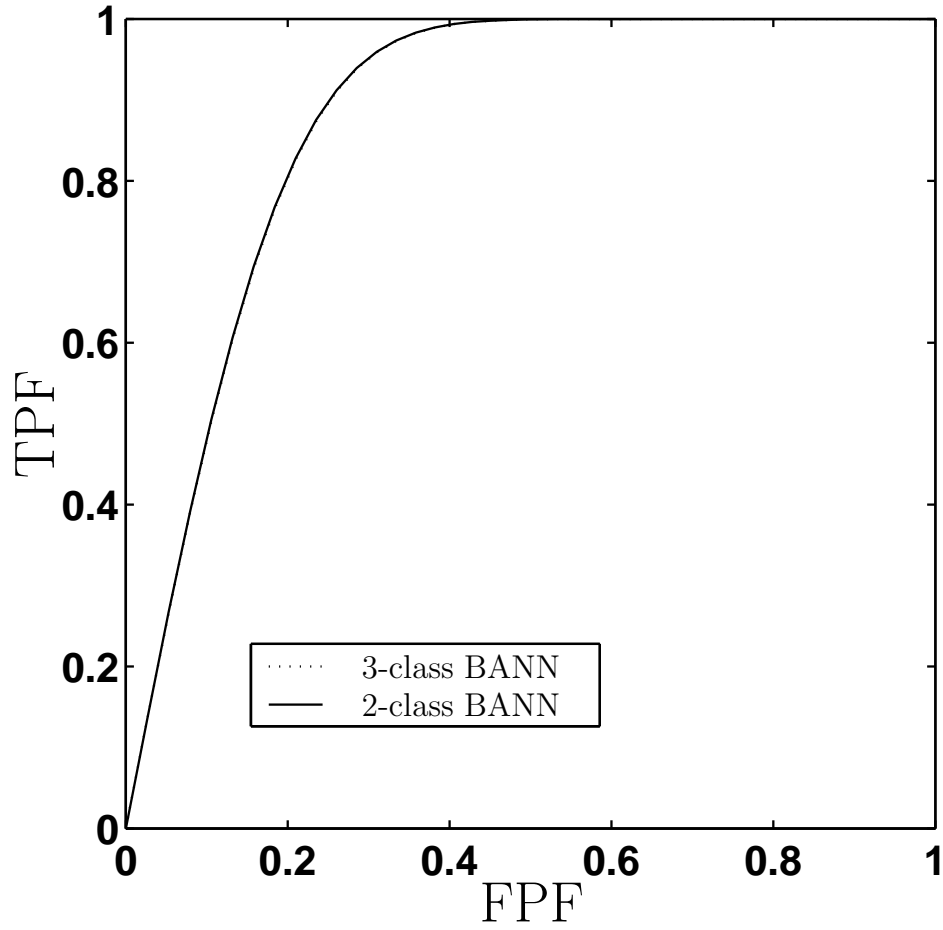


FIG. 3: ROC curves produced by PROPROC averaged over 200 jackknifings for the first output of the three-class BANN, and for a two-class BANN trained specifically to distinguish malignant from nonmalignant detections. Values are shown for BANN architectures with ten hidden units. (The two curves shown are virtually indistinguishable.)

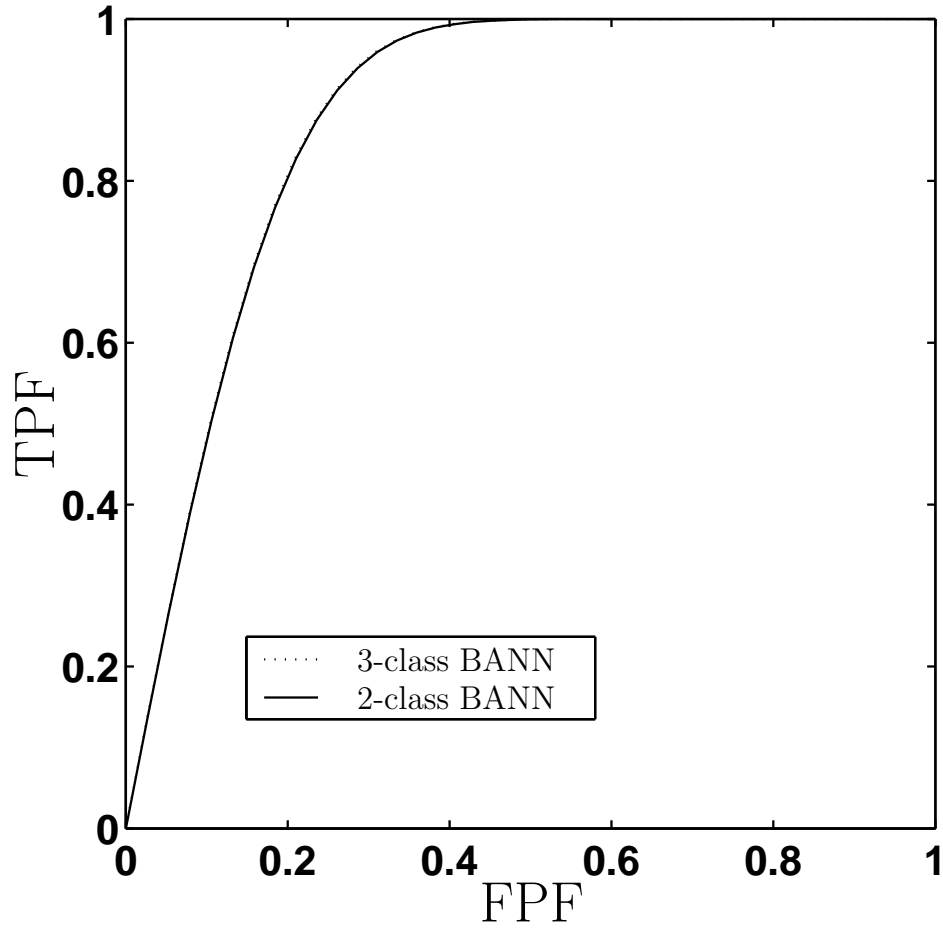


FIG. 4: ROC curves produced by PROPROC averaged over 200 jackknifings for the sum of the outputs of the three-class BANN, and for a two-class BANN trained specifically to distinguish TP from FP computer detections. Values are shown for BANN architectures with ten hidden units. (The two curves shown are virtually indistinguishable.)

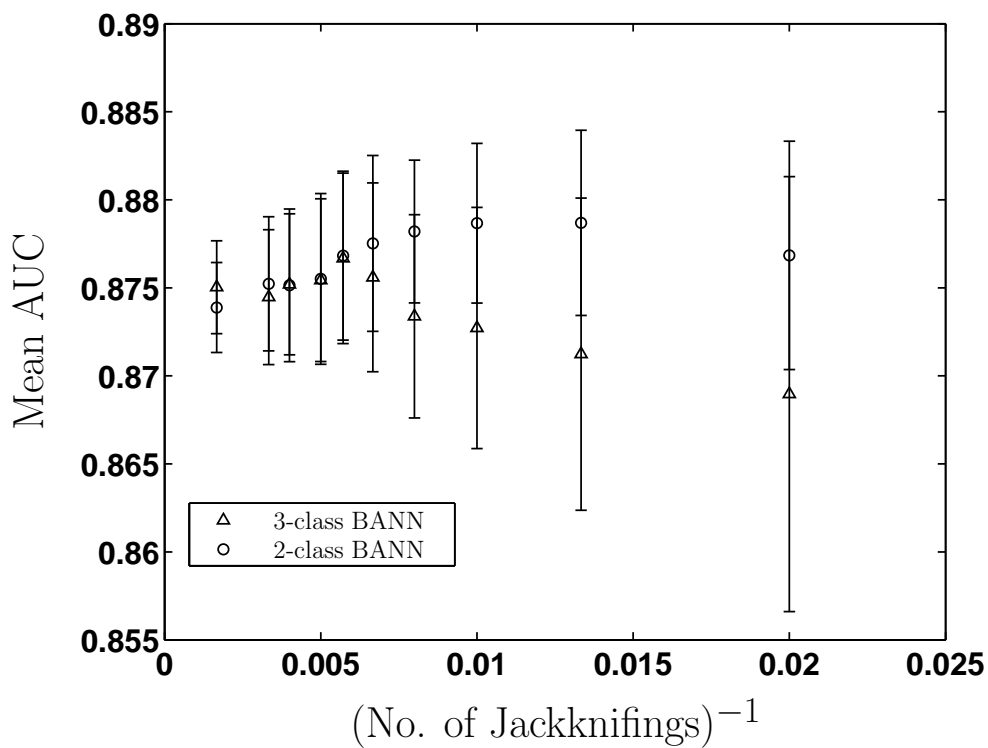


FIG. 5: Mean AUC, plus and minus two standard errors in the mean, for three-class and two-class BANNs with seven hidden units for the task of distinguishing malignant from nonmalignant detections, *vs.* the inverse of the number of jackknifings of the data set into training and testing sets. The numbers of jackknifings used were 50, 75, 100, 125, 150, 175, 200, 250, 300, and 600.

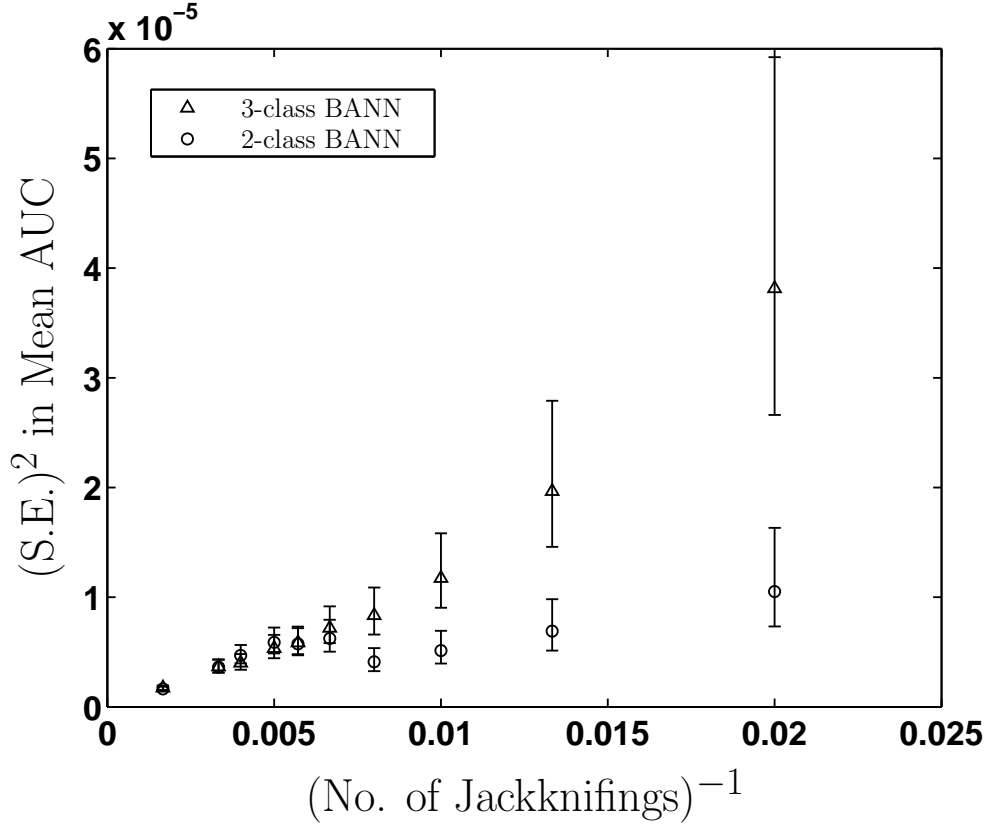


FIG. 6: Square of the standard error in the mean AUC for three-class and two-class BANNs with seven hidden units for the task of distinguishing malignant from nonmalignant detections, *vs.* the inverse of the number of jackknifings of the data set into training and testing sets. The numbers of jackknifings used were 50, 75, 100, 125, 150, 175, 200, 250, 300, and 600, and the error bars are $[\frac{n-1}{\chi_{0.975}^2(n-1)} \widehat{\mathbf{SE}}_{\text{AUC}}^2, \frac{n-1}{\chi_{0.025}^2(n-1)} \widehat{\mathbf{SE}}_{\text{AUC}}^2]$ where n is the number of jackknifings.

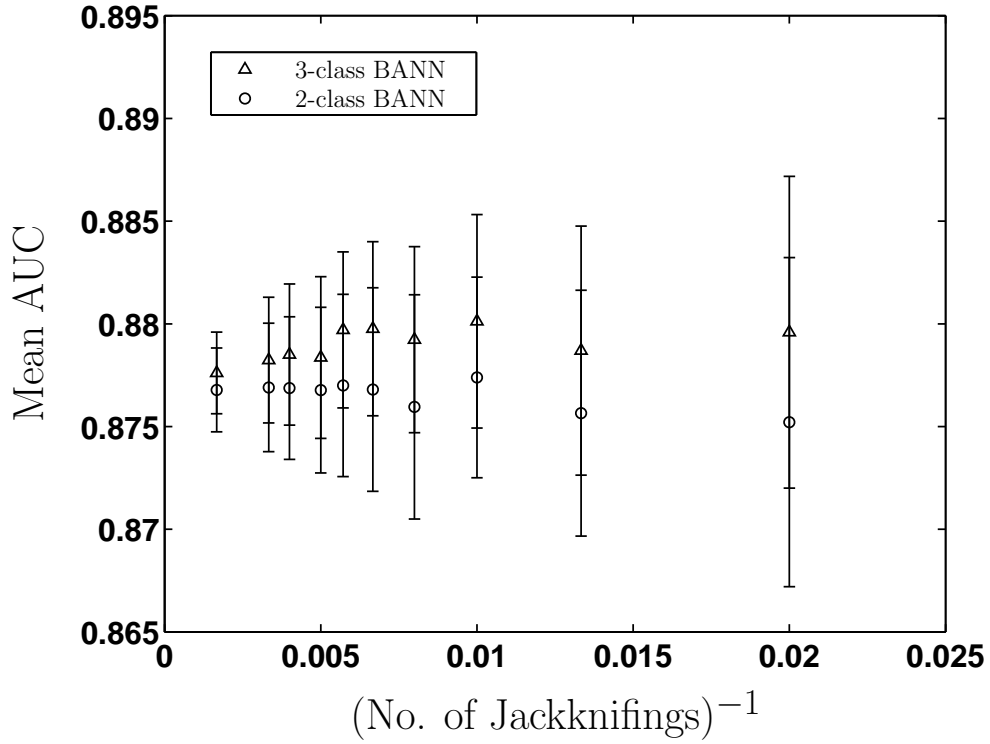


FIG. 7: Mean AUC, plus and minus two standard errors in the mean, for three-class and two-class BANNs with seven hidden units for the task of distinguishing TP from FP computer detections, *vs.* the inverse of the number of jackknifings of the data set into training and testing sets. The numbers of jackknifings used were 50, 75, 100, 125, 150, 175, 200, 250, 300, and 600.

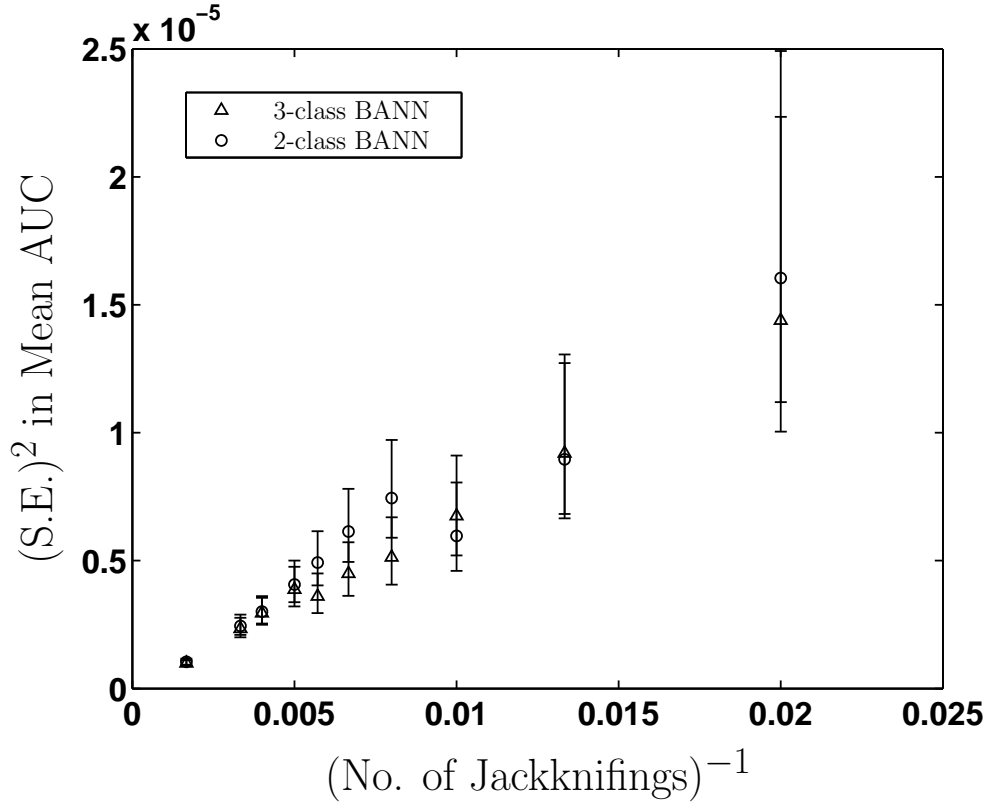


FIG. 8: Square of the standard error in the mean AUC for three-class and two-class BANNs with seven hidden units for the task of distinguishing TP from FP computer detections, *vs.* the inverse of the number of jackknifings of the data set into training and testing sets. The number of jackknifings used were 50, 75, 100, 125, 150, 175, 200, 250, 300, and 600, and the error bars are $[\frac{n-1}{\chi_{0.975}^2(n-1)} \widehat{\mathbf{SE}}_{\text{AUC}}^2, \frac{n-1}{\chi_{0.025}^2(n-1)} \widehat{\mathbf{SE}}_{\text{AUC}}^2]$ where n is the number of jackknifings.