

Evaluating Bayesian ANN estimates of ideal observer decision variables by comparison with identity functions

Darrin C. Edwards* and Charles E. Metz

Department of Radiology, The University of Chicago, Chicago, IL 60637

ABSTRACT

Bayesian artificial neural networks (BANNs) have proven useful in two-class classification tasks, and are claimed to provide good estimates of ideal-observer-related decision variables (the *a posteriori* class membership probabilities). We wish to apply the BANN methodology to three-class classification tasks for computer-aided diagnosis, but we currently lack a fully general extension of two-class receiver operating characteristic (ROC) analysis to objectively evaluate three-class BANN performance. It is well known that “the likelihood ratio of the likelihood ratio is the likelihood ratio.” Based on this, we found that the decision variable which is the *a posteriori* class membership probability of an observational data vector is in fact equal to the *a posteriori* class membership probability of that decision variable. Under the assumption that a BANN can provide good estimates of these *a posteriori* probabilities, a second BANN trained on the output of such a BANN should perform very similarly to an identity function. We performed a two-class and a three-class simulation study to test this hypothesis. The mean squared error (deviation from an identity function) of a two-class BANN was found to be 2.5×10^{-4} . The mean squared error of the first component of the output of a three-class BANN was found to be 2.8×10^{-4} , and that of its second component was found to be 3.8×10^{-4} . Although we currently lack a fully general method to objectively evaluate performance in a three-class classification task, circumstantial evidence suggests that two- and three-class BANNs can provide good estimates of ideal-observer-related decision variables.

Keywords: Bayesian artificial neural networks, ideal observers, three-class classification

1. INTRODUCTION

In the past, computerized methods for the detection¹⁻⁵ and classification⁶⁻¹¹ of mammographic mass lesions have been investigated at the University of Chicago. The classification scheme currently analyzes lesions which have been manually identified by a radiologist. We are attempting to develop a fully automated classification scheme by combining the existing detection and classification schemes; we have argued previously¹² that this will require a three-class classifier to account for the presence of false-positive (FP) computer detections, in addition to the malignant and benign lesions, in the output of the detection scheme.

For some time now we have explored the use of Bayesian artificial neural networks (BANNs) for a variety of detection^{5, 13, 14} and classification¹¹ tasks in computer-aided diagnosis (CAD). Our motivation for investigating BANNs is based, first, on our theoretical observation that, in the limit of infinite training data, a BANN will yield an ideal observer decision function for that data population;¹⁵ 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 (We note that the BANN implementation we are using is that of MacKay,¹⁷ which employs a multivariate normal function for the prior distribution on the network weight values.) We have also performed simulation studies showing that BANNs can accurately estimate ideal observer decision variables in a three-class classification task.¹⁵ Moreover, we showed recently that a three-class BANN could produce decision variables for actual mammographic mass lesion feature data, and that these decision variables are related to two-class BANN decision variable data in a particular way consistent with a theoretical relationship between three-class and two-class ideal observer decision variables.¹² We consider this to be strong circumstantial evidence for the ability of a BANN to estimate three-class ideal observer decision variables, though we currently lack a fully general method for evaluating three-class classifiers (*i.e.*, a three-class extension to receiver operating characteristic (ROC) analysis).

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

In this work, we present further circumstantial evidence toward the claim that a BANN can provide good estimates of three-class ideal observer decision variables. We develop a theoretical relationship between the *a posteriori* class membership probabilities of a given observational data variable and the *a posteriori* class membership probabilities of those *a posteriori* probabilities treated as a set of observational data in their own right. (It is known that *a posteriori* class membership probabilities are equivalent to ideal observer decision variables in a two-class task,¹⁶ and related in a straightforward way to the ideal observer decision variables in a task with three or more classes.¹⁵) We then describe simulation studies to train and test a set of BANNs, and present results of such a simulation study verifying that the BANNs we examined did indeed obey the theoretical relationship predicted for ideal observer decision variables, to within experimental error. In the final section, we present our conclusions drawn from this work.

2. THEORY

It is well known that the ideal observer decision variable, *i.e.*, the likelihood ratio or any monotonic transformation of this value, yields optimal performance in a two-class classification task.¹⁸ It can also be shown, in a classification task with N classes ($N > 2$), that the ideal observer decision rule becomes more complicated than a simple threshold on a single decision variable, but that the optimal decision variables remain a set of $N - 1$ likelihood ratios.^{18,19}

We can define the i th likelihood ratio as

$$\Lambda_i \equiv \text{LR}_i(\vec{x}) \equiv \frac{p(\vec{x}|\pi_i)}{p(\vec{x}|\pi_N)}, \quad (1)$$

where \vec{x} represents statistically variable observational data (which we assume to have dimensionality n), and π_j represents one of the N classes from which the data are drawn (here $1 \leq i \leq N - 1$). Clearly the vector (of dimensionality $N - 1$) of decision variables Λ_i is itself statistically variable, and one might ask what the likelihood ratios of these variables are. In fact,²⁰

$$\begin{aligned} p(\vec{\Lambda}|\pi_i) &= \int \dots \int \sum_j \frac{p(\vec{x}_j|\pi_i)}{|J(\vec{x}_j)|} dx^N \dots dx^n \\ &= \int \dots \int \sum_j \text{LR}_i(\vec{x}_j) \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|} dx^N \dots dx^n, \end{aligned} \quad (2)$$

where we have assumed that $N - 1 < n$; if $N - 1 = n$, then no integration is performed. (If $N - 1 > n$, then at least one of the likelihood ratio decision variables will be expressible as a function of the others; we will not consider this degenerate case here.) The sum is over all solutions to Eq. 1 for a given $\vec{\Lambda}$; this yields

$$\begin{aligned} p(\vec{\Lambda}|\pi_i) &= \int \dots \int \sum_j \Lambda_i \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|} dx^N \dots dx^n \\ &= \Lambda_i \int \dots \int \sum_j \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|} dx^N \dots dx^n \\ &= \Lambda_i p(\vec{\Lambda}|\pi_N) \\ \frac{p(\vec{\Lambda}|\pi_i)}{p(\vec{\Lambda}|\pi_N)} &\equiv \text{LR}_i(\vec{\Lambda}) = \Lambda_i, \end{aligned} \quad (3)$$

the source of the well-known adage that “the likelihood ratio of the likelihood ratio is the likelihood ratio.”

Consider now a different set of decision variables, the *a posteriori* class membership probabilities considered as functions of the statistically variable observational data

$$\mathbf{y}_i \equiv P(\pi_i|\vec{x}). \quad (4)$$

(Since $P(\pi_N|\vec{x}) = 1 - \sum_{i=1}^{N-1} P(\pi_i|\vec{x})$, we still have $N-1$ decision variables.) Note that in a two-class classification task, this decision variable is known to be a monotonic function of the likelihood ratio, and is therefore an ideal observer decision variable;¹⁶ while in a classification task with more than two classes, the *a posteriori* class membership probabilities can be shown to be related to the likelihood ratios in a straightforward way.¹⁵

Reasoning as above, we may ask what the *a posteriori* class membership probability of these decision variables, or $P(\pi_i|\vec{y})$, is. In fact,

$$\begin{aligned} P(\pi_i|\vec{x}) &= \frac{p(\vec{x}|\pi_i)P(\pi_i)}{p(\vec{x})} \\ &= \frac{p(\vec{x}|\pi_i)P(\pi_i)}{\sum_{k=1}^N p(\vec{x}|\pi_k)P(\pi_k)} \\ &= \frac{\text{LR}_i(\vec{x})P(\pi_i)/P(\pi_N)}{1 + \sum_{k=1}^{N-1} \text{LR}_k(\vec{x})P(\pi_k)/P(\pi_N)}, \end{aligned} \quad (5)$$

and this relation can also be inverted to yield

$$\begin{aligned} \text{LR}_i(\vec{x}) &= \frac{P(\pi_i|\vec{x})}{1 - \sum_{k=1}^{N-1} P(\pi_k|\vec{x})P(\pi_k)/P(\pi_N)} \\ &= \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)}. \end{aligned} \quad (6)$$

We again start with Eq. 2, this time obtaining

$$\begin{aligned} p(\vec{y}|\pi_i) &= \int \cdots \int \sum_j \text{LR}_i(\vec{x}_j) \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|} dx^N \cdots dx^n \\ &= \int \cdots \int \sum_j \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)} \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|} dx^N \cdots dx^n \\ &= \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)} \int \cdots \int \sum_j \frac{p(\vec{x}_j|\pi_N)}{|J(\vec{x}_j)|} dx^N \cdots dx^n \\ &= \frac{y_i}{1 - \sum_{k=1}^{N-1} y_k P(\pi_k)/P(\pi_N)} p(\vec{y}|\pi_N), \end{aligned} \quad (7)$$

where the sums in j are over all solutions to Eq. 4 for a given \vec{y} . (The fraction can be taken out of the integral because the relations in Eqs. 5 and 6 are one-to-one, and thus the set of all solutions to Eq. 4 correspond to a single value of $\text{LR}_i(\vec{x}_j)$.) This again yields

$$\text{LR}_i(\vec{y}) = \text{LR}_i(\vec{x}_j) \quad (8)$$

where \vec{y} is the vector of *a posteriori* class membership probabilities of \vec{x} from Eq. 4, and \vec{x}_j is any solution to that equation for a given \vec{y} .

It follows that

$$\begin{aligned} P(\pi_i|\vec{y}) &= \frac{\text{LR}_i(\vec{y})P(\pi_i)/P(\pi_N)}{1 + \sum_{k=1}^{N-1} \text{LR}_k(\vec{y})P(\pi_k)/P(\pi_N)} \\ &= \frac{\text{LR}_i(\vec{x}_j)P(\pi_i)/P(\pi_N)}{1 + \sum_{k=1}^{N-1} \text{LR}_k(\vec{x}_j)P(\pi_k)/P(\pi_N)} \\ &= P(\pi_i|\vec{x}_j) = y_i, \end{aligned} \quad (9)$$

where \vec{x}_j is again any solution to Eq. 4 for a given \vec{y} . This shows that a similar adage to that for likelihood ratios holds true, namely that “the *a posteriori* class probabilities of the (data) *a posteriori* class probabilities are the (data) *a posteriori* class probabilities.”

3. MATERIALS AND METHOD

We have shown in the past¹⁶ that a BANN can provide good estimates of the *a posteriori* class membership probabilities in a two-class classification task, and we have presented the results of simulation studies¹⁵ and experiments with real mammographic feature data¹² strongly suggesting that the same holds true for three-class BANNs as well. The theoretical relationship given by Eq. 9, derived in the preceding section, provides a basis for another simulation study which should provide further circumstantial evidence for the claim that two-class and three-class BANNs can provide good estimates of the two- and three-class *a posteriori* class membership probabilities (directly related to the ideal observer decision variables *via* Eq. 5), respectively.

Specifically, for the two-class simulation study, we drew 500 samples pseudorandomly from each of two distributions:

$$p(x|\pi_1) \equiv N(x; \mu_1 = 1, \sigma_1^2 = 2) \quad (10)$$

$$p(x|\pi_2) \equiv N(x; \mu_2 = 0, \sigma_2^2 = 1). \quad (11)$$

We then trained a two-class BANN with one input, five hidden units, and one output on this data, obtaining a classifier we denote by

$$y = B_1^2(x). \quad (12)$$

(The superscript denotes the number of classes being classified.) We then used this output, given the known truth states for the original observations x from which it was obtained, as training data for a second BANN with one input, five hidden units, and one output:

$$z = B_2^2(y). \quad (13)$$

Finally, we pseudorandomly sampled an independent testing set of 500 observations x from each of the two classes given in Eqs. 10 and 11. This testing set was used as input to the first BANN to obtain a testing set y^{test} ; this in turn was given as input to the second BANN, for which the output was z^{test} .

Given Eq. 9, together with the assumption that an adequately trained two-class BANN yields good estimates of the *a posteriori* class membership probabilities of the observations being classified, it should be the case that z^{test} estimates y^{test} at least to within experimental error. To verify this, we plotted z^{test} as a function of y^{test} for each of the two classes, and we computed the mean squared error

$$\text{MSE}_2 = \frac{1}{1000} \sum (z^{\text{test}} - y^{\text{test}})^2, \quad (14)$$

where the sum is over all the observations in the two classes.

Similarly, for the three-class simulation study, we drew 500 two-dimensional samples pseudorandomly from each of three distributions:

$$p(\vec{x}|\pi_1) \equiv N\left(\vec{x}; \vec{\mu}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 4 & .75 \times 2 \\ .75 \times 2 & 1 \end{bmatrix}\right) \quad (15)$$

$$p(\vec{x}|\pi_2) \equiv N\left(\vec{x}; \vec{\mu}_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & -.4 \times 1.5 \\ -.4 \times 1.5 & 2.25 \end{bmatrix}\right) \quad (16)$$

$$p(\vec{x}|\pi_3) \equiv N\left(\vec{x}; \vec{\mu}_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \quad (17)$$

We then trained a three-class BANN with two inputs, five hidden units, and two outputs on this data, obtaining a classifier we denote by

$$\vec{y} = B_1^3(\vec{x}). \quad (18)$$

We then used this output, given the known truth states for the original observations \vec{x} from which it was obtained, as training data for a second BANN with two inputs, five hidden units, and two outputs:

$$\vec{z} = B_2^3(\vec{y}). \quad (19)$$

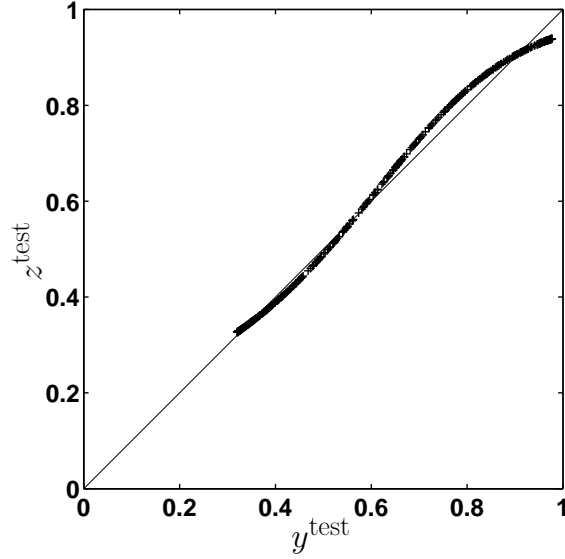


Figure 1. Output of the second two-class BANN as a function of its input for the observations actually drawn from class π_1 in the two-class simulation study.

Finally, we pseudorandomly sampled an independent testing set of 500 observations \vec{x} from each of the three classes given in Eqs. 15-17. This testing set was used as input to the first BANN to obtain a testing set \vec{y}^{test} ; this in turn was given as input to the second BANN, for which the output was \vec{z}^{test} .

Again, given Eq. 9, together with the assumption that an adequately trained two-class BANN yields good estimates of the *a posteriori* class membership probabilities of the observations being classified, it should be the case that z_1^{test} estimates y_1^{test} , and z_2^{test} estimates y_2^{test} , at least to within experimental error. To verify this, we plotted z_1^{test} as a function of y_1^{test} , and z_2^{test} as a function of y_2^{test} , for each of the three classes, and we computed the mean squared errors

$$\text{MSE}_{3i} = \frac{1}{1500} \sum (z_i^{\text{test}} - y_i^{\text{test}})^2, \quad (20)$$

$\{i : 1, 2\}$, where the sum is over all the observations in the three classes.

4. RESULTS

Figure 1 shows z^{test} as a function of y^{test} for the observations in class π_1 , and Fig. 2 shows z^{test} as a function of y^{test} for the observations in class π_2 from the two-class simulation study. The mean squared error for the complete set of 1000 observations was 2.5×10^{-4} .

Figure 3 shows the components of \vec{z}^{test} as a function of the corresponding components of \vec{y}^{test} for the observations in class π_1 . Similarly Fig. 4 shows the components of \vec{z}^{test} as a function of the corresponding components of \vec{y}^{test} for the observations in class π_2 , and Fig. 5 shows the components of \vec{z}^{test} as a function of the corresponding components of \vec{y}^{test} for the observations in class π_3 . The mean squared error for the complete set of 1500 observations was 2.8×10^{-4} for the first component and 3.8×10^{-4} for the second component.

5. DISCUSSION AND CONCLUSIONS

We developed a theoretical relationship between the *a posteriori* class membership probabilities, directly related to ideal observer decision variables, and the *a posteriori* class membership probabilities of those *a posteriori* class membership probabilities treated as statistically variable observer data in their own right. The identity relationship found is, perhaps unsurprisingly, quite similar in spirit to the identity relationship between the likelihood ratio decision variables and the likelihood ratio of those likelihood ratio decision variables for a given task.

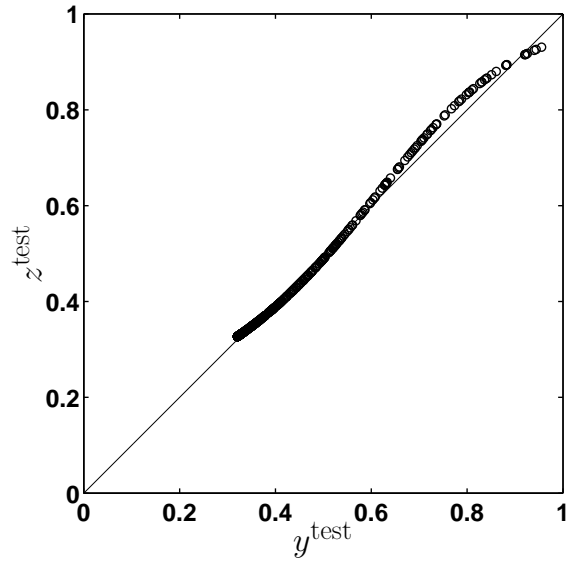


Figure 2. Output of the second two-class BANN as a function of its input for the observations actually drawn from class π_2 in the two-class simulation study.

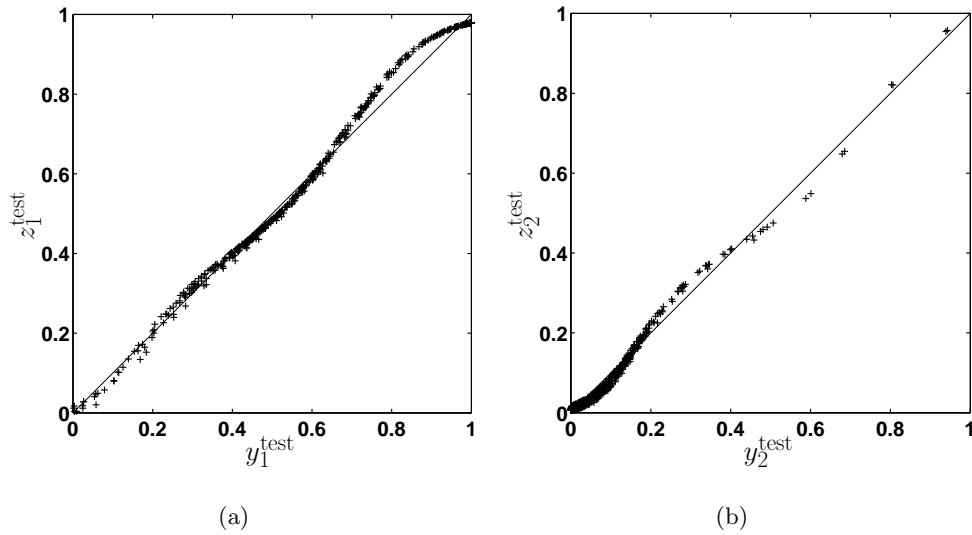


Figure 3. The (a) first and (b) second components of the output of the second three-class BANN as a function of the corresponding component of its input for the observations actually drawn from class π_1 in the three-class simulation study.

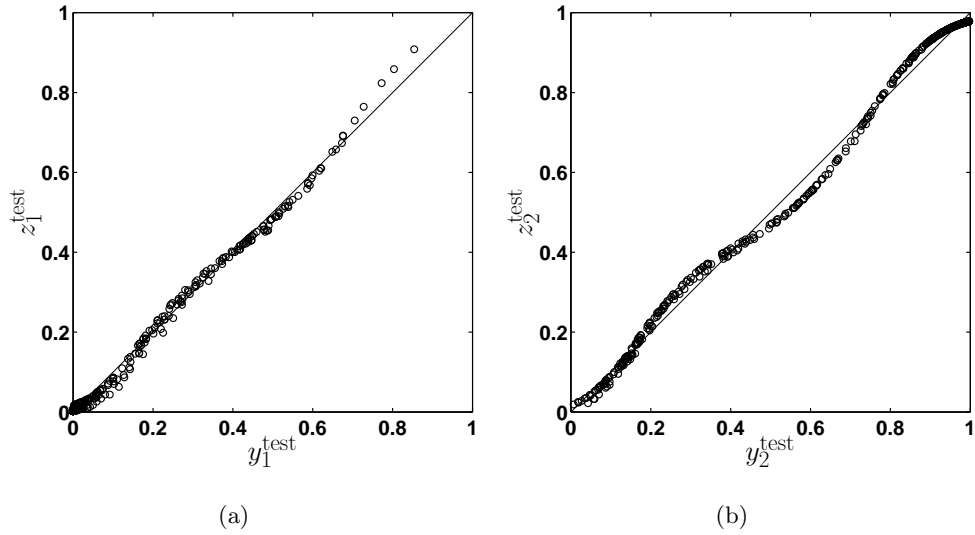


Figure 4. The (a) first and (b) second components of the output of the second three-class BANN as a function of the corresponding component of its input for the observations actually drawn from class π_2 in the three-class simulation study.

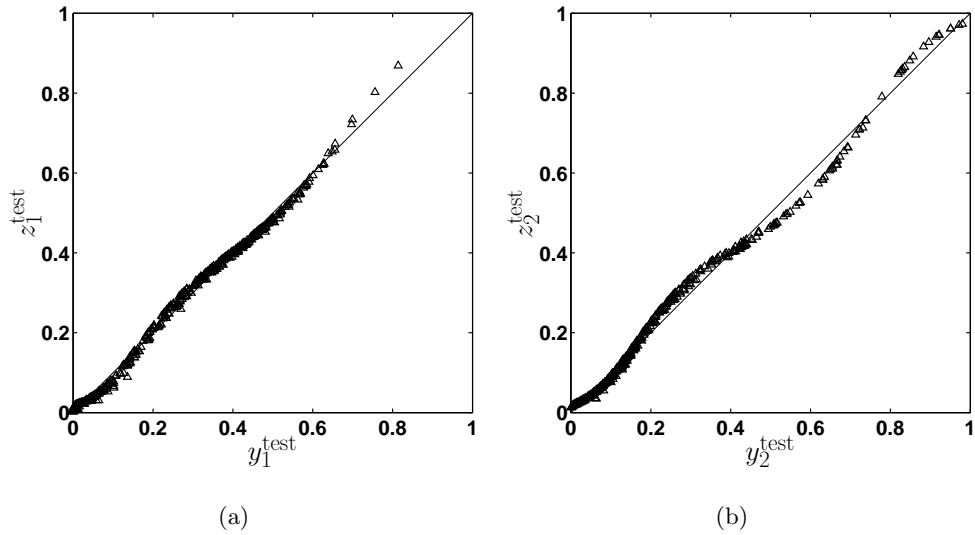


Figure 5. The (a) first and (b) second components of the output of the second three-class BANN as a function of the corresponding component of its input for the observations actually drawn from class π_3 in the three-class simulation study.

We currently lack a fully general method for three-class classification or for practically evaluating the performance of a three-class classifier. As a first step toward such a classification method, we are investigating the use of BANNs to estimate three-class ideal observer decision variables for such a task. Since, in a practical situation, we will not have access to the underlying probability distributions from which the observational data are drawn, we must rely on circumstantial evidence in support of our claim that a three-class BANN can adequately estimate decision variables directly related to ideal observer decision variables.

Previously, we presented work relating the output of a three-class BANN to the outputs of two-class BANNs trained for various “simplified” cases in which the three-class classification task was reduced to a two-class classification task, and showed that the relationships found were consistent with the relationship between three- and two-class ideal observers for the same tasks.¹² In the present work, we showed that the output of two- and three-class BANNs was consistent, to within experimental error, with the theoretical relationship developed for actual *a posteriori* class membership probabilities. This is of limited practical use in the complete development of a three-class classifier, mainly because the three-class ideal observer decision rule is considerably more complicated than its two-class counterpart (a simple threshold on a single decision variable). It does, however, bolster our confidence in the choice of the BANN as an appropriate tool for estimating the decision variables which would eventually be incorporated in such a classifier.

ACKNOWLEDGMENTS

This work was supported by grant W81XWH-04-1-0495 from the US Army Medical Research and Materiel Command (D. C. Edwards, principal investigator). Charles E. Metz is a shareholder in R2 Technology, Inc. (Sunnyvale, CA).

REFERENCES

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**, pp. 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**, pp. 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**, pp. 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**, pp. 445–452, 1994.
5. M. A. Kupinski, *Computerized Pattern Classification in Medical Imaging*. Ph.D. thesis, The University of Chicago, Chicago, IL, 2000.
6. 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**, pp. 155–168, 1998.
7. 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**, pp. 2579–2595, 1999.
8. 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**, pp. 1077–1084, 2000.
9. 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**, pp. 1285–1292, 2001.
10. 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**, pp. 560–568, 2002.

11. Z. Huo and M. L. Giger, "Effect of case mix on feature selection in the computerized classification of mammographic lesions," in Proc. SPIE Vol. 4684 *Medical Imaging 2002: Image Processing*, Milan Sonka and J. Michael Fitzpatrick, eds., pp. 762–767, (SPIE, Bellingham, WA), 2002.
12. D. C. Edwards, L. Lan, C. E. Metz, M. L. Giger, and R. M. Nishikawa, "Estimating three-class ideal observer decision variables for computerized detection and classification of mammographic mass lesions," *Med. Phys.* **31**, pp. 81–90, 2004.
13. 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*, M. J. Yaffe, ed., *Proceedings of the Workshop*, pp. 168–173, Medical Physics Publishing, (Madison, WI), 2001.
14. 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 Proc. SPIE Vol. 4322 *Medical Imaging 2001: Image Processing*, Milan Sonka and Kenneth Hanson, eds., pp. 1954–1960, (SPIE, Bellingham, WA), 2001.
15. 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 Proc. SPIE Vol. 4686 *Medical Imaging 2002: Image Perception, Observer Performance, and Technology Assessment*, Dev P. Chakraborty and Elizabeth A. Krupinski, eds., pp. 1–12, (SPIE, Bellingham, WA), 2002.
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**, pp. 886–899, 2001.
17. D. J. S. MacKay, *Bayesian Methods for Adaptive Models*. Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1992.
18. H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*, John Wiley & Sons, New York, 1968.
19. D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N -class classification," *IEEE Trans. Med. Imag.* **23**, pp. 891–895, 2004.
20. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., New York, 1991.