A multitarget training method for artificial neural network with application to computer-aided diagnosis

Bei Liu[a)]
Department of Radiation Oncology, University of Southern California, Los Angeles, California 90033 and Department of Radiology, University of Chicago, Chicago, Illinois 60637

Yulei Jiang
Department of Radiology, University of Chicago, Chicago, Illinois 60637

(Received 29 June 2011; revised 20 November 2012; accepted for publication 22 November 2012; published 26 December 2012)

Purpose: The authors propose a new training method for artificial neural networks (ANNs) in two-class classification tasks such as classifying breast lesions on a mammogram as malignant or benign.
Methods: Whereas the conventional binary training method uses binary training target values based on the diagnostic truth of a lesion being malignant or benign, the authors use multiple training target values based on more detailed histological diagnosis that presumably are related to the posterior probability of a lesion being malignant. The authors performed Monte Carlo simulation studies in which training target values were assigned based on posterior probability, and they also performed a mammography study in which training target values were assigned according to histological subtypes.
Results: These studies showed that the multitarget training method produced less variability in the ANN outputs than the binary training method. The simulation studies also showed that except for when the number of training cases was extremely large, the multitarget training method produced improved overall classification performance over the binary training method.
Conclusions: Therefore, the multitarget ANN training method is potentially useful for ANN applications in computer-aided diagnosis of breast cancer. © *2013 American Association of Physicists in Medicine*. [http://dx.doi.org/10.1118/1.4772021]

Key words: artificial neural network (ANN), ANN training, breast cancer, histological subtype, computer-aided diagnosis (CAD)

I. INTRODUCTION

Artificial neural networks (ANNs) are frequently used in computer-aided diagnosis (CAD) methods, which aim to help radiologists detect cancers or classify malignant from benign lesions.[1–4] An ANN can be viewed as a multivariate mathematical function that consists of interconnected computational nodes (or neurons), in which the connections between a particular node and nodes in the proceeding layer represent weighted linear combination of nodes in the proceeding layer (usually accompanied by a nonlinear activation function). ANN weights are determined during ANN training, in which the ANN is applied to a set of training cases and the weights are adjusted, iteratively, in an attempt to match the ANN output to known target value for each case. In CAD, three-layer backpropagation networks (also known as multilayer perceptrons) are commonly used. These ANNs consist of an input layer with an equal number of nodes as the number of input image features, a hidden layer of an empirically chosen number of nodes, and an output layer of typically one node producing a continuous output. The ANN output is compared with a binary outcome (e.g., target values of 0 and 1) if the ANN is used for classification purpose, and it is compared with continuous outcomes if the ANN is used for regression purpose. These ANNs are typically fully connected between every node in adjacent layers.

The universal approximation theorem states that the three-layer ANN described above is capable of approximating any multidimensional continuous function to any desired accuracy.[5] In a two-class classification problem, the optimal classification performance is achieved by the ideal observer who uses the likelihood ratio, or any monotonic transformation thereof, as the decision variable.[1,6] Therefore, the task for the ANN is to approximate the likelihood ratio accurately. In CAD applications, one often chooses to limit the number of hidden nodes in favor of better generalizability due to limited number of cases available for training.[7,8] Once the number of hidden nodes is fixed, the ANN architecture or, equivalently, the form of the multivariate mathematical function that transforms the ANN inputs to its output is also fixed. In this situation, it is no longer true in general that the ANN can approximate the likelihood ratio to arbitrary accuracy. Furthermore, the usually limited number of training cases can prevent the ANN from achieving full optimization of its weights. Under these conditions, the ANN output only approximates the likelihood ratio and the ANN underperforms in comparison with the ideal observer, especially when the number of training cases is small and/or the dimensionality of the input features is large.[9] In addition, the stochastic nature of ANN training implies that the ANN outputs are inherently variable.[10]

In CAD applications, one typically employs ANNs for classification tasks that involves binary outcomes (e.g., malignant vs benign diagnosis), which are, therefore, represented

by binary training target values (e.g., 0 and 1). However, histological reports often provide more detailed information on lesion subtypes than the malignant vs benign diagnosis. In this paper, we investigate a method that uses lesion histological subtypes in ANN training. We divide breast lesions into six histological subtypes, which correspond to six ANN training target values. We refer to this method as multitarget training and the method of using binary target values (e.g., 0 vs 1 which corresponds to benign vs malignant) as binary training. We compare these two training methods in terms of overall classification performance, which is measured by the area under the receiver operating characteristic (ROC) curve (AUC), and in terms of variability in ANN outputs.

II. BACKGROUND AND THEORY

II.A. Posterior probability as an optimal decision variable

In a two-class classification task, one can define a likelihood ratio (LR) that can be used as a decision variable to obtain the optimal classification performance, which can be measured with AUC.[11–13] Let $\boldsymbol{x}$ represent an arbitrary case (i.e., a vector of feature values), we define LR($\boldsymbol{x}$) as

$$\mathrm{LR}(\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\mathrm{positive})}{P(\boldsymbol{x}|\mathrm{negative})}. \tag{1}$$

According to the Bayes theorem, the posterior probability of $\boldsymbol{x}$ being positive is

$$\begin{aligned} &P(\mathrm{positive}|\boldsymbol{x}) \\ &= \frac{P(\boldsymbol{x}|\mathrm{positive})P(\mathrm{positive})}{P(\boldsymbol{x}|\mathrm{positive})P(\mathrm{positive}) + P(\boldsymbol{x}|\mathrm{negative})P(\mathrm{negative})} \\ &= \frac{P(\mathrm{positive})}{P(\mathrm{positive}) + P(\mathrm{negative})/\mathrm{LR}(\boldsymbol{x})}, \end{aligned} \tag{2}$$

where P(positive) and P(negative) are the prevalence of positive and negative cases, respectively, and are unknown constants. Because the posterior probability is a monotonic transformation of the LR, it is also an optimal decision variable.

II.B. Neural network prediction of posterior probability

Theoretically, the output of a classification ANN is an estimate of the posterior probability, provided that the number of training cases is sufficiently large and the ANN architecture is sufficiently "complex."[5,14,15] We describe briefly here the proof provided originally by Rojas.[14] For a set of n-dimensional feature vector s{$\boldsymbol{x}$} of ANN inputs, the space of the input data can be partitioned into a lattice of differential volume s{d$\boldsymbol{x}$}. Let N_{px} and N_{nx} denote the number of positive and negative training cases, respectively, in the differential volume d$\boldsymbol{x}$, then the sum-of-square error function of the ANN output y($\boldsymbol{x}$) can be written as

$$\Delta = \int [N_{px}(t_p - y(\boldsymbol{x}))^2 + N_{nx}(t_n - y(\boldsymbol{x}))^2] d\boldsymbol{x}, \tag{3}$$

where $t_p = 1$ and $t_n = 0$ are the training target values for positive and negative cases, respectively. Because the integrand is non-negative, minimizing the error function Δ is equivalent to minimizing the integrand. From setting the first derivative of the integrand with respect to y($\boldsymbol{x}$) to zero, one obtains the ANN output that minimizes the sum-of-square error

$$y(\boldsymbol{x}) = \frac{N_{p\boldsymbol{x}}t_p + N_{n\boldsymbol{x}}t_n}{N_{p\boldsymbol{x}} + N_{n\boldsymbol{x}}} = \frac{N_{p\boldsymbol{x}}}{N_{p\boldsymbol{x}} + N_{n\boldsymbol{x}}}, \tag{4}$$

which is the fraction of positive cases in the differential volume d$\boldsymbol{x}$. As the size of the training dataset becomes sufficiently large, the fraction of positive cases converges asymptotically to the posterior probability; therefore, the ANN outputs converge to posterior probabilities. Consequently, assuming that the ANN architecture is sufficiently "complex," ANN outputs do not show variability.[10] However, if the number of training cases is not sufficiently large, ANN output maybe a poor estimator of the posterior probability because the ratio $N_{px}/(N_{px} + N_{nx})$ in each differential volume d$\boldsymbol{x}$ cannot be guaranteed to equal the posterior probability. Furthermore, the common practice of limiting the number of hidden nodes to avoid over training also limits the ability of the ANN to minimize its cost function, i.e., the sum-of-square error. Therefore, with finite training data, ANNs often do not achieve the classification performance of the ideal observer[9] and, due to the stochastic nature of ANN training, ANNs trained from different initial random weights, or with different numbers of training epochs, will not converge to unique output values but, rather, exhibit variability in their outputs.[10]

We hypothesize that, with finite training data, ANN training using posterior probability as target values instead of the binary target values will help the ANN output to better approximate the posterior probability. If this is true, then it follows that ANNs trained with posterior probability as target values will in general produce better classification performance than ANNs trained with the binary target values. However, the obvious problem is that the posterior probability is generally unknown and, therefore, cannot be used to train the ANN. We observe that, in some situations, certain surrogate information that is strongly correlated with the posterior probability maybe available. For example, diagnostic information on histological subtypes of breast lesions may correlate more strongly with the continuous quantity of posterior probability than the binary malignant/benign designation. This information could be used to train ANNs. However, information on histological subtypes is usually discarded in CAD applications in favor of the binary malignant/benign training target values. In this paper, we study the performance of ANNs trained with six target values based on histological subtypes of breast lesions as presumed surrogate of the posterior probability and compare the performance of these ANNs with that of ANNs trained with the binary target values based on the malignant/benign diagnostic information. Note that we assume histological data correlate strongly with the posterior probability, but we do not assume that the histological data represent, or are good estimates of, the posterior probability. It is not possible to evaluate the latter assumption without access to the actual posterior probability. There are also cases in which histological data clearly do not agree with the posterior probability. For example, if a malignant case and a benign

case share the same feature values, then their posterior probabilities are identical, but their histological data are not. However, strong correlation between histological data and the posterior probability may reveal enough information contained in the posterior probability to improve ANN training.

III. MATERIALS AND METHODS

We studied two-class classification problems with both Monte Carlo simulations and an example classification task in a CAD application. In our simulation study, posterior probability can be calculated from the underlying distributions of positive and negative cases; therefore we trained ANNs with either the continuous quantity posterior probability or its discrete version as training target values. However, in our mammography CAD study, as in other real-world classification tasks, both the underlying distributions of the two classes of cases and the posterior probability are unknown. In this paper, we used six training target values based on lesion histological subtypes as presumed surrogates of the posterior probability.

We compared both overall classification performance and variability in the ANN output between the binary and multitarget training methods. Overall classification performance was measured with ROC analysis and with AUC as a summary index. Variability in ANN outputs that originates from the stochastic nature of the ANN training process was measured with the standard deviation of the outputs in a single given test case from eight ANNs with identical architecture and trained on a common set of cases but with different initial random weights, which was then averaged across all test cases.[10] We used three-layer feedforward and error-backpropagation neural networks,[16] which are used commonly in CAD applications. The number of input nodes was equal to the number of input features, the number of hidden nodes was chosen empirically, and there was a single output node. Generally, the number of hidden nodes increased as the number of input nodes increased. We use the notation n_1-n_2-n_3 to denote an ANN with n_1 input nodes, n_2 hidden nodes, and n_3 output nodes.

III.A. Simulation study

We studied two-class classification tasks. Examples of two-class classification tasks in medical diagnostic applications include classifying positive cases from negative cases for a particular disease and classifying malignant from benign breast lesions in mammograms. We carried out Monte Carlo simulations, in which simulated negative and positive cases were drawn randomly from two distinct and known distributions. We considered only the situation in which the training and testing cases contain equal numbers of positive and negative cases. The total number of training cases was varied between 10 and 2000, but the total number of test cases was fixed at 2000 to minimize the effect of the finite number of test cases on the estimation of classification performance.

We calculated the posterior probability according to Eq. (2) from the known underlying distributions of positive and negative cases. Because we simulated the situation in which positive and negative cases had equal prevalence, i.e., $P(\text{positive}) = P(\text{negative})$, it follows that for an arbitrary case $\boldsymbol{x}$:

$$P(\text{positive}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\text{positive})}{P(\boldsymbol{x}|\text{positive}) + P(\boldsymbol{x}|\text{negative})}. \quad (5)$$

We constructed training target values based on Eq. (5) as follows. Given a particular number of training target values, n, we partitioned the range of the posterior probability, i.e., the interval [0, 1] evenly into n equal subintervals: $[(i-1)/n, i/n]$, where $i = 1, 2, \ldots, n$. Then, for a training case with posterior probability within the subinterval $[(i-1)/n, i/n]$, we set the training target value to be $(i-1)/(n-1)$. Table I shows the training target values for $n = 5$. For the special case $n = \infty$, the continuous quantity posterior probability was used as training target values.

TABLE I. Assignment of five target values according to the posterior probability in the simulation study.

Range of posterior probability	0–0.2	0.2–0.4	0.4–0.6	0.6–0.8	0.8–1.0
Assigned target value	0.0	0.25	0.5	0.75	1.0

We simulated two types of classification tasks. In the first classification task, negative and positive cases follow multivariate isotropic normal distributions. Although normal distributions are idealized compared with real-world classification problems, they have been used previously in ANN simulation studies and their results often agree, at least qualitatively, with results of real-world ANN applications.[10,17,18] Furthermore, an advantage of these simple distributions is that one can calculate the theoretical performance of the ideal observer and compare that with the ANN performance.

We simulated both 2D and 8D input feature dimensions because 2D classification problems are easy to visualize and our mammography CAD classification task was an 8D classification problem. We simulated each dimension using a pair of multivariate isotropic normal distributions for negative and positive cases independently and, without loss of generality, we let negative cases follow the standard normal distribution $N(0, 1)$, and let positive cases follow a normal distribution $N(a/b, 1/b)$ with mean of a/b and standard deviation of $1/b$. Figure 1(a) shows random samples of 1000 cases drawn from such a pair of 2D normal distributions where $a = 1$ and $b = 1$. The ideal observer forms linear decision boundaries in these classification tasks. We used 8-6-1 and 2-2-1 ANNs for the 8D and 2D classifications tasks, respectively.

The second type of classification tasks that we simulated is known as the 2D exclusive or (XOR) classification problem. In the 2D Cartesian coordinate space of input feature data, positive cases were represented by the sum of two 2D isotropic normal distributions—one centered at $(-1, 1)$ and the other centered at $(1, -1)$, each with isotropic standard deviation of 1, whereas negative cases were represented by the sum of two other 2D isotropic normal distributions—one centered at $(1, 1)$ and the other centered at $(-1, -1)$, each with isotropic standard deviation of 1. Figure 1(b) shows random samples of 1000 cases from the 2D XOR distribution. The ideal observer decision boundary for this classification task

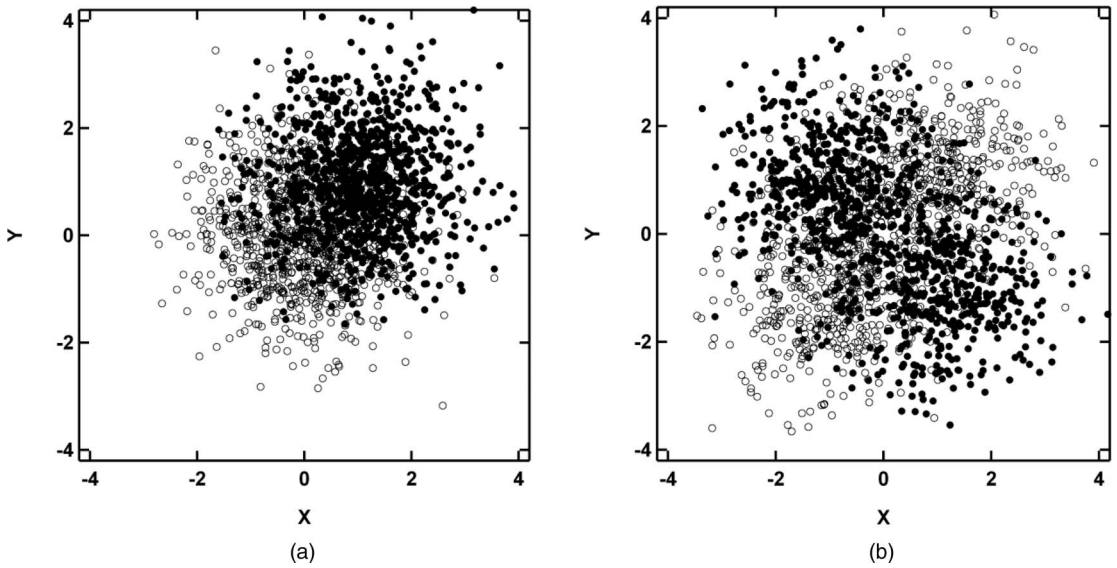

FIG. 1. Examples of (a) a 2D multivariate normal dataset and (b) a 2D XOR dataset. Solid symbols and hollow symbols are simulated positive and negative data points, respectively.

is quadratic. We used 2-10-1 ANNs for the XOR classification task, which was more complex than what we used for the 2D multivariate normal classification problem. Whereas the AUC value of the ideal observer was calculated analytically in the multivariate normal classification task, it was calculated empirically in XOR classification task from the posterior probability of 2000 randomly sampled test cases using the LABROC4 algorithm.[19]

III.B. Mammography study

A mammography study was also carried out to compare the binary ANN training method and the multitarget training method. The classification task was to classify clustered microcalcifications on mammograms as malignant or benign based on eight image features.[3] As in other real-world classification problems, the underlying distributions of positive and negative cases were unknown and, therefore, the posterior probability could not be calculated. Therefore, we did not use posterior probability as training target values. Instead, we used the diagnostic information of histological subtype of lesions as presumed surrogates of the posterior probability. Specifically, we divided malignant diagnosis into four histological subtypes: noncomedo ductal carcinoma *in situ* (DCIS), comedo DCIS, DCIS with microinvasion, and invasive carcinoma (including invasive ductal and invasive lobular carcinomas); and divided benign diagnosis into two histological subtypes: benign and benign with high risk (including atypical ductal hyperplasia and lobular carcinoma *in situ*). Thus, we used six training target values that correspond to six histological subtypes. This sequence of histological subtypes is one (of more than one) possible representation of an ordered spectrum from benign tissue in one extreme to aggressive cancer in the other.[20] The mammogram image dataset we used consisted of 96 cases, of which 42 cases were malignant and 54 were benign. Every case had a mediolateral oblique (MLO)-view and a craniocaudal (CC)-view mammograms. Table II lists the number of cases of each histological subtype together with training target values for both the binary and the multitarget training method.

We used an 8-6-1 ANN architecture as in our previous study.[21] We used the jackknife method in which both malignant and benign cases were partitioned randomly into two halves: one half (21 malignant and 27 benign cases) was used for training and the other half (21 malignant and 27 benign cases) for testing. All mammograms of a patient were kept as a unit for data partitioning to ensure true independence between the training and test cases. Further, we partitioned the cases independently 20 times. For each jackknife partition, eight ANNs with identical architecture were trained with arbitrarily different initial random weights. The average ANN performance was estimated by averaging over the 20 jackknife partitions of the mean AUC value of the eight ANNs in each jackknife partition. ANN output variability (i.e.,

TABLE II. Assignment of target values for the binary training and multitarget ANN training in the mammography study.

Histological subtype	Benign	Benign with high risk	Noncomedo DCIS	Comedo DCIS	DCIS with microinvasion	Invasive carcinoma
Number of cases	50	4	20	13	3	6
Target values of multitarget method	0.1	0.2	0.6	0.7	0.8	0.9
Target values of binary training method	0.1	0.1	0.9	0.9	0.9	0.9

standard deviation in the eight ANN output values on a particular test case) was compared with Student *t*-test for paired data on the test cases for an arbitrarily chosen jackknife partition.

To further understand the effect of multitarget training, we also trained ANNs using three patients (i.e., six mammograms) as the training cases for each histology subtype. Other cases were used as test data and the results of binary training and multitarget training were compared. Exactly as in the jackknife experiment, eight ANNs of the 8-6-1 architecture were trained to the same epochs from different initial random weights, the average AUC values of the eight ANNs were used to characterize the classifier performance, and the standard deviation of the eight ANN outputs in a single given test case was calculated to characterize ANN output variability. Compared with the jackknife experiment, the number of training cases was uniform across all histology subtypes. The CLABROC algorithm[22] was used to compare the AUC values between the two training methods.

IV. RESULTS

IV.A. Simulation results

Figure 2 shows the AUC values of ANNs trained with the multitarget training method and the binary training method. The numbers of training cases were 100 positive and 100 negative cases. The AUC values of the ideal observers were 0.8413 for both the 2D and 8D multivariate normal classification tasks and 0.8061 for the XOR classification task. Classification performance of ANNs in the 2D multivariate normal classification task is slightly closer to that of the ideal observer than that in the 8D multivariate normal classification task, probably because the ANN architecture in the 8D task (8-6-1) is more complex than in the 2D task (2-2-1), thus, requiring more training cases than in the 2D task to approach the ideal-observer performance. With a greater number of training cases than those shown in Fig. 2, the 8-6-1 ANN trained with posterior probabilities achieved AUC value of 0.839 with 400 training cases, whereas the same ANN trained with the conventional method needed 2000 training cases to achieve the same classification performance. Figure 2 also shows that multitarget training provides greater benefit to classification performance (i.e., the AUC values) for the XOR dataset than for the multivariate normal datasets, probably because the optimal classifier for the XOR dataset (quadratic) is more complicated than that for the multivariate normal dataset (linear) and, thus, relatively more difficult to train and easier to see the effect of a better training method.

Standard deviations of the AUC values of the eight ANNs that were trained with different initial random weights were on the order of 0.0001 or 0.001 (not shown in Fig. 2). However, variability in ANN output reduced markedly from ANNs trained with the binary training method to ANNs trained with the multitarget training method. For the binary training method, standard deviation of ANN output on individual cases was on the order of either 0.01 or 0.1, and the standard deviation averaged across 2000 test cases was just under 0.1. ANN output variability of this magnitude is not negligible considering the range of ANN output is from 0.0 to 1.0. Standard deviations of outputs of ANNs trained with the multitarget training method were smaller as shown in Fig. 3. Furthermore, the standard deviation decreased as the number of training target values increased (Fig. 3).

Note that in Figs. 2 and 3, there are corresponding data points of both solid symbols that represent multitarget

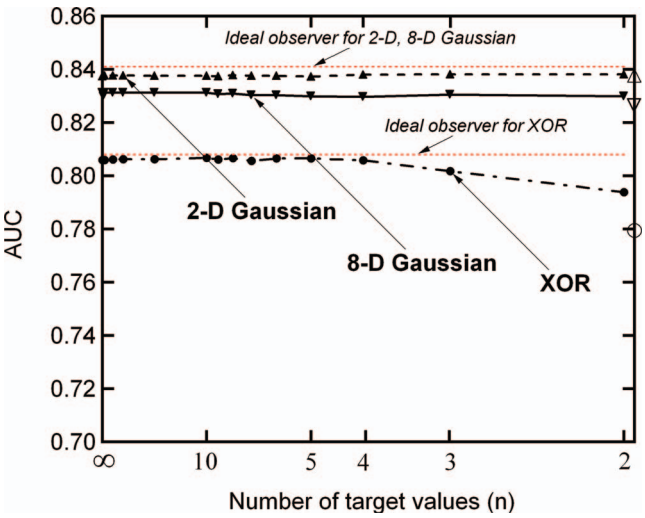

FIG. 2. ANN classification performance in simulation studies measured with AUC values of the multitarget training method (solid symbols) and the binary raining method (hollow symbols). Standard deviations of AUC values from eight ANNs with identical architecture but trained with different initial random weights (not shown in this figure) are on the order of 0.0001 or 0.001. The standard errors of the AUC value of each individual ANN, which reflects the number of test cases, are approximately 0.01 (also not shown in this figure). The numbers of training cases were 100 positive and 100 negative cases. The scale of the horizontal axis is 1/n.

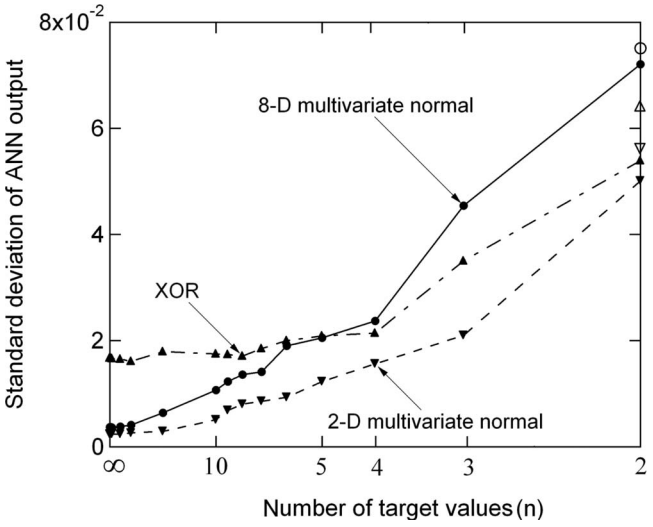

FIG. 3. Variability in ANN outputs, characterized by the standard deviation in a single given test case of the outputs of eight ANNs trained identically but from different initial random weights, which was subsequently averaged across all test cases, is shown for the multitarget training method (solid symbols) and the binary training method (hollow symbols) in simulation studies. The numbers of training cases were 100 positive and 100 negative cases. The scale of the horizontal axis is 1/n.

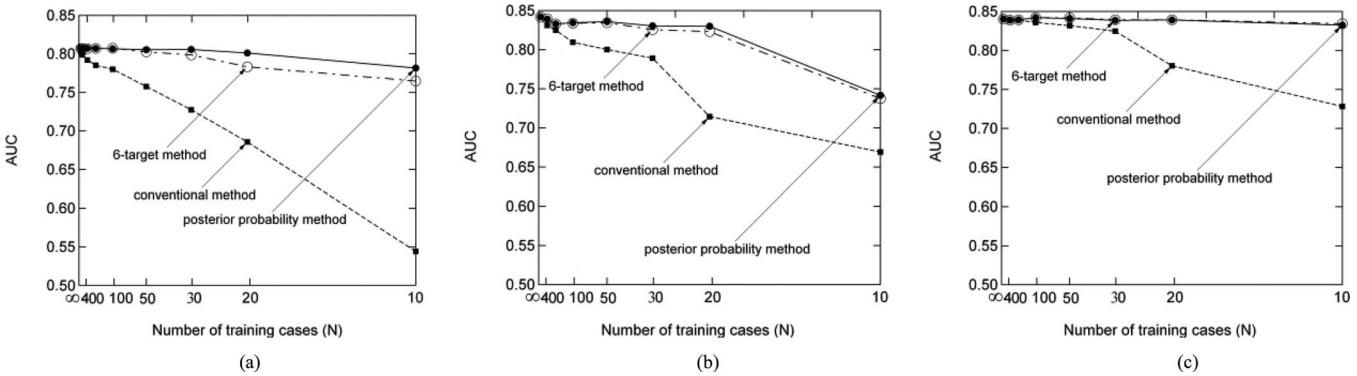

FIG. 4. Dependence of ANN performance (AUC) on the number of training cases for binary ANN training, six-target training, and posterior probability training in the (a) XOR, (b) 8D multivariate normal, and (c) 2D multivariate normal simulation studies. The standard errors of the AUC value of each individual ANN (not shown), which reflect the number of test cases, are approximately 0.01. The scale of the horizontal axis is 1/N.

training and hollow symbols that represent the binary training at $1/n = 0.5$, that is $n = 2$. The difference between these corresponding data points is that the binary training uses the binary truth as target values whereas multitarget training when $n = 2$ uses two discrete target values that represent the continuous quantity of posterior probability. For example, for a positive training case with posterior probability of less than 0.5, binary training will assign a target value of 1, whereas multitarget training method will assign a target value of 0. Similarly, the binary training method will assign a target value of 0 for a negative training case with posterior probability greater than 0.5, whereas the multitarget training method will assign a target value of 1. It is interesting to note that the multitarget training method with just two target values produces better overall classification performance and less variability in ANN output than the binary training method (Figs. 2 and 3).

Figure 4 shows the effect of the number of training cases on the overall classification performance of ANNs trained with three different sets of training target values in the XOR, 8D multivariate normal, and 2D multivariate normal classification task. In general, classification performance of ANNs increased as the number of training cases N increased, eventually reaching the classification performance of the ideal observer. However, compared with ANNs trained with the binary training method, ANNs trained with the six-target method or with the posterior probability method achieved higher classification performance with smaller numbers of training cases, as one can see from Fig. 4 for the XOR, 8D multivariate normal, and 2D multivariate normal studies. One can also see that the simpler classifiers for 2D multivariate normal dataset needed fewer training cases to approach the performance of the ideal observer.

IV.B. Results of mammography study

The average AUC values across 20 jackknife resamplings of cases from the two training methods were similar: 0.777 for the binary method and 0.778 for the six-target method. However, variability in ANN output was different between the two training methods. Standard deviation in ANN output in individual test cases, averaged across all test cases and then averaged over the 20 jackknife partitions, was 0.083 for the binary training method, and 0.045 for the six-target method, which was a reduction of approximately one half. Figure 5 compares the mean, minimum, maximum, and 25th and 75th percentiles of the ANN output standard deviation from each jackknife partition and from the binary and the six-target training methods, and evidently shows reduced ANN output variability from the six-target training method. For an arbitrarily chosen jackknife partition, we found a standard deviation of 0.084 for the binary training method and 0.045 for the six-target training method, which differ significantly ($p < 10^{-15}$, Student t-test for paired data).

Training ANNs with three patients in each histology subtype (i.e., 18 patients total) resulted in an AUC value of 0.697 with binary training and 0.723 with the six-target training. The improvement in the AUC values was 0.026, smaller than the estimate standard errors of the AUC values (0.05), and was not statistically significant ($p = 0.78$). The standard deviation in ANN outputs in individual test cases, averaged across all test cases, was 0.082 and 0.043 for binary training and six-target training, respectively, which are similar to the results of the jackknife study.

V. DISCUSSION

The binary ANN training method seeks to minimize the cost function: $\sum_{i=1}^{N_p}(t_p - y_{pi})^2 + \sum_{j=1}^{N_n}(t_n - y_{nj})^2$, where t_p and t_n are, respectively, the target values for positive and negative cases, and y_{pi} and y_{nj} are the ANN outputs for positive and negative cases. In ANN training, a positive and a negative training case that have identical feature values will produce identical ANN output, but they will have different target values t_p and t_n in the optimization scheme above. In this particular situation, minimizing the contribution of the positive case to the cost function will maximize the contribution of the negative case to the cost function, and vice versa. This situation makes ANN training difficult because the ANN must find on its own an appropriate balance that minimizes the combined contribution to the cost function from both the positive and the negative cases, which clearly differs from what minimizes the contribution to the

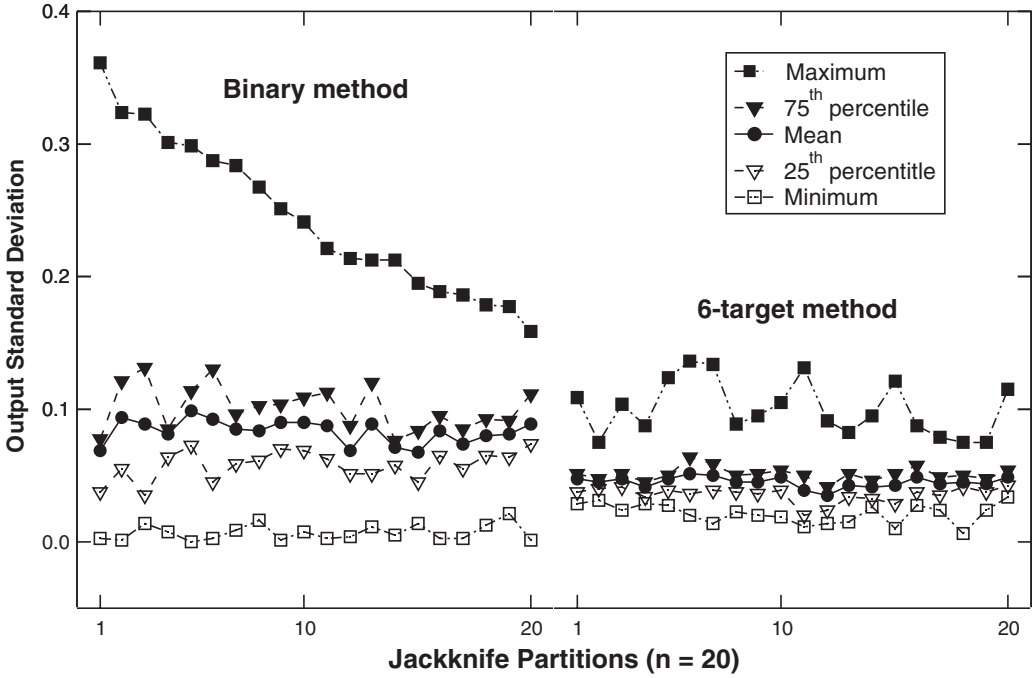

FIG. 5. Comparison between binary and six-target training methods across 20 jackknife partitions in the mammography study of the mean, minimum, maximum, and 25th and 75th percentiles of ANN output standard deviations on individual test cases. Reduced ANN output variability is evident from the six-target training method. For better legibility, corresponding jackknife partitions between the binary and six-target training methods are shown in decreasing order of the maximum output standard deviations of the binary training method.

cost function from either one of the two cases. When posterior probability is used as training target values, the cost function is: $\sum_{i=1}^{N_p}(P_{post.i}-y_{pi})^2+\sum_{j=1}^{N_n}(P_{post.j}-y_{nj})^2$. With this method, a positive and a negative training case that have identical feature values will also correspond to identical training target values. Consequently, minimizing the contribution to the cost function from one case also simultaneously minimizes the contribution from the other case. Therefore, at least with regard to the situation discussed above, training with posterior probability as target values is potentially better than the binary training method.

Consider the ANN training cost function within a differential volume of the input feature space, $s\{d\boldsymbol{x}\}$, the cost function of binary training can be written as: $\sum_{i=1}^{n}(t_i-y)^2$, where t_i is the training target values ($t_i=1$ for positive and $t_i=0$ for negative cases), n is the total number of training cases in $s\{d\boldsymbol{x}\}$, and y is the ANN output for all cases in $s\{d\boldsymbol{x}\}$. Clearly, t_i is a Bernoulli variable with expected value of $P_{post.}$ and variance of $P_{post.}(1-P_{post.})$. With binary training, the ANN tries to minimize $\sum_{i=1}^{n}(t_i-y)^2$ and, when successful, produces $y=(\sum_{i=1}^{n}t_i)/n=t_{mean}$, which is another random variable also with expected value of $P_{post.}$ but variance of $P_{post.}(1-P_{post.})/n$. That is to say, binary training produces estimates of the posterior probability with a variance of $P_{post.}(1-P_{post.})/n$. One needs to increase the size of the training cases to reduce this variance. However, with posterior probability training, the cost function is: $\sum_{i=1}^{n}(P_{post.}-y)^2$, and, when this cost function is successfully minimized, the ANN produces simply $y=P_{post.}$, even for $n=1$. As the number of training cases increases, the binary training method improves and the difference between the two training methods reduces. However, one can still expect better ANN training results from using posterior probability as the training target, as Fig. 4 shows.

It is plausible that multitarget training is beneficial to ANNs in two-class classification tasks because they allow more information to be incorporated into the training process. In the simulation study, the underlying distributions of positive and negative cases were known, thus the posterior probability was also known. Therefore, we were able to use the posterior probability as training target values, which represent substantially more information than the binary target values of the binary training method. In practice, we generally have no access to the posterior probability and, therefore, cannot use it as training target values. However, in some situations, information in addition to the binary "truth" is available. For example, histological reports on breast lesions typically contain more detailed diagnostic information than simply the binary malignant versus benign diagnosis. We grouped the more detailed histological diagnosis information into six histological subtypes, which we presume that, because they are correlated with the natural progression of the aggressiveness and disease burden of breast cancer, they are correlated with the posterior probability of a lesion being malignant. Note, however, we do not assume that the histological subtypes correlate with the posterior probability perfectly; in certain cases, the histological subtypes clearly do not agree with the posterior probability. Although our grouping of the histological subtypes and our assignment of the training target values are *ad hoc*, it is reasonable to expect that the incorporation of the presumed

correlation between the histological subtypes and the true, unknown, posterior probability into ANN training would help improve the results of ANN training compared with the binary training method. There may be also other ways to group the histologic data for multitarget ANN training purposes.

In our mammography study, the ANN produced separate outputs for MLO- and CC-view mammograms. These two ANN outputs in a given patient need to be combined into a single estimate of the likelihood of malignancy of the lesion. Liu *et al.*[23] showed that each of the methods of taking the average, the maximum, or the minimum can produce the best overall classification performance, depending on the shape of the single-view mammogram ROC curve. However, in this study, because the shape of the single-view mammogram ROC curve may not be consistent across different jackknife resamplings of the cases, it is unclear which method is optimal for merging the two ANN outputs. Therefore, for simplicity and to be consistent, we took the maximum when merging all ANN outputs from two views of a patient. Although this method may not have produced the best overall classification results, the use of this method should not affect the comparison between the multitarget and the binary training methods, because taking the maximum ANN outputs of the two-view mammograms was used in both training methods.

Whereas our simulation and mammography studies showed that multitarget training can reduce variability in ANN outputs and the simulation studies also showed improved overall classification performance from multitarget training, our jackknife mammogram study showed no AUC gain for the multitarget training method. This is probably due to the limited number of cases in the mammography study, thus limited number of test cases, which yielded large uncertainty with standard deviation approximately 0.05 in the AUC values of the ANN classification performance, making it difficult to find any difference in the AUC values between the two training methods. However, our mammogram study using a small training dataset (18 patients) showed a nonstatistically significant AUC gain of 0.026 for multitarget ANN training over binary ANN training. Note that a small number of test cases make it difficult to show statistically significant gain in the AUC value whereas a small number of training cases may show the effect of multitarget training more clearly. In our simulation studies, we have 2000 test cases and a variable number of training cases, which was the best of both worlds. But compromise is unavoidable with the mammography dataset as in other real-world datasets.

One possible approach to further investigate the effect of the number of training and test cases in our mammography study is to repeat it multiple times with varying numbers of training and test cases, and analyze the trend in the results as a function of the number of training and test cases. In our study, we conducted two experiments: in one of which half of the total 96 cases were used for training and the other half used for test, and in another 18 cases were used for training and the rest used for test. It is possible to conduct additional repeat experiments with intermediate numbers of training and test cases, but these were not done in our study. Although with small numbers of training and test cases the results may be inconclusive due to large uncertainties in the observed results, if uncertainties reduce as the numbers of training and test cases increase, then a trend may emerge that better reveals the effect that we intended to find. We expect this approach more likely to produce convincing results if we had a larger total number of cases. Because the limitations in the number of cases, our mammography study serves only as an example of how training ANNs with multiple target values could be applied to practical classification tasks, and not a validation of it.

The variability in the ANN output that we calculate is caused by the stochastic nature of ANN training. Fixing the ANN weights at the end of training masks this variability but does not eliminate it.[10] Although this variability in the ANN output may be secondary to the overall accuracy of the ANN in many situations, this variability in the ANN output can be important if a fixed decision threshold is applied to the ANN output in the vicinity of large variability in that output, and in situations in which the ANN output is interpreted by human observers.[3,4] Thus, reducing this variability in the ANN output with multitarget training is desirable.

Our analysis of theoretical considerations and empirical studies with simulated datasets supports strongly benefits of training ANNs with multiple target values, and our empirical studies with mammograms show limited support (due to limited number of cases) of training ANNs with histological subtypes as multiple target values. Our proposal of using histological subtypes as multiple training target values is based in large part on the empirical correlation between histology subtypes and cancer progression and, by extension, the presumed correlation between histological subtypes and cancer posterior probability. This study did not focus on providing strong empirical evidence in support of that presumed correlation, and further studies with larger numbers of cases are needed for validation.

VI. CONCLUSION

For two-class classification tasks, our simulation study shows that ANNs trained with the multitarget method and the binary training method produce similar overall classification performance if the number of training cases is large, whereas the multitarget method can produce better overall classification performance with smaller numbers of training cases. Both our simulation and mammography studies show that ANNs trained with the multitarget method produce less variability in ANN output in individual cases than ANNs trained with the binary training method. Therefore, the multitarget training method is potentially useful for ANN applications in computer-aided diagnosis for breast cancer.

ACKNOWLEDGMENTS

This work was supported in part by National Cancer Institute/National Institutes of Health (NCI/NIH) through Grant Nos. R21-CA93989 and R01-CA092361, and by U.S. Army Medical Research and Materiel Command through Grant No. DAMD17-00-0197.

[a]Author to whom correspondence should be addressed. Electronic mail: beiliu@usc.edu

[1]J. A. Baker, P. J. Kornguth, J. Y. Lo, and C. E. J. Floyd, "Artificial neural network: Improving the quality of breast biopsy recommendations," Radiology **198**, 131–135 (1996).

[2]H. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: Texture analysis using an artificial neural network," Phys. Med. Biol. **42**, 549–567 (1997).

[3]Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," Acad. Radiol. **6**, 22–33 (1999).

[4]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).

[5]M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate Bayesian *a posteriori* probabilities," Neural Comput. **3**, 461–483 (1991).

[6]H. H. Berry, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality: III. ROC metrics, ideal observers and likelihood-generating functions," J. Opt. Soc. Am. A **15**, 1520–1535 (1998).

[7]K. Swingler, *Applying Neural Networks: A Practical Guide* (Academic, London, 1996).

[8]M. J. A. Berry and G. Linoff, *Data Mining Techniques* (Wiley, New York, 1997).

[9]H. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," Med. Phys. **26**, 2654–2668 (1999).

[10]Y. Jiang, "Uncertainty in the output of artificial neural networks," IEEE Trans. Med. Imaging **22**, 913–921 (2003).

[11]J. Egan, *Signal Detection Theory and ROC Analysis* (Academic, New York, 1975).

[12]J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," Invest. Radiol. **14**. 109–121 (1978).

[13]C. E. Metz, "ROC methodology in radiologic imaging," Invest. Radiol. **21**, 720–733 (1986).

[14]R. Rojas, "A short proof of the posterior probability property classifier neural networks," Neural Comput. **8**, 41–43 (1996).

[15]D. E. 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 Netw. **1**, 296–298 (1990).

[16]D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in Microstructure of Cognition Volume I: Foundations*, Computational Model of Cognition and Perception Vol. 1, edited by D. E. Rumelhart, J. L. McClell, and T. P. R. Group (MIT, Cambridge, MA, 1986), pp. 318–362.

[17]S. V. Beiden, R. F. Wagner, and G. Campbell, "Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects, receiver operating characteristic analysis," Acad. Radiol. **7**, 341–349 (2000).

[18]S. V. Beiden, R. F. Wagner, G. Campbell, and H. Chan, "Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis," Acad. Radiol. **8**, 616–622 (2001).

[19]C. E. Metz, B. A. Herman, and J. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat. Med. **17**, 1033–1053 (1998).

[20]P. P. Rosen, *Rosen's Breast Pathology* (Lippincott Williams & Wilkins, Philadelphia, Pennsylvania, 2001).

[21]Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: Automated feature analysis and classification," Radiology **198**, 671–678 (1996).

[22]C. E. Metz, P. L. Wang, and H. B. Kronman, "*A new approach for testing the significance of differences between ROC curves measured from correlated data*," in *Information Processing in Medical Imaging*, edited by F. Deconinck (Martinus Nijhoff, The Hague, 1984), pp. 432–445.

[23]B. Liu, C. E. Metz, and Y. Jiang, "An ROC comparison of four methods of combining information from multiple images of the same patient," Med. Phys. **31**, 2552–2563 (2004).