Ultrasound in Med. & Biol., Vol. 41, No. 12, pp. 3148–3162, 2015

0301-5629/$ - see front matter

http://dx.doi.org/10.1016/j.ultrasmedbio.2015.07.020

● *Original Contribution*

# GOING BEYOND A FIRST READER: A MACHINE LEARNING METHODOLOGY FOR OPTIMIZING COST AND PERFORMANCE IN BREAST ULTRASOUND DIAGNOSIS

Santosh S. Venkatesh,* Benjamin J. Levenback,† Laith R. Sultan,† Ghizlane Bouzghar,† and Chandra M. Sehgal†

*Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA; and †Department of Radiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

(*Received* 27 *February* 2015; *revised* 16 *June* 2015; *in final form* 16 *July* 2015)

**Abstract—The goal of this study was to devise a machine learning methodology as a viable low-cost alternative to a second reader to help augment physicians' interpretations of breast ultrasound images in differentiating benign and malignant masses. Two independent feature sets consisting of visual features based on a radiologist's interpretation of images and computer-extracted features when used as first and second readers and combined by adaptive boosting (AdaBoost) and a pruning classifier resulted in a very high level of diagnostic performance (area under the receiver operating characteristic curve = 0.98) at a cost of pruning a fraction (20%) of the cases for further evaluation by independent methods. AdaBoost also improved the diagnostic performance of the individual human observers and increased the agreement between their analyses. Pairing AdaBoost with selective pruning is a principled methodology for achieving high diagnostic performance without the added cost of an additional reader for differentiating solid breast masses by ultrasound. (E-mail: sehgalc@uphs.upenn.edu)** **© 2015 World Federation for Ultrasound in Medicine & Biology.**

***Key Words:*** **Breast ultrasound, Breast cancer, Adaptive boosting, Computer-aided diagnosis, Artificial intelligence.**

## INTRODUCTION

Considerable effort is being devoted to improve breast ultrasound for differentiating solid malignant and benign masses (Candelaria et al. 2013; Sehgal et al. 2006). Progress toward this goal is being made *via* the integration of ultrasound and mammography imaging modes (Padilla et al. 2013) and the introduction of new modes of ultrasound imaging like elastography (Chang et al. 2013; Golatta et al. 2013), 3-D imaging (Cho et al. 2005, 2006; McDonald 2011; Ruiter et al. 2012; Watermann et al. 2005) and computer-aided tomography (Duric et al. 2007). These technological developments have spurred the evolution of new computer-based algorithms to assist radiologists with breast cancer diagnosis using clinical ultrasound. These studies have been reviewed (Huang 2009; Sehgal et al. 2006). More recent investigations have extended the use of computer features for automated mass detection and classification in three dimensions with ultrasound images (Cheng et al. 2010; Liu et al. 2010). The state of the art is, however, still not entirely satisfactory; despite these advances in both breast imaging technology and image analysis, the biopsy yield continues to be low, and as many as 70% to 85% of biopsies prove to be benign. The costs of this low yield are the emotional trauma experienced by patients whose masses are ultimately determined to be benign and the socioeconomic costs to society as a whole imposed by a large number of unneeded procedures (Kopans 1992). One of the main reasons for such low yield is that the false negatives have major consequences related to patient mortality. Improving the accuracy of prediction to reduce the number of unneeded biopsies while keeping the false-negative rates to a minimum, possibly approaching zero, continues to be an important objective in the state of the art.

Several studies have indicated that multiple readings of mammograms improve diagnostic performance for breast cancer diagnosis (Georgian-Smith 2007; Gromet 2008; Taylor and Potts 2008; Waldmann et al. 2012). Waldman et al. reported that double reading of

Address correspondence to: Chandra M. Sehgal, Department of Radiology, University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104, USA. E-mail: sehgalc@uphs.upenn.edu

mammograms increased the tumor detection rate from 14.6 to 16.4 per 1000 cases. On the basis of their results they concluded that double reading is crucial not only for screening, but also for lesion characterization. Gromet also observed that with the double reading process, sensitivity rose from 81.4% to 88.0%, with a gain of 8.2% cancers detected. Similarly, Taylor et al. reported that double reading with arbitration increased the detection rate and decreased the recall rate. Experience with mammography suggests that the benefits of double or multiple readings can be expected to extend to ultrasound imaging. A recent study found that double reading of breast ultrasound improved diagnostic performance (Bouzghar et al. 2014). Unfortunately, it is not a simple matter to deploy multiple readers in clinical settings. As may be expected, the generation of multiple readings is labor intensive, costly, time consuming and limited by the scarcity of specialized radiologists skilled in interpreting breast ultrasound images. It is in this context that the advent of automated methods for feature extraction provides a viable real-time, low-cost alternative to a second reader to help augment physicians' interpretations of images.

The availability of this computerized resource opens the door to a principled combination of visual and machine-generated features using an adaptive machine learning procedure that incorporates the strengths of each feature set, while simultaneously identifying the small set of cases for whom the images are intrinsically ambiguous and merit further evaluation by additional imaging. In this study we describe such a computer-based system to complement a radiologist's interpretation of the ultrasound Breast Imaging Reporting and Data System (BI-RADS$_{US}$) as a second reader. The automated process of machine generating a second feature set is not only a low-cost procedure, but has the added advantage of generating features that are independent of those generated visually by a trained radiologist in the sense that the features extracted are qualitatively different. The computerized system combines *a priori* information and expert human knowledge in Bayesian settings with the logistic regression probability of computerized features to improve diagnostic decisions.

Next, we provide an overview of methodologic tools and algorithms. Then, we describe the results obtained by using the algorithms and patient data. We discuss the role of adaptive boosting and selective pruning in reducing cost and improving diagnostic performance and make our conclusions.

## METHODS AND ALGORITHMS

### *Overview*

Two independent feature sets representing readers 1 and 2 were constructed and classified for each image in the library of breast ultrasound images using a radiologist's interpretation of ultrasound BI-RADS$_{US}$ and computerized features extracted from the images (Fig. 1, step a). The radiologist's interpretations of ultrasound images (BI-RADS$_{US}$) were combined with the computer-generated features using *adaptive boosting* or the consensus method (Fig. 1, step b). This process was implemented to expand the discriminatory region of each feature set by incorporating the strengths of each set. Despite combining the regions of strength of each of the two independent feature sets, some cases remained persistently ambiguous. These cases representing the low-confidence group were pruned from the data set for further evaluation by additional imaging (Fig. 1, step c). The remaining cases, representing the high-confidence group, were evaluated for their diagnostic performance (Fig. 1, step d).

### *General methods*

Ultrasound images of 264 solid masses from 246 patients were analyzed for this study with approval

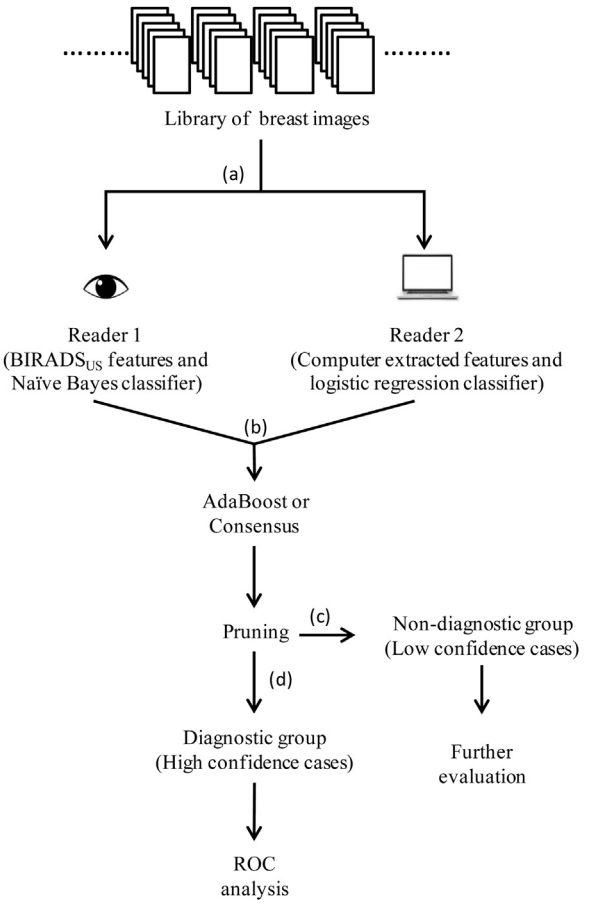


Fig. 1. Overview of the methods and procedures. BIRADS$_{US}$ = ultrasound Breast Imaging Reporting and Data System, ROC = receiver operating characteristic.

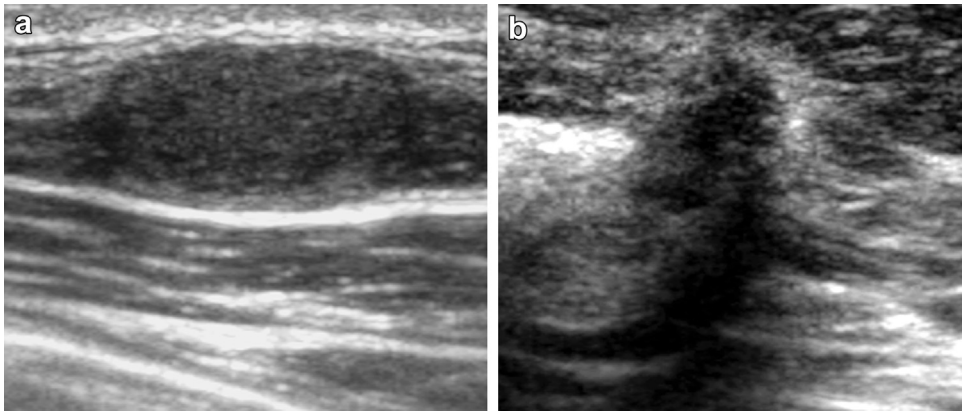


Fig. 2. (a) Oval-shaped fibroadenoma with benign characteristics, including circumscribed margin, parallel orientation, iso-echogenicity and posterior acoustic enhancement. (b) Infiltrative ductal carcinoma with malignant characteristics, including spiculated margin, irregular border, taller-than-wide non-parallel orientation, marked hypo-echogenicity and strong posterior acoustic shadow.

from the institutional review committee. The sonographic images for each mass consisted of five to seven views of the lesion in radial and anti-radial planes. In Figure 2 are examples of breast masses with malignant and benign characteristics. Based on a separate mammographic examination, each of the patients in the study had also undergone a biopsy. Of the 264 masses, 179 and 85 were designated as benign and malignant, respectively. The biopsy outcomes provided labels for each image in the study, these labels representing the ultimate ground truth. These biopsy-generated labels constitute the de facto gold standard in the training and testing of machine learning algorithms. Among the malignant lesions were invasive carcinoma (76%), invasive lobular carcinoma (8%), ductal carcinoma *in situ* (8%) and adenocarcinoma (4%). The remaining 4% were mixed including poorly differentiated carcinomas and mucinous mammary carcinoma. Of the benign masses, 44% were fibroadenomas, 33% were miscellaneous fibrocystic changes, 6% were sclerosing adenosis and the remaining 17% were benign lesions without atypia in the histopathology report. The mean age of the patients studied was 51.5 ± 14.7 y.

*Feature extraction*

Two qualitatively different types of features were extracted from sonograms. The first type consisted of ultrasound BI-RADS$_{US}$ features that are used routinely for evaluating suspicious breast lesions in breast ultrasound images. BI-RADS$_{US}$ is the standardized lexicon proposed by the American College of Radiology (ACR) for reporting and characterizing breast masses (ACR 2003). Images were reviewed independently by two physicians (B.G. and L.S.) using ACR guidelines (ACR 2003). The reviewers were not privy to other patient information including age, medical history and biopsy results. The BI-RADS$_{US}$ features consisted of 10 different features that characterize margin properties, echo patterns, posterior shadowing and enhancement of the lesions (ACR 2003; Bouzghar et al. 2014).

The second set of features was extracted by an automated procedure from manual tracings. For each breast ultrasound image reviewed by the physicians, the lesion was manually traced on a computer display using a mouse. Eight features describing gray scale, shape and coarseness of the margin were automatically computed from the traced margin. These features were extracted by partitioning the lesion into $N$ sectors and then comparing the difference between the inside and the outside of each sector (Sehgal et al. 2004, 2012). The features used in the analysis included the brightness difference between the lesion interior and immediate exterior, margin sharpness, angular variation in brightness, depth-to-width ratio, axis ratio, tortuosity, radius variation and elliptically normalized skeleton.

In addition to ultrasound image features, patient age and mammographic BI-RADS$_{US}$ category (1–5 representing probabilities of increasing malignancy) were also included in the analysis.

*Classifier selection tuned to the feature sets*

We model an underlying chance process generating lesions in which $M$ represents the event that a lesion is malignant, and its complement $B$ represents the event that it is benign. From the image of each lesion, a set of features $F = F_1, F_2, \ldots, F_N$ are extracted. These extracted features constitute the "measurements" representing the characteristics of the underlying lesions.

*Naïve Bayes classifier.* In the naïve Bayes formulation the features are assumed to be conditionally independent (hence the appellation "naïve"). In the case of nominal features taking discrete values this can be expressed mathematically as

$$P(F_1,\ F_2, \ldots,\ F_N \mid M) = P(F_1 \mid M) \times P(F_2 \mid M) \times \cdots \times P(F_N \mid M) \quad (1)$$

where an expression of the form $P(\ \cdot\ \mid\ \cdot\ )$ stands for a conditional probability, with $P(\cdot)$ representing an unconditional probability. By Bayes's rule, this means that the *a posteriori* probability of malignance is given by

$$P(M \mid F) = \frac{1}{Z} P(M) \prod_{j=1}^{N} P\left(F_j \mid M\right) \quad (2)$$

where, by total probability, we may write the normalizing constant $Z$ representing the unconditional probability of observing the given features in the form

$$Z = P(F) = P(F \mid M)P(M) + P(F \mid B)P(B) \quad (3)$$

The quantities $P(M)$ and $P(B)$ constitute *a priori* probabilities in the Bayesian settings of malignant and benign cases. These probabilities are not known ahead of time, but are estimated in the usual way by the relative frequency of occurrence (malignancy) in the training data set.

Although the naïve Bayes procedure makes strong independence assumptions on the feature set, it has turned out to be remarkably robust and effective in practice, especially in settings where the dimensionality is high (Hastie et al. 2009). The nominal nature of the BI-RADS$_{US}$ features, the relatively high dimensionality, the presumption of a degree of independence across features and the low complexity of the classifier guided our selection of this classifier to process the BI-RADS$_{US}$ data. Training and testing were performed by leave-one-out cross-validation.

*Logistic regression classifier.* Logistic regression models arise in settings where features take a continuum of values in a finite-dimensional space out of a desire to have the *a posteriori* class probabilities expressible in terms of simple linear functions of the features. Formally, the logarithm of the odds ratio is assumed to be in the form

$$\log \frac{P(M \mid F)}{P(B \mid F)} = c_0 + \sum_{j=1}^{N} c_j F_j \quad (4)$$

which, in view of the fact that $P(B \mid F) = 1 - P(M \mid F)$, we may express as

$$P(M \mid F) = \frac{1}{1 + \exp(-z(F))} \quad (5)$$

where $z(F)$ is given by the linear form

$$z(F) = c_0 + \sum_{j=1}^{N} c_j F_j \quad (6)$$

The virtues of the logistic regression classifier lie in its great simplicity—only the linear parameters $c_0, c_1, \ldots, c_N$ need to be evaluated—and its ready interpretability. For these reasons the classifier has enjoyed sustained popularity in medical diagnosis (Hastie et al. 2009). The continuous nature of the computer features guided our selection of this classifier to process the computer-derived data. Similar to BI-RADS$_{US}$, data training and testing were performed by leave-one-out cross-validation.

*Classifier combination* via *AdaBoost*

Given two classifiers, how best should they be combined? Although many *ad hoc* approaches have been advocated over the years, most of these have been superseded by a mathematically elegant approach developed by Freund and Schapire (1997). The procedure, commonly known as AdaBoost in the machine learning literature (short for adaptive boosting), was originally developed to systematically combine a family of weak learning algorithms, each of which did only slightly better than random guessing, to form a strong learning algorithm with very high performance levels. In our setting, the individual classifiers are in themselves quite strong (Bouzghar et al. 2014; Chen et al. 1999; Horsch et al. 2004; Sahiner et al. 2004), with a performance level of 80% or greater as measured by the area under the receiver operating characteristic (ROC) curve. But the gravity of the consequences of error in cancer diagnosis means that these performance levels, which are fairly strong in many applications, are quite inadequate in our present context of breast cancer differentiation. The rationale underlying AdaBoost continues to be relevant in this setting, however, and the combined classifier inherits the virtues of both of our moderately strong classifiers. AdaBoost and the variations have proved to be efficacious in a diverse range of applications (Schapire and Freund 2012). As described next, we use a variant of the original procedure to combine the two feature sets arising out of human and computer analyses using classifiers with real-valued outcomes.

In essence, AdaBoost minimizes an exponential loss criterion on a training sample using a forward, stagewise additive model to combine multiple constituent classifiers. The procedure begins with a *tabula rasa* in which the elements of the training sample are weighted equally. In the first round of training, the first classifier in the sequence is trained on the uniformly weighted sample, and at the end of the round, the sample is reweighted with greater weight placed on those cases that the

classifier found difficult to label correctly. Training now proceeds in a sequence of rounds, each constituent classifier trained sequentially on a reweighted sample with greater weight placed on the cases that the previous classifier in the sequence found difficult to classify accurately. At the end of each round of training, the sample error (or, more precisely, risk) of the classifier being trained on the (reweighted) training sample is recorded before the sample is reweighted yet again and passed on to the next classifier in the sequence. The final boosted classifier forms its prediction as a convex combination of the predictions of each constituent classifier, with the more accurate classifiers (with respect to training sample error) given more weight. Next we specialize the procedure to our setting.

The representation of a lesion in feature space and its (biopsy-certified) label form a pair $(x, y)$, where $x$ represents a feature vector and $y$ is the associated label. In our current context of features extracted by two readers, we may represent the feature vector in the form $x = (F^{(1)}, F^{(2)})$ where $F^{(1)} = F_1^{(1)}, \ldots, F_{N_1}^{(1)}$ represents the collection of (nominal-valued) BI-RADS$_{US}$ features and $F^{(2)} = F_1^{(2)}, \ldots, F_{N_2}^{(2)}$ represents the collection of (continuum-valued) computer-generated features. The associated label $y$ takes one of two numeric values, 1 representing malignant (M) and 0 representing benign (B). In this setting a real-valued classifier is a function $f(x)$ that maps each vector $x$ in feature space into a real value that nominally represents the classifier's estimate of the *a posteriori* probability of malignancy conditioned on the observed feature vector $x$. The goal in classifier design is to minimize, in some suitable sense, the error $|f(x) - y|$.

Let $f_1(F^{(1)})$ and $f_2(F^{(2)})$ represent the outputs of the naïve Bayes and logistic regression classifiers operating on their respective feature sets. Although, nominally, each of these classifiers operates on the entire feature space $x = (F^{(1)}, F^{(2)})$, in practice they have been selected to match the characteristics of the two rather different types of features that have been generated. Accordingly, the naïve Bayes classification $f_1(x) = f_1(F^{(1)})$ depends only on the nominal BI-RADS$_{US}$ features $F^{(1)} = F_1^{(1)}, \ldots, F_{N_1}^{(1)}$, whereas, in a similar fashion, the logistic regression classification $f_2(x) = f_2(F^{(2)})$ depends only on the computer-generated features $F^{(2)} = F_1^{(2)}, \ldots, F_{N_2}^{(2)}$. Thus, each of these classifiers operates in a natural lower-dimensional subspace of the entire feature space. We may view these procedures in a formal sense as using domain knowledge to reduce both the effective dimensionality of the feature space and the effective complexity of the resulting classifier $f(\cdot)$ obtained by boosting from $f_1(\cdot)$ and $f_2(\cdot)$. This, in effect, accomplishes a "practitioner's complexity regularization" by using domain knowledge to mitigate both Bellman's curse of dimensionality and the danger of over fitting.

The boosted classifier $f(\cdot)$ is obtained from the constituent classifiers $f_1(\cdot)$ and $f_2(\cdot)$ by training on a random sample of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ obtained (presumptively) by independent sampling from an underlying probability distribution governing the class-conditional distribution of features and the *a priori* probabilities of the two classes in the population at large. We train first on the naïve Bayes classifier $f_1(\cdot)$ and then on the logistic regression classifier $f_2(\cdot)$. The order may of course be reversed without any essential change in the algorithm.

We begin in round 1 by placing equal weight on each element of the data sample: $w_{1,j} = 1/n$ for $1 \le j \le n$. The classifier $f_1$ is fitted to the training data using weights $w_{1,j}$, where fitting is with respect to the squared error. The optimum classifier $f_1$ is selected (in our case from the class of naïve Bayes classifiers), and its weighted (minimum) squared error computed:

$$err_1 = \sum_{j=1}^{n} w_{1,j} \left| f_1(x_j) - y_j \right|^2 \tag{7}$$

The weight (or importance) $c_1$ of classifier $f_1$ is now deduced from the squared error *via* the logarithm of the odds ratio,

$$c_1 = \frac{1}{2} \ln \left( \frac{1 - err_1}{err_1} \right) \tag{8}$$

Round 2 (and, in our case, the final round of training) begins with a reweighting of the data sample. For $1 \le j \le n$, the data point $(x_j, y_j)$ is given weight

$$w_{2,j} = \frac{w_{1,j} e^{c_1 \left| f_1(x_j) - y_j \right|^2}}{Z_1} \tag{9}$$

where $Z_1$ is a normalization factor chosen so that the sum of the weights is 1. The second classifier $f_2$ is now fitted to the training data using weights $w_{2,j}$, fitting again with respect to squared error. The squared error $err_2$ and the weight (importance) $c_2$ of the best logistic regression classifier $f_2$ are now computed with analogous formulas:

$$err_2 = \sum_{j=1}^{n} w_{2,j} \left| f_2(x_j) - y_j \right|^2 \text{ and } c_2 = \frac{1}{2} \log \left( \frac{1 - err_2}{err_2} \right) \tag{10}$$

This concludes the training stage. It should be noted that without loss of generality, we may assume $err_1$ and $err_2$ are both $<1/2$ (if necessary, by interchanging the roles of 0 and 1), so that the classifier weights $c_1$ and $c_2$ are both positive.

The AdaBoost classifier $f$ is formed as a convex combination of the two constituent classifiers $f_1$ and $f_2$ weighted in accordance with their relative importance $c_1$ and $c_2$, respectively. Thus, we set

$$f(x) = \frac{c_1}{c_1+c_2} f_1(x) + \frac{c_2}{c_1+c_2} f_2(x) \qquad (11)$$

Although in our study we used two classifiers, the above process can be expanded to include an arbitrary number of constituent classifiers, reweighting the training cases for each constituent classifier based on the weights and performance of the previous classifier.

*Selective pruning*

In the standard mode of operation, once the AdaBoost classifier $f$ is determined, a hard-limited classification of a given lesion with feature vector $x$ is made by selecting a threshold $t$ of operation and mapping $x$ to 1 (malignant) or 0 (benign) in accordance with whether $f(x) > t$ or $f(x) \leq t$, respectively. By varying $t$ we obtain the ROC curve characteristic of the classifier.

If the AdaBoost classifier output $f(x)$ represents an estimate of the *a posteriori* probability of malignancy for a given a vector of features $x$, then values near 0 or 1 represent high confidence that the cases are benign or malignant, respectively, whereas values near ½ represent ambiguous cases that need to be pruned for further evaluation by other imaging methods. Pruning of the low-confidence cases was performed by specifying an *ambiguity interval* $(t_l, t_u)$ with a lower threshold $t_l$ and an upper threshold $t_u$. Lesions for which the feature vector $x$ satisfies $t_l < f(x) < t_u$ were pruned and set aside for further testing by other imaging methods. On the other hand, lesions for which the feature vector $x$ was outside the ambiguity interval were recorded as 1 or 0 by the pruned classifier depending on whether $f(x) \geq t_u$ or $f(x) \leq t_l$, respectively. The cases that are pruned determine the *drop rate* of the pruned classifier. With the ambiguity region centered on 1/2 (corresponding to the maximum uncertainty region of the classifier), drop rate was determined for different ambiguity intervals. As cases that are pruned will require additional testing, the drop rate stands in the role of a quantifiable surrogate for a cost; the performance of the classifier on the disambiguated cases that are retained provides a second quantifiable attribute. The pruned classifier hence operates effectively in a 2-D cost–performance rubric and allows users to achieve desired levels of performance or drop rates by varying the location and size and of the ambiguity interval.

*Consensus classifiers.* The idea that there are a number of cases that are persistently difficult for both classifiers to label accurately and that these hence continue to be problematic for the AdaBoost combined classifier suggests that *consensus* may provide a useful, low-complexity method for combining classifiers. A threshold, $t_{th}$, for the unweighted probability estimates of the constituent classifiers is chosen to call lesions as malignant or benign. The cases are retained if both classifiers agree on the classification and are pruned otherwise. The operational threshold $t_{th}$ is varied over the entire range of probabilities, and at each threshold, the drop rate is determined. An ROC performance curve is obtained by varying the thresholds of operation with the caveat that there is in general a varying drop rate for each operational point.

## RESULTS

*Diagnostic performance of visual and computer features*

In previous studies we had compared the diagnostic performance of two observers using visual features described by the BI-RADS lexicon (Bouzghar et al. 2014). These studies provide the baseline for the present study, and the results of the analysis of the two observers (O1 and O2), along with the results of computer analysis, are illustrated in Figure 3.

Both visual and computer features were high performers, with areas under the ROC curve ($A_z$) ranging between 0.866 and 0.924 (Table 1, rows A and B). The areas under the ROC curve, $A_z$, for the two observers differed markedly: 0.924 for observer 1 versus 0.866 for observer 2, with observer 2 underperforming observer 1 in all parts of the curve (Fig. 3). The difference of 0.058 in $A_z$ between the two observers was significant ($p = 0.006$) as measured by the method described earlier (DeLong et al. 1988).

The ROC curve for computer features was between the performance of the two observers; $A_z$ for computer features was lower than that of observer 1, 0.887 versus

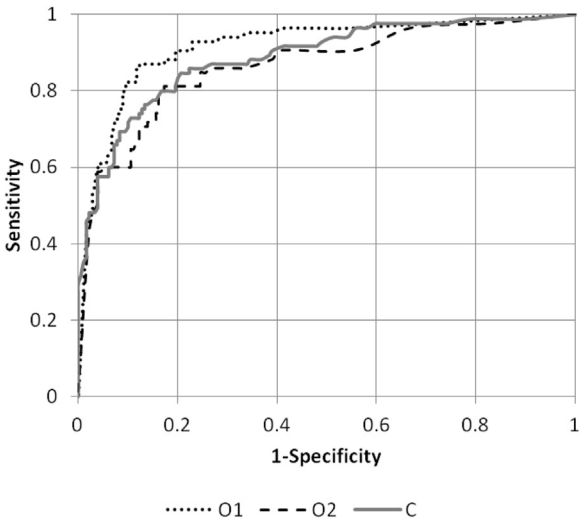


Fig. 3. Comparison of diagnostic performance of human observers 1 (O1) and 2 (O2) with computer based image analysis (C) in the differentiation of breast masses.

Table 1. Area under receiver operating characteristic curves ($A_z$) for visual and computer features using different learning strategies*

| | | $A_z$ | | |
|---|---|---|---|---|
| | | Observer 1 | Computer | Observer 2 |
| A | Visual features, Naïve Bayes | 0.924 ± 0.021 | | 0.866 ± 0.027 |
| B | Computer features, logistic regression | | 0.887 ± 0.025 | |
| | | Observer 1 + computer | | Observer 2 + computer |
| C | AdaBoost, visual features (A) → computer features (B) | 0.937 ± 0.018 | | 0.906 ± 0.023 |
| D | AdaBoost, computer features (B) → visual features (A) | 0.936 ± 0.019 | | 0.905 ± 0.023 |

* The arrows represent the direction of boosting.

0.924, but higher than that of observer 2, 0.887 versus 0.866. The difference in $A_z$ between observers and the computer analysis was not significant: $p = 0.334$ for observer 1 versus computer, and $p = 0.279$ for observer 2 versus computer.

*Enhancing diagnostic performance*

Combining visual and computer features with AdaBoost increased $A_z$ for both observers: 0.924 to 0.937 for observer 1 and 0.866 to 0.905 for observer 2 (Table 1, rows C and D). The difference in performance between the observers was reduced to 0.031 after boosting, compared with the difference of 0.058 observed without boosting. The improvement in performance after boosting was significant ($p = 0.016$).

As expected from theoretical considerations, the order of selection of individual classifiers in the boosting procedure did not significantly affect outcomes. The $A_z$ values obtained by taking visual features first, followed by computer-generated features, was comparable to those obtained by taking computer-generated features first, followed by visual features: 0.937 and 0.936 for observer 1 and 0.906 and 0.905 for observer 2 (Table 1, rows C and D).

Figures 4 and 5 provide a key validation of the AdaBoost principle. In Figure 4 we plot the estimate of the *a posteriori* probability $P(M \mid F)$ engendered by observer 1 versus that for observer 2; these probability estimates were obtained by using the naïve Bayes procedure on the visual features generated by the two observers for each case. In Figure 5 we plot the revised estimates of the two *a posteriori* probabilities $P(M \mid F)$ obtained by AdaBoost by adaptively boosting visual features for the two observers with computer-generated features for each image. Without boosting, there is a marked difference between the probability estimates of the two observers. The *dotted line* in the figure represents the linear ($y = mx$) least-squares fit of the data with $R^2$ of 0.44. The concordance correlation coefficient ($\rho_c$), estimating the degree to which pairs of observations fall on the 45° line through the origin, was 0.80. After AdaBoosting the probability estimates of the two observers were uniformly distributed and became better correlated with an $R^2$ of 0.64 and a $\rho_c$ of 0.93 (Fig. 5). The difference between the unboosted and boosted groups for both measures, $R^2$ ($p < 0.006$) and $\rho_c$ ($p < 0.0001$), was significant. These figures illustrate a key feature: As predicted by theory, *adaptive boosting results in a greater consensus with a concomitant reduction in variability across observers*.

*Pruning by margin classifier*

Figure 6 illustrates the effect of using a banded threshold, where the cases within the pruning threshold band are dropped and the diagnostic decision is postponed for additional testing because of the low diagnostic performance within the band. The ROC curves for three drop rates ranging from 0% to 40% reveal a uniform improvement in diagnostic performance with increase in drop rate. Indeed, as the drop rate increases, it is evident that the curves converge toward unit sensitivity and specificity.

In Figure 7 we plotted the diagnostic performance for each observer for the AdaBoost classifier as a function of the pruning rate. As is evident, the area under the ROC

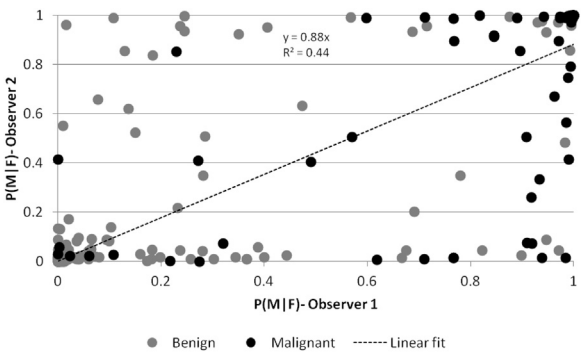


Fig. 4. Case-by-case distribution of probability of malignancy $P(M \mid F)$ estimates obtained from the raw visual features generated by the two observers.

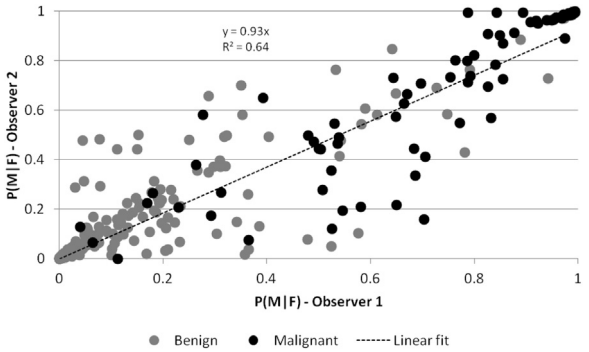


Fig. 5. Case-by-case distribution of probability of malignancy $P(M \mid F)$ estimates of the two observers after combination of visual and computer-generated features using AdaBoost.

curve increases monotonically with drop rate for both observers (though performance on the feature set engendered by observer 1 dominates at all drop rates). The increase, however, is non-linear, with rapid improvement initially moderating to a more gradual improvement at drop rates above 20%. Eventually, the curves plateau with minimal benefits from further increases in drop rates above 50%. A 20% drop rate provides a reasonable compromise between drop rate and performance improvement; at this drop rate, the area under the ROC curve increases from 0.937 ± 0.018 to 0.974 ± 0.012 for observer 1 and from 0.906 ± 0.023 to 0.952 ± 0.017 for observer 2. *In both cases we see a rather dramatic improvement in performance at a moderate cost in terms of pruned cases requiring further evaluation.*

The change in specificity at a fixed sensitivity for different drop fractions is illustrated in Figure 8. This graph indicates that a user can choose an operation point to achieve different sensitivities and specificities. For example, at a drop rate of 20% (*vertical arrow*), a specificity of 0.95 can be achieved at a sensitivity of 0.90. For the same drop rate, the specificity drops to 0.88 and 0.48 at sensitivities of 0.95 and 1.0, respectively.

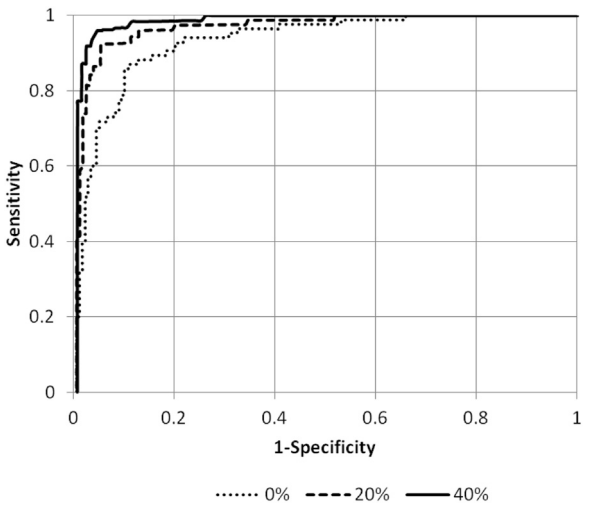


Fig. 6. Receiver operating curves at different drop rates of 0%, 20 % and 40%.

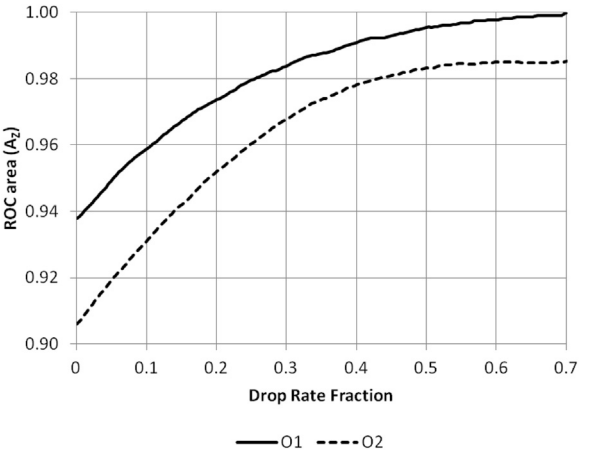


Fig. 7. Effect of drop rate on the diagnostic performance based on the AdaBoost *a posteriori* probability estimates. Diagnostic performance was measured as the area under the ROC curve. O1 and O2 represent data for observers 1 and 2, respectively. ROC = receiver operating characteristic.

*Pruning by consensus*

Figure 9 illustrates the results of diagnostic performance when consensus between visual and computer analyses was used for differentiating malignant and benign masses with cases on which there was no consensus being pruned. For both observers the use of computer analysis increased the area under the ROC curves generated by varying the threshold for classification of malignant cases: For observer 1, $A_z$ increased to 0.973 ± 0.012, whereas for observer 2, $A_z$ increased to 0.955 ± 0.016.

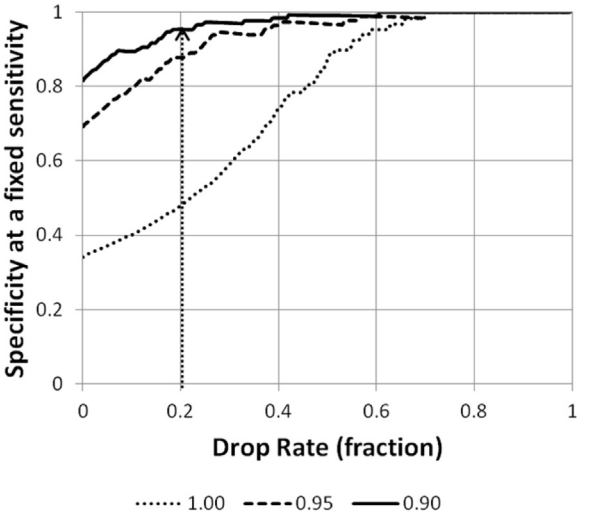


Fig. 8. Change in specificity with drop rate at different fixed specificity values. *Dotted*, *dashed* and *solid* curves correspond to fixed sensitivities of 1.0, 0.95 and 0.90, respectively.

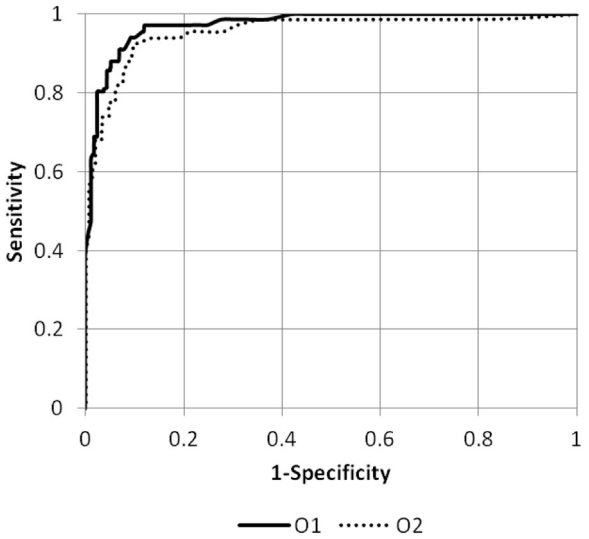


Fig. 9. Receiver operating characteristic curves of the consensus probability estimates of observers 1 and 2.

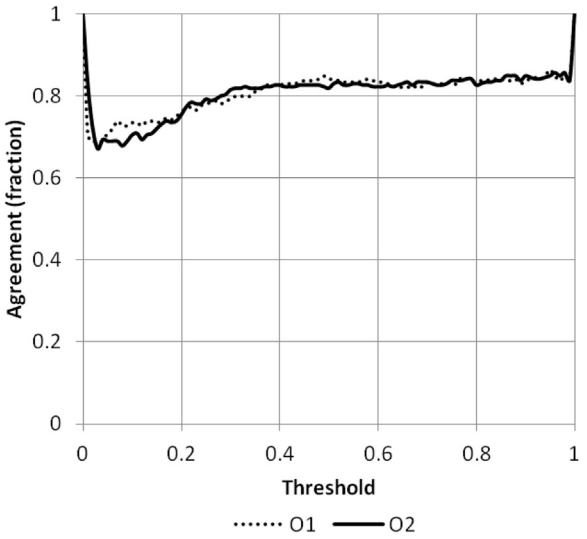


Fig. 10. Percentage agreement between visual and computer analyses at different probability thresholds for differentiating malignant and benign masses. O1 and O2 represent data for observers 1 and 2, respectively.

The difference of 0.028 in $A_z$ between the two observers was not significant ($p = 0.0873$). Figure 10 illustrates the results of agreement between the visual and computer analyses as a function of the selected threshold for malignancy. The average agreement between visual and computer analyses was 80.9 ± 5.2% and 80.7 ± 5.7% for observers 1 and 2. The difference was not significant ($p = 0.77$).

Although the drop rate for a consensus-based procedure is not directly controllable, Figure 10 illustrates that we are operating at about a 20% drop rate over a wide range of thresholds. This suggests that consensus is, in effect, providing a heuristic approximation to a boosting procedure coupled with a 20% drop rate. This interpretation supports the results in Table 2, which indicates that the performance of the consensus-based procedure with a drop rate of approximately 20% (inherited from cases on which there is no consensus) is essentially the same as that of the AdaBoost procedure coupled with selective pruning of 20% of the cases (rows A and B in Table 2 correspond to a drop rate of 20% in Fig. 7; rows C and D in Table 1 correspond to a drop rate of 0% in Fig. 3).

## DISCUSSION

As described earlier, the machine learning algorithm uses two feature sets generated from a library of biopsy-classified ultrasound images. The extraction of the feature sets from each ultrasound breast image in the database follows two distinct pathways: The visual classification of features by a radiologist follows the standard BI-RADS$_{US}$ paradigm, whereas the generation of the computer-based feature set is automated and focuses on various geometric and gray-scale attributes of the lesion. These feature extraction processes, although critical, are by now well understood and have been described before (Bouzghar et al. 2014; Sehgal et al. 2004). The results of this study indicate that both BI-RADS$_{US}$ features, derived from a human observer by visual inspection of the ultrasound images, and computer-generated features can be used individually to predict malignancy of breast masses with comparable accuracy. The area under the ROC curve ranged from 0.866 to 0.924 for human observers versus 0.887 for computer analysis. The difference in performance between the human observers was observed to be larger than the difference observed between the computer and individual observers. The close similarity in the performance of computer-based analysis and human experts yields a strong rationale for using a computer-based system as a second reader to improve the accuracy of either human observer individually. This premise was tested by combining computer-generated and BI-RADS$_{US}$ features as discussed earlier. Next we discuss the theoretical and experiential rationales underlying this choice.

### *Combining classifiers*

Prior experience with classification of solid breast ultrasound lesions indicates that essentially any commonly used generic learning algorithm can achieve solid levels of performance, but it is quite hard to improve outcomes significantly beyond a certain point. For instance, if we adopt the area under the ROC curve as a measure of performance, then performance levels of 85% or so can be attained relatively easily, but improvements beyond this are hard to come by even with

Table 2. Areas under receiver operating characteristic curves ($A_z$) for visual and computer features*

| | Features/classifier | $A_z$ | |
|---|---|---|---|
| | | Observer 1 + computer | Observer 2 + computer |
| A | AdaBoost, visual features (a) → computer features (b), drop rate 20% | 0.975 ± 0.018 | 0.956 ± 0.023 |
| B | Consensus, visual features ↔ computer features, drop rate ~ 20% | 0.973 ± 0.012 | 0.955 ± 0.016 |

* In row A are the $A_z$ values for the two observers for AdaBoost combined with selective pruning at a drop rate of 20%. In row B are the $A_z$ values for the two observers using consensus; the drop rate that is inherited by the procedure is approximately 20%. The arrows represent the direction of boosting.

determined algorithmic tweaking. This observation is quite robust with respect to choice of algorithmic vehicle: neural nets, logistic regression, support vector machines and Bayes' methods all yield performances in this ball park (Bouzghar et al. 2014; Cary et al. 2012; Song et al. 2005). This suggests, on the one hand, that the characteristics of malignant and benign masses are, generally speaking, well separated in feature space, thus permitting any reasonably complete classification mechanism to quickly infer the essential discriminating features. On the other hand, it also means that associated with each classifier is a non-trivial set of obdurate, ambiguous images constituting about 10% to 20% of the cases that the classifier cannot easily resolve. This leads to the option of combining two (or more) classification methodologies.

There are potentially three possible ways of combining data: (i) by training classifiers on both feature sets together, (ii) by using boosting algorithms to enhance the outputs of the individual models and (iii) by using a consensus or voting between models or observers. The first option has two drawbacks. First, the observer-generated visual features are discrete in nature, whereas the computer-generated features are continuous, meaning that a classifier best suited for one will not necessarily be well suited to the other. Second, some of the features in the two data sets are very similar; for example, the "depth-to-width ratio," a computer-generated feature, measures something similar to parallel versus non-parallel orientation in the BI-RADS$_{US}$ lexicon. Putting both of these in the same feature set will bias results by double counting the same underlying property of the mass.

In this study the second option was attempted using AdaBoost (Freund and Schapire 1997). AdaBoost has been found to enhance the performance of an ensemble of a large number of weakly performing individual features for breast masses (Takemura et al. 2010). Our use of AdaBoost differed from its conventional use in several ways. Although AdaBoost is typically used to enhance collective performance of an ensemble of "weak classifiers" whose individual performance is little better than random guessing, nothing in principle prevents its use when the individual classifiers are relatively strong. In fact, the better the individual classifiers, the lower the error rate, all other things being equal. This is precisely the condition we encounter in this study in which two base classifiers individually have strong accuracies in absolute terms (in excess of 85% area under the ROC curve). Although these levels of performance are strong, they are not sufficient for breast diagnosis because of the gravity of the consequences resulting from an incorrect diagnosis. This study indicates that the performance is enhanced by using just two high-accuracy classifiers represented by human and machine observers, rather than a large number of lower-accuracy classifiers.

The function of AdaBoost is to examine the independent diagnosis of the individual observer and that of the computer algorithm. After this examination, it assesses the definitiveness of diagnosis by measuring the distance from midprobability of 0.5. Cases that have a less definitive diagnosis by one approach (say, human observer) receive greater weight in the final diagnosis from the second approach (computer algorithm) to boost performance. In essence AdaBoost is a "weighting machine" that boosts performance by selectively emphasizing the diagnosis from one or the other of two independent sources: human observer and computer algorithm. Because AdaBoost is an "arbitrator" that treats the two sources of information as independent, it does not involve joint training for each individual with the computer algorithm.

The deployment of AdaBoost in this study differs from the normal use of the procedure in two additional ways. The error function was modified to account for the continuous nature of the classifiers used and to limit overweighting cases with a large difference between the predicted and actual outcomes. Finally, the individual classifiers each exploit expert domain knowledge and operate on a portion of the entire feature space, thus partly obviating Bellman's "curse of dimensionality," as well as reducing the accumulation of insidious bias from correlated features in different sets.

The results indicate that the use of AdaBoost had a modest improvement in the ROC area for observer 1 with the stronger initial performance ($A_z = 0.924$), with a more marked improvement for observer 2 who had weaker initial performance ($A_z = 0.866$). In other words, as predicted by theory, the improvements conferred by AdaBoost were stronger if the initial accuracy was lower.

In particular, this implies that the proposed technique could have an even more pronounced impact in an environment where the observers are less expert.

The use of AdaBoost also had another interesting effect on the performance of individual observers. Although the diagnostic performance of two observers on aggregate was high, 0.866 versus 0.924, there were significant differences in the probability estimates on a case-by-case basis, as illustrated by the scatter in Figure 4. These differences are to be expected given the significant biological variability in the characteristics of breast lesions, the limitations of imaging systems in depicting these lesions accurately and differences in observer expertise. However, after AdaBoost was used, the probability estimates of the two observers became more uniformly distributed and correlated, as illustrated in Figure 5. Although further investigations are needed, these results suggest that AdaBoost could be a useful means of improving consistency in the diagnosis between different observers.

Although the robustness of classification performance with respect to choice of algorithm may suggest that the particular selection of classifiers (naïve Bayes, logistic regression, neural net) is not critical (Cary et al. 2012; Song et al. 2005), the distinctive natures of the two feature sets that we use suggests that we may as well get some small benefit by using domain knowledge to select our initial classifiers to match the characteristics of the feature set. The first reader feature set comprising the BI-RADS$_{US}$-derived features is intrinsically nominal in nature and hence well-suited to a classification methodology operating on nominal data. One of the simplest, effective algorithms in this class is the naïve Bayes procedure, and we adopted it as a first-stage classifier. The computer-generated feature set (second reader), on the other hand, is intrinsically continuous in nature and is suited to a panoply of algorithms operating on real-valued data. One of the simplest algorithms in this class, and the one we adopted in the work described here, is logistic regression. Both these procedures have the added virtue of producing results that are readily interpretable as providing approximations to *a posteriori* probability estimates of malignancy for given features. It is important to emphasize that either or both of these algorithms may be replaced by other algorithms in these classes without essential change in the nature of the results; the impact resides not in the specific algorithmic choices, but in their principled combination accompanied by pruning. In this study, the data available to us were purely BI-RADS$_{US}$ features. In principle, one can imagine a nuance of data interpretation where radiologists also provide confidence in the assessment in addition to BI-RADS categorization. Because giving confidence in the interpretation of the images is not a part of routine clinical practice, there is currently no standardized way to obtain this information. However, we envision that if such was available in a standardized form, it could also be folded into the algorithm to further enhance the diagnostic performance.

Our results indicate that pruning increased the accuracy on cases for which a prediction is made, at the quantifiable cost of making no prediction on a fraction of cases, the drop rate. The extent of improvement increased with the increase in the drop rate fraction. A key discovery was that near-perfect classification performance is achievable in a majority of cases at the cost of a modest drop rate of ambiguous cases pruned in a computationally effective manner for additional testing. For example, at a drop rate of 20%, an area under the ROC curve of $A_z$ = 0.975 can be achieved for observer 1 (Figs. 7 and 8). A more detailed examination (Figs. 7 and 8) reveals that for sensitivities between 0.90 and 0.95, one can achieve high specificities between 0.975 and 0.88 for the visual feature set generated by observer 1 coupled with computer-generated features *via* Adaptive Boosting and selective pruning at a 20% drop rate.

The area under the ROC curve, $A_z$, has the virtue of providing a single, easily understandable metric for purposes of comparison. But to achieve the improvement we pay a price in a modest pruning rate. As we may anticipate that the cases pruned will be sent for biopsy, a natural question that arises is how the number of unnecessary biopsies has been affected with and without pruning. It is essential, however, that we make our comparisons at the same false negative rate: These cases connote mortality and constitute the single biggest cost in the procedure.

With this as background, for purposes of illustration consider a 20% drop rate (the middle curve in Fig. 6). For illustrative purposes, suppose that there are 200 cases with 100 benign and 100 malignant and that the distribution of malignant and benign cases is the same in the pruned subsample. Thus, at a 20% drop rate there will be 80 malignant and 80 benign cases in the retained subpopulation of 160 high-confidence cases and 20 malignant and 20 benign cases in the low-confidence subpopulation of 40 pruned cases (all presumed to be sent for a biopsy). In the high-confidence group, if we operate at 80% specificity, we obtain a sensitivity (true positive fraction) of 98% from the middle curve of Figure 6, leading to two missed malignant cases (rounded up from 1.6) out of the total of 80 in this group. This is the regrettable false-negative rate at this level of specificity. At an operational point of 80% specificity, 16 of the 80 benign cases in the high-confidence group are misdiagnosed (false-positive results), and together with the 20 pruned benign cases, this leads to 36 unnecessary biopsies out of a total of 134 biopsies performed with 2 missed malignant cases (Table 3).

Table 3. Calculations illustrating the impact of pruning on biopsy yield

| | | | | | | |
|---|---|---|---|---|---|---|
| **20% Pruning rate** | | | | | | |
| | High confidence group | | | | Low confidence group | |
| | Malignant cases 80 | | Benign cases 80 | | Malignant cases 20 | Benign cases 20 |
| | Sensitivity = 0.98 | | Specificity = 0.8 | | | |
| | False positive (2%) | True positive (98%) | False positive (20%) | True positive (80%) | No diagnosis | No diagnosis |
| Number of cases | 2 | 78 | 16 | 64 | 20 | 20 |
| Action | No biopsy | Biopsy | Biopsy | No biopsy | Biopsy | Biopsy |
| Result | Missed malignancy | Necessary biopsies | Unnecessary biopsies | Biopsies saved | Necessary biopsies | Unnecessary biopsies |
| **No pruning** | | | | | | |
| | Malignant cases 100 | | Benign cases 100 | | Malignant cases 0 | Benign cases 0 |
| | Sensitivity = 0.98 | | Specificity = 0.45 | | | |
| | False positive (2%) | True positive (98%) | False positive (55%) | True positive (45%) | 0 | 0 |
| Number of cases | 2 | 98 | 55 | 45 | | |
| Action | No biopsy | Biopsy | Biopsy | No biopsy | | |
| Result | Missed malignancy | Necessary biopsies | Unnecessary biopsies | Biopsies saved | | |

To compare this procedure with operation without pruning, we must function at the same true positive rate (sensitivity 98%) leading to 2 missed malignant cases. As Figure 6 illustrates, the corresponding specificity for this operating point for the lowest curve (representing no pruning) is somewhere between 0.4 and 0.50 because the curve is very flat at high sensitivities. At a sensitivity of 0.98 and specificity of 0.45, for definiteness, there will be 55 unnecessary biopsies out of the 153 biopsies performed again with 2 missed malignant cases (Table 3). Thus, by keeping the false-negative result rate low at 2%, pruning has reduced the number of unnecessary biopsies from 55 to 36 (a reduction of 35%).

The purpose of our analysis in the above example was to illustrate purposes the scope for reduction in the number of unnecessary biopsies at a very low false-negative result rate. Many factors affect the actual gains in practice. In our study there were two benign masses for every malignant case. If this *a priori* information is taken into consideration, the benefit is even greater than in our numerical example: For 2 missed malignancies, there are 34 unnecessary biopsies out of 99 biopsies performed for the pruned case, compared with 73 unnecessary biopsies out of 139 biopsies performed when there was no pruning. Similarly, the benefits will be even greater than in the preceding example if malignant lesions, which are often more difficult to characterize, are present in larger numbers in the ambiguous low group that is pruned.

*AdaBoost and consensus models*

The option to treat the two classifiers as independent observers, and make predictions only on cases where the two agree is also studied. The cases with no agreement are considered to represent low-confidence cases that needed more evidence for decision and were pruned from the set of cases. Similar to Adaptive Boosting with selective pruning, this has the advantage of increasing accuracy on those cases for which a prediction is made. The results indicate that the agreement fraction is fairly constant over a broad range of threshold values used to differentiate malignant and benign masses (Fig. 10). The average percentage agreement over all thresholds was 81% (drop rates 19%) between the computer analysis and observers 1 and 2, respectively. Requiring consensus between computer and visual analyses markedly improved the diagnostic performance of each observer: For observer 1, $A_z$ increased from 0.924 $\pm$ 0.021 to 0.973 $\pm$ 0.012, whereas for observer 2, it increased from 0.866 $\pm$ 0.027 to 0.955 $\pm$ 0.016 (Table 1). Comparison of Figures 3 and 9 reveals that the diagnostic performances of the two observers, which were noticeably different initially (Fig. 3), became comparable to one another after cases of disagreement were pruned by the consensus seeking procedure (Fig. 9). This is consistent with the smoothing of scatter seen in Figure 5 when AdaBoost is used to couple the visual features generated by the two observers with computer-generated features.

Although, the consensus model is simple to use and provides insights into the rationale for leaving some cases out, it has the disadvantage that the number of cases left out is not controlled *a priori*, but is inherited from the agreement or disagreement of the classifiers on the sample, though Figure 10 illustrates that the inherited drop rate is approximately 20% over a wide range of operational thresholds. In adaptive boosting coupled with selective pruning, on the other hand, any desirable feasible operating point in a 2-D performance versus drop rate rubric may be chosen depending on the permissible drop rate and the desired performance. As a practical matter, a consensus classifier may be viewed as approximately implementing one operational point in the 2-D feasible region of operating points of an adaptively boosted classifier incorporating selective pruning. For the present study, a consensus-based procedure implements a computationally simple *ad hoc* boosting method at a drop rate of approximately 20%.

## CONCLUSIONS

Despite considerable efforts, the goal of reducing unneeded biopsies continues to be an important objective of breast cancer imaging. This study describes a machine learning methodology involving adaptive boosting and selective pruning for optimizing cost and performance for differentiating solid breast masses with ultrasound imaging. These techniques enable a computer-based analysis to complement and enhance the performance of human observers; the cases with low diagnostic confidence are pruned, and accurate diagnoses are made on the remaining cases. In short, the machine learning methodology described in this study goes beyond the first reader by providing a second opinion that improves diagnosis significantly without the added cost of a human observer.

*Acknowledgments*—We thank Theodore W. Cary, Susan M. Schultz and Karen Apadula for help with image acquisition and data analysis. The images used in this study were also used to evaluate the baseline performance of individual observers in our previous publication (Bouzghar et al. 2014). This work was supported in part by National Institutes of Health Grant CA130946.

## REFERENCES

American College of Radiology (ACR). BI-RADS: Ultrasound. In: Breast imaging reporting and data system: BI-RADS atlas. 4th ed. Reston, VA: Author; 2003.

Bouzghar G, Levenback BJ, Sultan LR, Venkatesh SS, Cwanger A, Conant EF, Sehgal CM. Bayesian probability of malignancy with breast ultrasound BI-RADS features. J Ultrasound Med 2014;33:641–648.

Candelaria RP, Hwang L, Bouchard RR, Whitman GJ. Breast ultrasound: current concepts. Semin Ultrasound CT MR 2013;34:213–225.

Cary TW, Cwanger A, Venkatesh SS, Conant EF, Sehgal CM. Comparison of naïve Bayes and logistic regression for computer-aided diagnosis of breast masses using ultrasound imaging. In: Bosch JG, Doyley MM, (eds). Medical Imaging 2012: Ultrasonic Imaging, Tomography, and Therapy, 8320. Proc SPIE; 2012:83200(M1–M7).

Chang JM, Won JK, Lee KB, Park IA, Yi A, Moon WK. Comparison of shear-wave and strain ultrasound elastography in the differentiation of benign and malignant breast lesions. AJR Am J Roentgenol 2013;201:W347–W356.

Chen DR, Chang RF, Huang YL. Computer-aided diagnosis applied to US of solid breast nodules by using neural networks. Radiology 1999;213:407–412.

Cheng HD, Shan J, Ju W, Guo Y, Zhang L. Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recog 2010;43:299–317.

Cho KR, Seo BK, Lee JY, Pisano ED, Je BK, Lee JY, Choi EJ, Chung KB, Whan Oh Y. A comparative study of 2-D and 3-D ultrasonography for evaluation of solid breast masses. Eur J Radiol 2005;54:365–370.

Cho N, Moon WK, Cha JH, Kim SM, Han BK, Kim EK, Kim MH, Chung SY, Choi HY, Im JG. Differentiating benign from malignant solid breast masses: Comparison of two-dimensional and three-dimensional US. Radiology 2006;240:26–32.

DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988;44:837–845.

Duric N, Littrup P, Poulo L, Babkin A, Pevzner R, Holsapple E, Rama O, Glide C. Detection of breast cancer with ultrasound tomography: First results with the computerized ultrasound risk evaluation (CURE) prototype. Med Phys 2007;34:773–785.

Freund Y, Schapire RE. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. J Comput System Sci 1997;55:119–139.

Georgian-Smith D, Moore RH, Halpern E, Yeh ED, Rafferty EA, D'Alessandro HA, Staffa M, Hall DA, McCarthy, Kopans DB. Blinded comparison of computer-aided detection with human second reading in screening mammography. AJR Am J Roentgenol 2007;189:1135–1141.

Golatta M, Schweitzer-Martin M, Harcos A, Schott S, Junkermann H, Rauch G, Sohn C, Heil J. Normal breast tissue stiffness measured by a new ultrasound technique: Virtual Touch tissue imaging quantification (VTIQ). Eur J Radiol 2013;82:e676–e679.

Gromet M. Comparison of computer-aided detection to double reading of screening mammograms: Review of 231,221 mammograms. AJR Am J Roentgenol 2008;190:854–859.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Berlin/New York: Springer; 2009.

Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. Acad Radiol 2004;11:272–280.

Huang YL. Computer-aided diagnosis using neural networks and support vector machines for breast ultrasonography. J Med Ultrasound 2009;17:17–24.

Kopans DB. The positive predictive value of mammography. AJR Am J Roentgenol 1992;158:521–526.

Liu B, Cheng HD, Huang J, Tian J, Tang X, Liu J. Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images. Pattern Recog 2010;43:280–298.

McDonald DN. 3-Dimensional breast ultrasonography: What have we been missing? Ultrasound Clin 2011;6:381–406.

Padilla F, Roubidoux MA, Paramagul C, Sinha SP, Goodsitt MM, Le Carpentier GL, Chan HP, Hadjiiski LM, Fowlkes JB, Joe AD, Klein KA, Nees AV, Noroozian M, Patterson SK, Pinsky RW, Hooi FM, Carson PL. Breast mass characterization using 3-dimensional automated ultrasound as an adjunct to digital breast tomosynthesis: A pilot study. J Ultrasound Med 2013;32:93–104.

Ruiter NV, Zapf M, Hopp T, Dapp R, Kretzek E, Birk M, Kohout B, Gemmeke H. 3-D ultrasound computer tomography of the breast: A new era? Eur J Radiol 2012;81(Suppl 1):S133–S134.

Sahiner B, Chan HP, Roubidoux MA, Helvie MA, Hadjiiski LM, Ramachandran A, Paramaqul C, LeCarpentier GL, Nees A, Blane C. Computerized characterization of breast masses on three-dimensional ultrasound volumes. Med Phys 2004;31:744–754.

Schapire R, Freund Y. Boosting: Foundations and algorithms. Cambridge: MIT Press; 2012.

Sehgal CM, Cary TW, Kangas SA, Weinstein SP, Schultz SM, Arger PH, Conant EF. Computer-based margin analysis of breast sonography for differentiating malignant and benign masses. J Ultrasound Med 2004;23:1201–1209.

Sehgal CM, Weinstein SP, Arger PH, Conant EF. Review of breast ultrasound. J Mammary Gland Biol Neoplasia 2006;11:113–123.

Sehgal CM, Cary SP, Cwanger A, Levenback BJ, Venkatesh SS. Combined naïve Bayes and logistic regression for quantitative breast sonography. Proc IEEE Int Ultrason Symp 2012;ULTSYM:1686–1689.

Song JH, Venkatesh SS, Conant EF, Arger PH, Sehgal CM. Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. Acad Radiol 2005;12:487–495.

Takemura A, Shimizu A, Hamamoto K. Discrimination of breast tumors in ultrasonic images using an ensemble classifier based on the AdaBoost algorithm with feature selection. IEEE Trans Med Imaging 2010;29:598–609.

Taylor P, Potts HW. Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer 2008;44:798–807.

Waldmann A, Kapsimalakou S, Katalinic A, Grande-Nagel I, Stoeckelhuber BM, Fischer D, Barkhausen J, Vogt FM. Benefits of the quality assured double and arbitration reading of mammograms in the early diagnosis of breast cancer in symptomatic women. Eur Radiol 2012;22:1014–1022.

Watermann DO, Foldi M, Hanjalic-Beck A, Hasenburg A, Lüghausen A, Prömpeler H, Gitsch G, Stickeler E. Three-dimensional ultrasound for the assessment of breast lesions. Ultrasound Obstet Gynecol 2005;25:592–598.

## APPENDIX: PSEUDOCODE FOR ROC CURVE AND AREA UNDER CURVE OF MARGINAL CLASSIFIER

*Function ROC Curve(boolean() d, double droprate)*

```
{
  //d is an array containing the true diagnosis for each case
  //assume that d is sorted in ascending order of the cases' ratings
  //so d(i) is the ith case's diagnosis, where i < j means rating(i) < rating(j)
  //droprate is the drop rate
  int n = length(d) //n is the number of cases
  int out = droprate*n //out is the number of cases dropped
  int in = n − out //in is the number of cases not dropped
  sen = new array(0 to in)
  spec = new array(0 to in)
  //array d is 0-indexed, so it goes from d(0) to d(n-1)
  for i = 0 to in
  {
    //if j < i, then cases are classified as benign.
    //if j ≥ i + out, they are classified as malignant
    //if i ≤ j < i + out, they are in the band and not classified
    TP = Number of cases j, with j ≤ i + out, and d(j) = true
    FP = Number of cases j, with j ≥ i + out, and d(j) = false
    TN = Number of cases j, with j < i and d(j) = false
    FN = Number of cases j, with j < i and d(j) = true
    sen(i) = TP/(TP + FN)
    spec(i) = TN/(TN + FP)
  }
  return (sen,spec)
}
```

*Function AreaUnderCurve(double() sen, double() spec)*

```
{
  double sum = 0
  int n = length(sen) //(length(sen) and length(spec) are equal)
  //sen, spec are 0 indexed, so they go from sen(0) to sen(n-1)
  for i = 1 to n-1
  {
    sum = (sen(i-1)-sen(i))*(spec(i-1)+spec(i-1))/2
  }
  return sum
}
```