34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Combining Multiple Feature Representations and AdaBoost Ensemble Learning for Reducing False-Positive Detections in Computer-aided Detection of Masses on Mammograms

Jae Young Choi, *Member*, *IEEE*, Dae Hoe Kim, Konstantinos N. Plataniotis, *Fellow*, *IEEE*, and Yong Man Ro, *Senior Member*, *IEEE*

Abstract— **One of the drawbacks of current Computer-aided Detection (CADe) systems is a high number of false-positive (FP) detections, especially for detecting mass abnormalities. In a typical CADe system, classifier design is one of the key steps for determining FP detection rates. This paper presents the effective classifier ensemble system for tackling FP reduction problem in CADe. To construct ensemble consisting of correct classifiers while disagreeing with each other as much as possible, we develop a new ensemble construction solution that combines data resampling underpinning AdaBoost learning with the use of different feature representations. In addition, to cope with the limitation of weak classifiers in conventional AdaBoost, our method has an effective mechanism for tuning the level of weakness of base classifiers. Further, for combining multiple decision outputs of ensemble members, a weighted sum fusion strategy is used to maximize a complementary effect for correct classification. Comparative experiments have been conducted on benchmark mammogram dataset. Results show that the proposed classifier ensemble outperforms the best single classifier in terms of reducing the FP detections of masses.**

I. Introduction

Breast cancer is the most common form of cancer among women and is the second leading cause of death. Early detection of breast cancer increases the survival rate and offers flexibility in terms of treatment option. Years of practice suggest that screening mammography [1] is a cost-effective approach for early detection of breast cancer. However, screening mammography is not a perfect diagnostic tool. On a screening mammogram, cancers can be missed (false-negative mammogram), and non-cancerous lesions can be mistaken as cancer, leading to a false-positive (FP) mammogram.

In general, clinical CADe systems have high sensitivity, but the difficulty is to achieve this at a low FP detection rate [2]. Especially, the FP rate for mass detection is much higher than that for detection of calcifications [2]. High FP detections not only reduce radiologists' productivity, but also increase the radiologists' recall rate [1-2]. In a typical CADe system, classifier is designed to perform the FP reduction phase where the detected suspicious regions, so-called region-of-interests (ROIs), are classified as mass vs. normal tissue [3]. Classifier design is, therefore, one of the key steps for determining FP detection rates [3]. Until so far, most previous studies on CADe have been mainly focused on the design of the single classifier system. It should be noted that there are two critical limitations within the classifier design process for CADe. First, the large variability in the appearance of mass patterns [4] ─ due to its irregular or obscured borders, and complex mixtures of margin types ─ makes this a quite difficult classification task. Second, research in mammography is characterized by a restricted training data due to cost, time, and availability to patient medical information and patient mammography images [4]. Due to the limitations mentioned above, it is extremely challenging to design a single classifier that properly classifies the entire instance (sample) space of a given mass input set.

Generally, different classifiers are likely to different errors on different input patterns [5], implying that classifiers differ in their decisions to complement each other. Thus, if the sets of mammographic mass patterns misclassified by different classifiers do not overlap to a certain extent, and also each classifier is accurate for a given particular local region in the instance space, suitably combining different classifiers could achieve better classification rates than those obtained using a single classifier. Further, using multiple classifiers, instead of a single classifier, can lead to improved generalization [6].

In this paper, we propose a new and novel multiple classifier system [5] (termed "classifier ensemble" hereafter) designed for reducing FP detections in CADe. Key characteristics of the proposed classifier ensemble solution are as follows. First, our method combines data resampling based on AdaBoost learning [5] with the use of different feature representations, aiming to create ensemble consisting of base classifiers that are as accurate as possible while disagreeing with each other as much as possible. Second, in order to extend the conventional AdaBoost framework to accommodate general (strong/weak) classifiers, we devise an effective strategy that controls the degree of weakness of base classifiers. Third, for combining multiple decision outputs of ensemble members, a weighted sum fusion strategy is used to maximize a complementary effect taken by classifier ensemble system. Experimental results using benchmark mammogram dataset show that the proposed classifier ensemble outperforms the best single classifier in terms of overall classification performance for test data, leading to the reduction of FP detections in CADe systems.

II. ROI Segmentation and Feature Extraction

General CADe algorithms consist of three stages: (1) segmentation of ROIs on mammogram; (2) feature extraction for generated ROIs; (3) classification of ROIs as mass or normal tissue. Note that segmentation of ROIs and feature extraction are prerequisite steps prior to performing classification of ROIs. In this section, for the sake of completeness, we briefly describe the segmentation and feature extraction approaches used in this paper. In our work, as recommend in [9], to perform a more realistic assessment of classification process, the ROI regions were automatically detected and segmented from each mammogram by using computer segmentation algorithm. For this purpose, one popular approach to using multi-level thresholding algorithm [10] was adopted for segmenting ROIs. We chose this segmentation

J. Y. Choi and K. N. Plataniotis are with Multimedia Lab, The Edward S Rogers Sr Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, M5S 3GA, Canada (jyg.choi@utoronto.ca; kostas@comm.utoronto.ca).

J. Y. Choi, D. H. Kim, and Y. M. Ro are with the Image and Video Systems Lab, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea (e-mail: jygchoi@kaist.ac.kr; dhkim10@kaist.ac.kr; ymro@ee.kaist.ac.kr).

TABLE I. FIVE TYPES OF FEATURES EXTRACTED FROM THE SEGMENTED ROIS.

Feature type	Description	No. of features
SGLD texture[a] [7]	Correlation, Energy, Entropy, Inertia, Inverse difference moment, Sum average, Sum variance, Sum entropy, Difference energy, Difference variance, Difference entropy, Information measure of correlation 1, Information measure of correlation 2	312
LBP texture [8]	LBP histograms were computed from core and margin regions of the segmented ROI. LBP operator with a circularly symmetric neighbourhood of P members on a circle radius of R was employed. Two LBP parameter settings, (P,R) =(8,1) and (P,R) =(8,3), were used, which results in two different LBP texture features.	118
GLDS texture[a] [7]	Contrast, Angular Second Moment, Entropy, and Mean	96
Morphological [7]	Circularity, Extent, Convexity, Solidity, Eccentricity, Elongatedness, Compactness, Area, NRL[b] mean, NRL standard deviation, NRL area ratio, NRL zero crossing count, NRL entropy	13
RBST texture[c] [9]	The SGLD matrices were calculated from the RBST image representation [10]. Eight texture measures, namely, "correlation, "energy", "difference entropy", "inverse difference moment", "entropy", "sum average", "sum entropy", and "inertia" were extracted from each SGLD matrix. The 40-pixel-wide band was used to construct the RBST images [9].	256

Note: [a]For SGLD and GLDS texture features, six different interpixel distances $d = \{1,2,4,6,8,10\}$ and four different angles $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ were used to produce 24 SGLD and 24 GLDS matrices, respectively [7]. [b]NRL=normalized radial length. [c]For RBST texture feature, eight different interpixel distances $d = \{1,2,3,4,6,8,12,16\}$ and four different angles $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ were used to calculate 32 SGLD matrices from the RBST image representation [9].

algorithm because it has been well-documented in previous publications that it provides "successful" segmentation results.

The mass or normal tissue ROIs are used as input for feature extraction. Five different types of feature representations, namely, SGLD, LBP, GLDS, RBST, and morphological features [7-9], were extracted from the segmented ROI to describe the texture, shape and margin characteristics in the breast tissue surrounding the mass. The features used in this study were summarized in Table I.

III. PROPOSED CLASSIFIER ENSEMBLE METHOD

The proposed classifier ensemble solution is based on the AdaBoost learning [5]. This algorithm is proven to be theoretically sound and shown to be empirically appealing because of its simplicity and superior performance in many domains. The proposed ensemble method differs from previous work on designing AdaBoost ensemble in the following aspects.

- The conventional AdaBoost ensemble is to use only a data resampling technique to create a set of base classifiers as ensemble members. The idea behind data resampling is that classifiers generated from different samples of the training data are likely to make errors in different ways. In the proposed method, to generate more diverse and accurate base classifiers, the combined use of different feature representations (of the same object) and data resampling has been developed.
- It is widely accepted that AdaBoost learning rules are not suited to a strong and stable classifier [5] such as Support Vector Machines (SVM). To overcome the limitation of weak classifier, the proposed method has an effective mechanism devised for regulating the degree of weakness of the classifiers by adjusting the size of a resampled set. This allows the conventional AdaBoost framework to be applicable to work with general (weak/strong) classifiers.

Fig. 1 provides a description for the proposed classifier ensemble algorithm. Let $\mathbf{T}$ be a training set composed of N instances (i.e., ROI images) each denoted by $x_i\,(i = 1\ldots, N)$ with a corresponding class label y_i, where $y_i \in \{0,1\}$. Assuming that a total of 'K' different feature representations of a given ROI are yielded from the feature extraction as explained in Section II, we then denote the m-th feature representation by f_m (e.g., LBP or SGLD texture features described in Table I) comprising a feature pool denoted by $\mathbf{F}$ for which $f_m \in \mathbf{F}$. To maintain a set of weights over the $\mathbf{T}$, the distribution denoted by $D_t(i)$ for each training sample x_i can be determined at every boosting round. Initially, values of $D_t(i)$ are set equally, but on each round, they are newly updated in such a way that base classifiers is forced to focus on the hard-to-classify training samples.

It should be noted that as described in Step 2.(2) in Fig. 1, parameter r is devised to determine the size of resamples set (used to train a base classifier) during the process of data resampling, which is a portion of the training set. Note that the amount of samples in a resampled set is directly proportional to the value of r. Hence, a smaller/large r value will equivalently lead to a weak/strong (i.e., more/less accurate) base classifier, given the same classifier model. From the diversity point of view, a resampled set with a smaller r value will lead to more diverse classifiers as ensemble members. Referring to Murua's bound [11], to achieve a low generalization error, the boosting procedure should not only create base classifiers with large expected margins, but also keep their dependence low. Here, large expected margins are directly linked to the classification accuracies of individual members in an ensemble [11]. Based on this fact, we find the optimal balance between the individual accuracies (i.e., the degrees of weakness) of ensemble members and their mutual dependence by adjusting the value of r. Considering overall classification accuracy in our experiments, a good compromise has been found by setting r in the range of [0.4, 0.6].

Note that much literature [5-6] dealing with classifier ensemble suggests that, as a vital requirement for the success of the ensemble, the ensemble members should be as correct as possible, at the same time, they should not make coincident errors. In light of this fact, base classifiers are produced by using both data resampling and different feature representations of the same input, as described in Step 2.(3) in Fig. 1. In particular, by using different and multiple feature representations, different base classifiers try to learn different parts of the input instance space during training. The underlying idea for this approach is that different feature representations make different characteristics apparent and an object ambiguous in one representation may be clearly recognizable in another representation.

Training phase for ensemble construction

0. (Input)

(1) Feature pool $\mathbf{F} = \{f_m, m = 1,\ldots,K\}$

(2) Training data set $\mathbf{T}$ consisting of N labeled instances $\{(x_i, y_i)\}_{i=1}^{N}$ with class labels $y_i \in \{0,1\}$

(3) Total number of boosting rounds T

1. (Initialization)

(1) Weight distribution $D_0(i) = 1/N$, for $i = 1,\ldots,N$ for N training samples included in a $\mathbf{T}$

(2) Weight vector $w_{1,i} = D_0(i)$ for $i = 1,\ldots,N$

(3) $\mathbf{E}_0 = \{\phi\}$ (Ensemble including classifier models)

2. (Repeat for $t = 1,\ldots,T$)

(1) Compute the distribution for each training sample $D_t(i) = \dfrac{w_{t,i}}{\sum_{i=1}^{N} w_{t,i}}$

(2) Select 'r' % hardest training samples per class according to the distribution to form a resampled (or training subset) set $\mathbf{T}_t$ ($\mathbf{T}_t \subset \mathbf{T}$)

(3) For $m = 1,\ldots,K$

- Build a base classifier $h_{t,m}$ for each feature f_m (along with the m-th feature) using $\mathbf{T}_t$
- Calculate the weighted classification error $\varepsilon_{h_{t,m}}$ each produced by $h_{t,m}$ using $\varepsilon_{h_{t,m}} = \sum_{i=1}^{N} D_i(i)\,|h_{t,m}(x_i) - y_i\,|$

(4) Construct candidate classifiers denoted by $\mathbf{H}_t = \{h_{t,m}\}_{m=1}^{K}$

(5) Determine the best base classifier h_t with the lowest error ε_{h_t} from $\mathbf{H}_t$, such that $h_t = \arg\min_{h_{t,m}} \varepsilon_{h_{t,m}}$

(6) Define the error ε_t of the best base classifier $\varepsilon_t = \varepsilon_{h_t}$

(7) If $\varepsilon_t = 0$ or $\varepsilon_t > 0.5$, ignore h_t, reinitialize the distribution $D_t(i)$ to $1/N$ and go to step 2.(2)

Else, calculate $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$ and $\mathbf{E}_t = \mathbf{E}_{t-1} \cup \{h_t\}$

(8) Update weight vector $w_{t+1,i} = w_{t,i}\beta_t^{1-h_t(x_i)-y_i|}$

3. (Output)

Classifiers ensemble $\mathbf{E}_t = \{h_t\}_{t=1}^{M}$ and corresponding weights $\{1/\beta_t\}_{t=1}^{M}$, where $M \le T$

Testing phase for classification

1. (Calculate the support (or confidence) for each sample by using the weighted average combiner)

$h^j_{\text{combined}} = \sum_{t=1}^{M} (1/\beta_t) h_t(x)$

2. (Use h^j_{combined} as decision variable in Receiver Operating Characteristic (ROC) analysis)

Fig. 1. Proposed classifier ensemble algorithm using the combination of AdaBoost data resampling with the different feature representations. Note that each resampled set contains 'r'% training samples per class of the original training data to control the weakness of classifiers. Also note that, as recommended by [5], if a classifier has an error rate greater than 1/2 in a trial, then we reinitialize the training set weights to the uniform distribution and continue drawing samples.

Thus, different feature representations are likely to provide complementary information for correct classification. Consequently, this allows producing ensemble that emphasizes more diversity in ensemble members.

It should be also noted that another key to successful ensemble methods is to construct base classifiers with small error rates [5]. To account for this, in our method, the best base classifier $h_t(\cdot)$ (at each boosting round t) for classifying a weighted version of $\mathbf{T}$ (i.e., weighted training samples) is determined as follows:

$$h_t = \arg\min_{h_{t,m}} \varepsilon_{h_{t,m}} \tag{1}$$

and

$$\varepsilon_{h_{t,m}} = \sum_{i=1}^{N} D_i(i)\,|h_{t,m}(x_i) - y_i\,| \tag{2}$$

where $D_i(i)$ denotes weight distribution for each training sample and $h_{t,m}(\cdot)$ is a base classifier trained with the m-th feature representation f_m. Note that in (2), without loss of generality, we can assume that the outputs of each classifier $h_{t,m}(\cdot)$ span the space in the range of [0,1]. Also note that $\varepsilon_{h_{t,m}}$ represents the weighted classification error produced by $h_{t,m}(\cdot)$. Using (1), among various individual base classifiers for each feature representation, we select the best base classifier (trained with a particular feature representation) that yield the most accurate results on a weighted training set according to the distribution where difficult samples have a great probability of being selected, and easy samples have less chance of being used for training. With this mechanism, we expect to produce more specialized base classifier each focusing on a smaller section of the instance input space consisting of hard-to-classify samples, and they can be more accurate than those generated using only a single feature representation during training.

After terminating our ensemble construction, a ensemble of M classifiers, denoted by $\mathbf{E}_t = \{h_t\}_{t=1}^{M}$, is used for performing classification. To combine the outputs of the M classifiers in an ensemble, weighted average combiner is used as follows:

$$h^j_{\text{combined}} = \sum_{t=1}^{M} (1/\beta_t) h_t(x) \tag{3}$$

TABLE II. COMPARISONS OF CLASSIFICATION ACCURACIES BETWEEN A SINGLE CLASSIFIER AND PROPOSED CLASSIFIER ENSEMBLE, WITH RESPECT TO SIX DIFFERENT FEATURE REPRESENTATIONS. IN OUR CLASSIFIER ENSEMBLE, THE PARAMETER r (DESCRIBED IN FIG. 1) WAS SET TO 0.4.

Feature representation	SVM base classifier		NN base classifier	
	Az for single classifier	*Az* for proposed ensemble	*Az* for a single classifier	*Az* for proposed ensemble
LBP (P=8,R=1)	**0.849 ± 0.024**	**0.917 ± 0.018**	0.814 ± 0.026	**0.911 ± 0.021**
LBP (P=8,R=3)	0.844 ± 0.011		**0.819 ± 0.022**	
SGLD	0.823 ± 0.092		0.804 ± 0.032	
GLDS	0.771 ± 0.026		0.734 ± 0.038	
RBST	0.834 ± 0.034		0.801 ± 0.041	
Morphological	0.831 ± 0.021		0.792 ± 0.023	

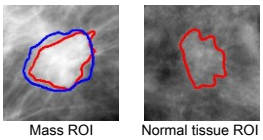

Fig. 2. Example of ROIs used in our experiments. The red line is segmented contour identified by segmentation algorithm, while the blue line is the mass outline (as ground truth) marked by experienced radiologists.

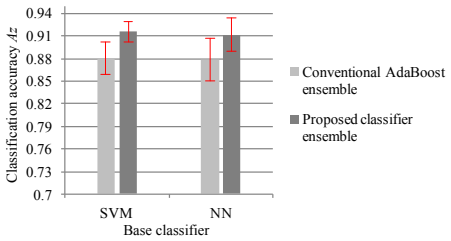

Fig. 3. Comparisons of classification accuracies between conventional AdaBoost ensemble [5] and proposed classifier ensemble. Note that in the proposed ensemble, the parameter 'r' was set to 0.4.

where $\beta_t = \varepsilon_t/(1-\varepsilon_t)$ and ε_t is classification error of the h_t. Note that in (3), the weights $1/\beta_t$ depend on the classifier's expertise in a given input instance region, and the fusion based on weights enables more competent classifiers (in terms of accuracy) to have a greater power in making the final decision.

IV. EXPERIMENTS

The public database of mammograms used in this study is Digital Database for Screening Mammography (DDSM) DB [4]. The 303 single-view mammograms were selected from the DDSM DB in our experiments. As described in Section II, using computer segmentation, a total of 2,742 ROIs were automatically generated (see Fig. 2): 246 masses and 2,496 normal tissues (i.e., FP regions). Note that a generated ROI was considered as a true mass only if it met the criteria proposed in [4], [10]. A total of 2,742 ROIs were randomly divided into two sets of equal size: training and testing sets. Note that, to guarantee stable classification results, 20 independent runs of aforementioned random partitions were executed. Thus, all of the results reported in this section were averaged over 20 runs. As for base classifiers, SVM which utilizes a Radial Basis Function [5] (as kernel) and Neural Network (NN) with back-propagation algorithm [5] were used. The classification outputs were used as the decision variable in Receiver Operating Characteristics (ROC) analysis [4] to evaluate the classification performance. In our experiments, the area under the ROC curve denoted by A_z was used as an index of classification accuracy. Note that the area under the ROC curve is the most widely used criterion to evaluate the overall performance of FP reduction in CADe systems [3-4].

Table II shows comparisons of classification accuracies using the area under ROC curves, A_z. For comparison purposes, a single classifier trained with a particular feature representation was used. For SVM and NN base classifiers, it can be seen that the proposed classifier ensembles yield much better classification accuracies than those obtained using all single classifiers. In particular, comparing to the best single classifier, the values of A_z considerably increase with about 0.068 and 0.092, in the order of SVM and NN, respectively.

In Fig. 3, we report the comparisons of classification accuracies between conventional AdaBoost ensemble [5] and proposed classifier ensemble. For comparison, classification accuracies obtained for the most accurate AdaBoost ensembles (for SVM and NN, respectively) were presented in Fig. 3. Specifically, for each base classifier, the most accurate AdaBoost ensemble was selected based on testing accuracies obtained using different AdaBoost ensembles ─ each learned with a particular feature representation. As shown in Fig. 3, the proposed approach to combining data resampling and different feature representations allows improving the ensemble accuracy for both cases of using SVM and NN. These results indicate that the proposed ensemble construction may be beneficial in terms of generating an ensemble of classifiers that are more accurate, as well as achieve better diversity.

V. CONCLUSION

In this paper, we propose new classifier ensemble solution for improved classification of mammographic masses from normal tissues. Comparative results show the potential clinical effectiveness of our method in terms of considerably reducing FP detections in CADe systems. For future work, we plan to incorporate ensemble evaluation module into our current ensemble construction framework to select an optimal subset of given ensemble members.

REFERENCES

[1] T.W. Freer and M.J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center," *Radiology*, vol. 220, no. 3, pp. 781-786, 2001.

[2] R. M. Nishikawa, "Current status and future directions of computer-aided diagnosis in mammography, " *Computerized Medical Imaging and Graphics,* vol. 31, no. 3, pp. 224-235, 2007.

[3] M. P. Sampat, "Computer-aided detection and diagnosis in mammography". in *Handbook of Image and Video Processing*, A.C. Bovik, Ed., 2nd ed. New York: Academic, 2005, pp. 1195-1217.

[4] J. S. Suri and R. M. Rangayyan, *Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer*, SPIE PRESS, Washington, 2006.

[5] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, WILEY, 2004.

[6] N. Ueda and R. Nakano, "Generalization Error of Ensemble Estimates, " *IEEE Int'l Conf. on Neural Networks*, 1996.

[7] H.D. Cheng, X.J. Shi, R. Min, L.M. Hu, X.P. Chi, and H.N. Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern Recognit*., vol. 39, no. 4, pp. 646-668, 2005.

[8] Jae Young Choi, Dae Hoe Kim, and Yong Man Ro, "Combining Multiresolution Local Binary Pattern Texture Analysis and Variable Selection Strategy Applied to Computer-Aided Detection of Breast Masses on Mammograms," *IEEE-EMBS Int'l Conf. on Biomedical and Health Informatics*, 2012.

[9] B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, and L.M. Hadjiiski, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Med. Phys*., vol. 25, no. 4. pp. 516-526, 1998.

[10] A. R. Dominguez and A. K. Nandi, "Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection," *Computerized Medical Imaging and Graphics*, vol. 32, no. 4, pp. 304-315, 2008.

[11] A. Mura, "Upper bounds for error rates of linear combinations of classifiers," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 24, no. 5, pp. 591-602, 2002.