Biomed Eng Lett (2015) 5:251-262
DOI 10.1007/s13534-015-0191-1

ORIGINAL ARTICLE

A Generalized Multiple Classifier System for Improving Computer-aided Classification of Breast Masses in Mammography

Jae Young Choi

Received: 1 December 2014 / Revised: 30 March 2015 / Accepted: 7 July 2015

Abstract
Purpose The objective of this paper is to present a generalized multiple classifier system for improved classification of mammographic masses in Computer-aided detection (CAD).
Methods To encourage different base (component) classifiers to learn different parts of an object instant space, we develop a novel base classifier generation algorithm which combines data resampling underpinning AdaBoost with the use of different feature representations. In addition, our proposed multiple classifier system can be generalized beyond the limitation of weak classifiers in conventional AdaBoost learning. To this end, our multiple classifier system has an effective and efficient mechanism for tuning the level of weakness of base classifiers.
Results Extensive experiments have been performed using benchmark mammogram data set to test the proposed method on classification between mammographic masses and normal tissues. In addition, to assess classification performance, we used the area under the receiver operating characteristic (AUC) and the normalized partial area under the curve ($_p$AUC). Results show that our method considerably outperforms (in terms of both AUC and $_p$AUC) the most commonly used single neural network (NN) and support vector machine (SVM) based classification approaches. In particular, the effectiveness of our method in terms of correct classification is much more significant over difficult mammogram cases with dense tissues that have higher risk of cancer incidences and cause higher false-positive (FP) detections.
Conclusions Our multiple classifier system shows quite promising results in terms of improving classification performances on the FP reduction application using classification between masses and normal tissues in mammography CAD systems.

Keywords Multiple classifier system, Mammographic masses, Generalization, False-positive reduction, Computer-aided detection (CAD)

Jae Young Choi (✉)
Biomedical Informatics and Pattern Recognition (BPR) Lab., Dept. of Biomedical Engineering, Jungwon University, 85 Munmu-ro Goesan-eup Goesan-gun, Chungcheongbuk-do 367-805, Republic of Korea
Tel : +82-42-830-8812 / Fax : +82-43-830-8812
E-mail : jyoung.choi@jwu.ac.kr

INTRODUCTION

Breast cancer is the most common form of cancer among women and is the second-leading cause of death [1-3]. To reduce the workload of radiologists and to improve the specificity and sensitivity in detection of breast cancer, Computer-aided detection (CAD) are being developed [1, 4-6]. Current mammography CAD systems have been clearly shown to be quite sensitive in its ability to detect cancer, but one of their main drawbacks is the high number of false-positive (FP) [4-6]. Hence, high FP rate for mass detection and diagnosis remains to be one of the major problems to be resolved in CAD study [4, 5].

In typical CAD systems, classifier design is one of the key steps for determining FP rates [4, 5]. Thus far, research efforts have mostly been focused on the design of the *single* classifier in CAD systems [4-7, 13, 23-26]. Wei *et al.* [13] used global and local texture features extracted from manually selected region of interest (ROI) of digitized mammograms, and linear discriminant analysis (LDA) to classify the masses from normal glandular tissues to minimize FP detections. Sahiner *et al.* [23] proposed a convolution neural network (NN) for the task of discriminating between masses and normal tissues using texture features. The authors in [24] developed a NN classifier based on multiresolution texture features extracted from the spatial gray level dependence (SGLD) matrices for distinguishing masses from normal tissues. In [25], the four texture features, namely contrast, coherence ratio, entropy of orientation, and variance of

coherence-weighted angular estimates, were extracted based on textual flow-field analysis and were used to reduce FP detections. Kupinski *et al.* [26] studied a regularized NN classifier to differentiate masses from normal tissues based on intensity, iso-intensity, location, and contrast features.

It should be noted that there are two critical limitations within the classifier design process in mammogram images. First, the large variability in the appearance of mass patterns [8, 9] – due to its irregular size, obscured borders, and complex mixtures of margin types – make classification task quite difficult. 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, 10, 11]. On the other hand, the number of available features (due to integration of multiple heterogeneous feature types) is large [8, 12, 13] (typically, in the thousands) relative to the number of training samples (curse of dimensionality [14]). For these reasons, a *single classifier design* may face a great challenge in achieving a level of FP reduction that meets the requirement of clinical applications.

In this paper, we propose a novel *multiple classifier scheme* for reducing false-positive detections in mammographic CAD system. Our multiple classifier system has the following key significances over existing multiple classifier based classification techniques [12, 15-18, 27-30].

- Key characteristics of our approach are to generate individual base classifiers each trained with a corresponding feature type and to select *the best base classifier* (learning with the best feature type) at each boosting round. This strategy enables *accommodating multiple, various feature types* for improved classification by alleviating the curse of dimensionality when the number of training samples is limited. This is also advantageous to produce more *specialized base classifiers* each focusing on a smaller section of the instance space consisting of particularly hard-to-classify object samples.
- It is generally believed that a typical AdaBoost learning would not be suited to a strong and stable classifier [19, 20], such as Support Vector Machine (SVM). A *weak learner limitation* may restrict the applicability of the AdaBoost learning in practical applications, especially for mammographic CAD in which most of the state-of-the-art classification approaches involve the use of a strong classifier [5, 6, 21, 22]. To break the aforementioned limitation, we design a *generalized multiple classifier* system that works well with general (both strong and weak) classifiers extensively used in mammographic CAD systems. For this, we devise a simple but effective strategy that regulates the *degree of weakness of base classifiers.* This can be achieved by adjusting the size of a resampled set.

METHODS AND MATERIALS

ROI segmentation and feature extraction

In typical mammography CAD systems, segmentation of ROIs and feature extraction for generated ROIs are prerequisite steps prior to performing classification of ROIs [5-6, 8]. Hence, in this section, we will briefly describe the segmentation algorithm and types of mammographic mass features used in our study before explaining in detail the proposed multiple classifier framework.

As recommended in [13, 23, 25] to perform a more realistic assessment of a classification process, the ROI regions were automatically detected and segmented from each mammogram by using a fully automated segmentation. For this purpose, one popular approach to using multi-level thresholding algorithm [32, 33] was adopted for segmenting masses. The implementation details on a used segmentation algorithm have been described in literature [25, 32, 33]. Fig. 1 shows example of a mammogram with detected suspicious regions, as well as segmented ROIs generated by used segmentation algorithm. Note that in Fig. 1, the red line is the successfully segmented ROI contour identified by segmentation algorithm and the blue line is the mass outline (as ground truth) marked by experienced radiologists.

The ROIs were used as input for feature extraction. The features used in our study can be divided into five feature subspaces (sets): texture, intensity, shape (or morphological), margin, and spiculation feature subspaces. The features sorted by the subspace are summarized in Table 1. Note that the features described in Table 1 were used as different *feature representations* used to construct a group of base classifiers as members in our multiple classifier system.

Base classifier generation

As shown in Fig. 2, the proposed multiple classifier

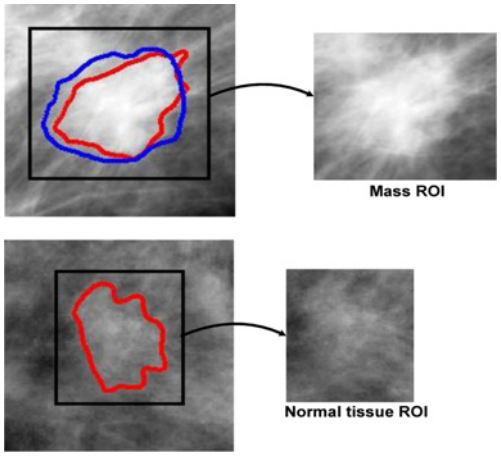

Fig. 1. Some examples of segmented mass and normal tissue ROIs. See text for explanation.

Table 1. Description for feature representations used to implement the proposed multiple classifier system. NC is abbreviation of 'number of components' for each feature representation.

Subspace	Feature representation description	NC
Texture	*SGLD Features* [13] 13 features, namely, "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" are extracted from each SGLD matrix at six different interpixel distances (d = 1, 2, 4, 6, 8, and 10) and in four directions (θ = 0°, 45°, 90°, and 135°), yielding 24 SGLD matrices; this results in a total of 312 features for each ROI (24 SGLD matrices x 13 features)	312
	Local Binary Pattern (LBP) Features [34] LBP histograms are computed from core and margin regions of the segmented object; LBP operator with a circularly symmetric neighbourhood of P members on a circle radius of R is employed; the three-resolution combination is used by setting LBP parameters (P,R) values of (8,1), (8,2), and (8,3)	357
	Run Length Statistics (RLS) Features [35] Five features, namely, "short run emphasis", "long runs emphasis", "gray-level nonuniformity", "run-length nonuniformity", and "run percentage" are obtained from the gray level run length matrices with four directions, θ = {0°, 45°, 90°, 135°}; hence, a total of 20 RLS-based features were calculated for each ROI image	20
	Gray-level Difference Statistics (GLDS) Features [23] Four features "contrast", "angular second moment", "entropy", and "mean" are extracted from the gray level difference statistics vector; six different interpixel distances (d = 1, 2, 4, 6, 8, and 10) and four directions (θ = 0°, 45°, 90°, and 135°) are used to calculate 24 GLDS vectors, yielding 96 GLDS features	96
	Rubber-band Straightening Transform (RBST) Features [36] Using eight different pixel pair distances (d = 1, 2, 3, 4, 6, 8, 12 and 16) and in four directions (θ = 0°, 45°, 90°, and 135°), SGLD matrices are calculated from the RBST image representation; eight features, namely, "correlation", "energy", "difference entropy", "inverse difference moment", "entropy", "sum average", "sum entropy", and "inertia" are extracted from each SGLD matrix; the 40-pixel-wide band was used to construct the RBST images	256
	Texture-flow field Features [25] Four features, namely, "contrast", "coherence ratio", "entropy of orientation", and "variance of coherence-weighted angular estimates" are extracted based on textual flow-field analysis	4
Shape	Circularity, Extent, Convexity, Solidity, Eccentricity, Elongatedness, Compactness, Area [8]	8
	Normalized Radial Length (NRL) Features [8] NRL mean, NRL standard deviation, NRL area ratio, NRL zero crossing count, NRL entropy	5
Intensity	Contrast measure, Average gray level, Standard deviation, Skewness, Kurtosis [8]	5

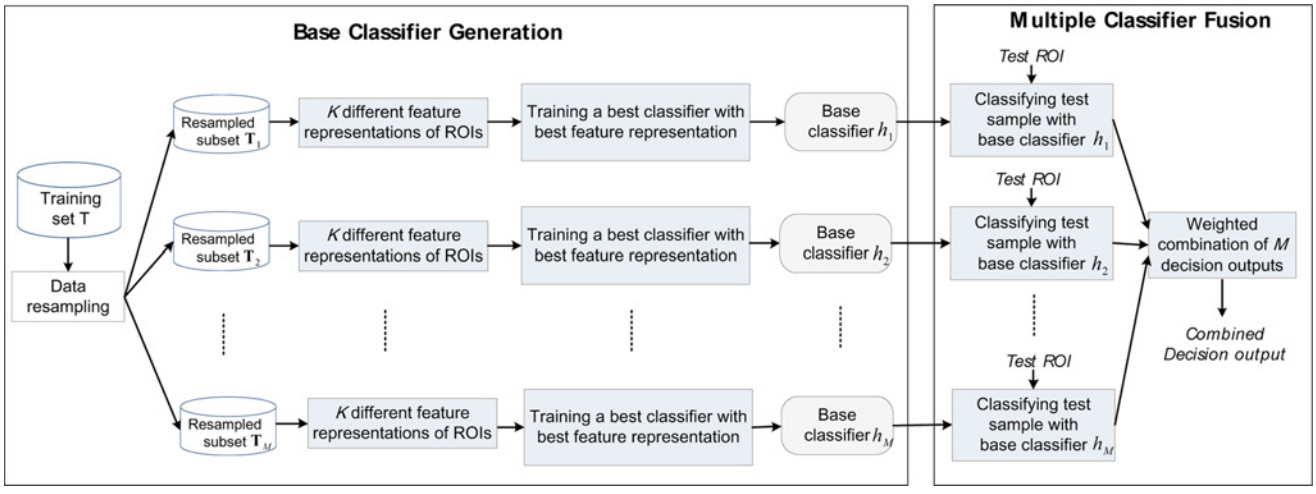

Fig. 2. Overview of the proposed multiple classifier framework.

framework largely consists of two parts: (a) base classifier generation and (b) multiple classifier fusion (or combination). Each of the two parts described in Fig. 2 will be explained in detail in this subsection and the following subsection, respectively. Let **T** be a training set composed of N samples (i.e., ROI images), each denoted by $x_i (i = 1..., N)$ with a corresponding class label l_i, where $l_i \in \{0,1\}$. Assuming that a total of K different feature representations of a given

ROI are yielded from the feature extraction as explained in Table 1, we then denote the m-th feature representation by f_m (e.g., LBP feature representation described in Table 1) comprising a feature pool denoted by $\mathbf{F}$ for which $f_m \in \mathbf{F}$. To maintain a set of weights over the $\mathbf{T}$ at each t-th boosting round, the distribution $D_t(x_i)$ on the training sample x_i can be determined as follows [10]:

$$D_t(x_i) = \frac{w_{t,i}}{\sum_{i=1}^{N} w_{t,i}}, \text{ for } i = 1,..., N \quad (1)$$

where $w_{t,i}$ denotes the weight for the i-th training sample on the t-th round. In Eq. (1), the weight values of initial distribution, denoted as $D_0(x_i)$, are set equally such that $D_0(x_i) = 1/N$. Note that $w_{t,i}$ in Eq. (1) is, in fact, computed based on classification error associated with each training sample (for details, please see Eq. (5)). Hence, the values of $D_t(i)$ increase as the likelihood of difficult samples for classification is increased. In light of this fact, during the base classifier generation, *hard-to-classify degree* for each sample has been measured by directly using the values of $D_t(i)$.

As described in Fig. 2, data resampling [14] underpinning AdaBoost learning is performed to form a *resampled subset* (denoted by $\mathbf{T}_t$) – which is a selectively sampled from the set $\mathbf{T}$. It is important to note that parameter $r(0 < r < 1)$ is devised for the formation of the $\mathbf{T}_t$, aiming to adjust the level of *weakness of base classifiers*. For this, a resampled subset $\mathbf{T}_t$ is formed in the following way:

$$\mathbf{T}_t = \{x_i | \Omega(D_t(x_i)) \le (r \times N)\} \quad (2)$$

where $\Omega(\cdot)$ is a function that returns a rank (order) index of $D_t(x_i)$, assuming that values of $D_t(x_i)$, $i = 1,..., N$, are sorted in the descending order on the interval $\left[\min_{i=1}^{N} D_t(x_i), \max_{i=1}^{N} D_t(x_i)\right]$ and N denotes the total number of training samples. Note that Eq. (2) means that a resampled subset is constructed by selecting $(r \times 100)$% hardest training samples, according to the distribution $D_t(i)$. The amount of samples contained in $\mathbf{T}_t$ is therefore directly proportional to the value of r such that $|\mathbf{T}_t| \approx r|\mathbf{T}|$.

The rationale behind the use of r is that referring to [37, 38], increasing the size of the training set generally leads to improved classification performance of classifier learning algorithms. In particular, in classification applications (e.g., in mammography) where a large number of different features are often used, and the decision rule is complex, it has been reported that there is stronger tendency that classification performances improve as the number of training samples becomes large [9]. Based on the aforementioned fact, it is reasonable to assume that a smaller/larger r value will equivalently lead to a weak/strong (i.e., more/less accurate) base classifier, given the same classifier model. The best classification performance was achieved when r was set between 0.4 and 0.6.

Let $h_{t,m}(\cdot)$ be a base classifier trained with the m-th feature representation f_m and the t-th resample subset $\mathbf{T}_t$. Without loss of generality, we assume that the outputs of the $h_{t,m}$ span the space in the range of [0, 1] such that $h_{t,m}: \mathbf{F} \times \mathbf{T}_t \rightarrow [0,1]$. It should be noted that main focus of our base learning is on the application of *classifier models* (such as SVM) into our proposed multiple classifier framework, rather than feature extraction using dimensionality reduction techniques such as Principal Component Analysis (PCA) [14]. In this context, $h_{t,m}$ should be built by using *the classifier models* suited for implementing AdaBoost framework. Also note that $h_{t,m}$ $(m = 1,..., K)$ are produced by combining data resampling and K different feature representations of the same input. The underlying idea behind this approach is that various and different feature representations make different characteristics apparent and a mass object ambiguous in one representation may be clearly recognizable in another different representation [12, 39]; hence, it should be understood that the use of multiple feature representations allows increasing diversity between base classifiers in the sense that they do not make coincident errors [40].

Among $h_{t,m}(m = 1,..., K)$, a best base classifier h_t (for each round t) for classifying weighted training samples is determined by selecting a best feature representation as follows:

$$h_t = \arg\min_{h_{t,m}} \varepsilon_{h_{t,m}} \quad (3)$$

and

$$\varepsilon_{h_{t,m}} = \sum_{i=1}^{N} D_t(x_i) \left| h_{t,m}(x_i) - l_i \right| \quad (4)$$

where $\varepsilon_{h_{t,m}}$ represents the weighted classification error produced by $h_{t,m}$. Using Eq. (3), among K individual base classifiers – each trained with a particular feature representation, we select a best base classifier h_t that yields the most accurate results on a weighted training set.

Based on the classification error of a best base classifier h_t (generated at t-th round), weight at $(t + 1)$-th round for each training sample x_i can be updated as follows [15]:

$$w_{t+1,i} = w_{t,i} \beta_t^{|1 - h_t(x_i) - l_i|} \quad (5)$$

$\beta_t = \varepsilon_t / (1 - \varepsilon_t)$ and ε_t is the corresponding classification error associated with the h_t [ε_t can be easily calculated by substituting h_t for $h_{t,m}$ shown in Eq. (4)]. In Eq. (5), the values of weights progress towards increasing the probability that difficult samples are being selected. This forces a *best base classifier* to be generated at next round to focus on the hard-to-classify training samples. The detailed implementation steps of the proposed multiple classifier generation are described in Fig. 3.

0. (Input)

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

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

(3) Total number of multiple classifier generation rounds T

1. (Initialization)

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

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

(3) $\mathbf{E}_0 = \{\phi\}$ (Generated multiple classifier including base classifiers)

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

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

(2) Using parameter r $(0 < r < 1)$, select $(r \times 100)\%$ hardest training samples per class (as the proportion of whole training samples) according to the distribution to form a resampled subset $\mathbf{T}_t$ $(\mathbf{T}_t \subset \mathbf{T})$

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

- Build a base classifier $h_{t,m}$ using both f_m and $\mathbf{T}_t$
- Calculate weighted classification error $\varepsilon_{h_{t,m}}$ for $h_{t,m}$ using $\varepsilon_{h_{t,m}} = \sum_{i=1}^{N} D_t(x_i) |h_{t,m}(x_i) - \ell_i|$

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

(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 a 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 $w_{t+1,i} = w_{t,i} \beta_t^{|1 - h_t(x_i) - \ell_i|}$

3. (Output)

Generated base classifiers $\mathbf{E} = \{h_t\}_{t=1}^{M}$ and corresponding confidences $\{1/\beta_t\}_{t=1}^{M}$, where $M \le T$

Fig. 3. Our proposed base classifier generation algorithm. As recommended by [14], if a best base classifier has an error rate greater than 1/2 in a trial (at each boosting round), then we reinitialize the weight distribution (for training samples) to the uniform distribution and continue drawing samples. Note that when the weight distribution is uniform, $(r \times 100)\%$ training samples per class are randomly selected from the training set **T**.

Multiple classifier fusion

Assuming that the M base classifiers are produced after terminating the base classifier generation process described in the aforementioned section, the weighted combination [14, 41] is adopted to perform the fusion of multiple base classifiers that aims at combining the M decision outputs as is described in Fig. 2. Based on weighted combination rule, multiple classifier fusion is performed as follows:

$$h_{\text{combined}} = \sum_{t=1}^{M} \alpha_t h_t(x) \tag{6}$$

where $\alpha_t = \dfrac{1/\beta_t}{\sum_{t=1}^{M} 1/\beta_t}$ (please refer to step 2. (7) in Fig. 3 for the definition of β_t). Note that since $1/\beta_t$ is monotonically increasing as ε_t becomes smaller, $1/\beta_t$ would be a reliable indicator of representing the confidence (or significance) of decision outputted by h_t [10]. In Eq. (6), weights α_t are normalized values of $1/\beta_t$ such that $0 < \alpha_t < 1$ and $\sum_{t=1}^{p} \alpha_t = 1$. The weights α_t depend on the classifier's expertise in a given input instance region. Thus, the fusion based on weights enable more competent classifiers (in terms of accuracy) to have a greater power in making the final decision.

Data set and performance evaluation

The public Digital Database for Screening Mammography (DDSM) database (DB) was in our evaluation study [42]. For data consistency purposes, all images were collected from the same type of scanner and resolution. We chose the scanner type Howtek 960 because a large number of cases are digitized by this type [42]. All images collected from the DDSM were subsampled to 200 μm and quantized to 8 bits per pixel for computational efficiency [43].

To evaluate the proposed multiple classifier, the data set was designed for assessing classification performances under clinical CAD application (i.e., classifying suspicious ROIs

into mass versus normal tissue). To this end, using a computer segmentation described in previous section, a total of 2,743 ROIs were automatically generated by using 303 mammograms collected from the DDSM. Referring to literature [4, 46], it can be desirable that high sensitivity rate should be maintained, prior to performing FP reduction stage. In light of this fact, during the generation of ROIs, we chose to operate threshold, which led to an average number of around 8.2 FPs per image at a detection sensitivity of about 82% [46]. With this fixed operating threshold, a total of 2,743 ROIs were automatically generated: 246 mass and 2,497 normal tissue ROIs

As described in Fig. 1, the DDSM provides manual annotations of the true masses presented in each image. These annotations were considered as the ground truth in our experiments. Using given manual annotations, a generated ROI was determined as a true positive (mass) only if it met the following two criteria [44, 47] (extensively used in CAD algorithms of breast masses): (1) the centroid of a segmented ROI region is included in the DDSM annotated area, and (2) a segmented ROI region intersects with the true mass region more than 25%. It is important to note that the masses with different shapes and density found in clinical practice were well represented in our data set by containing a wide variety of mass shapes, margin characteristics, and breast densities.

In order to guarantee the *stability* of evaluating classification approaches, we employed the most widely used 5 × 2-Fold cross-validation (cv) [14]. 5 × 2-Fold cv consists of repeating a two-fold cv procedure five times. In each cross-validation run, we divided the data set into training and testing halves. The roles are swapped at each fold to generate ten training and testing sets, yielding the final classification accuracy computed by averaging 10 corresponding results. In this way, the classification accuracy was estimated reliably and in an unbiased way.

We used the area under the receive operating characteristic (ROC) curve abbreviated as "AUC" [4, 45] to evaluate the classification performance. ROC analysis evaluates the relationship (i.e., trade-off) between the sensitivity and the false-positive rate (FPR) at differing classification decision thresholds. It should be noted that AUC has been most widely used index for evaluating the overall performances of classifiers in CAD systems [4, 45, 46]. For real-life applications, the specificities at high sensitivity levels are important because missing a cancer is a greater risk to patients than performing a biopsy to assess a lesion [3, 4, 46]. To address this issue, the normalized partial area index ($_{p}$AUC) was also evaluated, where p indicates the lowest acceptable sensitivity level. Further details on $_{p}$AUC are given in [45, 46]. In our experiments, a sensitivity level (threshold) of 90% was used for the computation of partial area index, denoted as "$_{0.9}$AUC".

RESULTS AND DISCUSSION

Evaluating classification of mass and normal tissues

The proposed multiple classifier solution was tested for assessing its effectiveness on classifying mass versus normal tissues. Note that nine types of features each described in Table 1 were used as different feature representations in this assessment [i.e., K (defined in Fig. 3) is set to 9]. As for base classifiers, SVM which utilizes a Radial Basis Function [47] (as kernel) and NN with the back-propagation training algorithm [14] was used.

Herein, the main focus of our comparative study is to investigate how well our multiple classifier system works well with *strong classifiers* – rather than feature extraction methods – in terms of improving classification performances. From this perspective, we compared the performance of the proposed multiple classifier against the single SVM and the single NN [21, 22]. It should be noted that SVM and NN are *representative strong classifiers* that are most commonly used in mammography CAD algorithms [5, 22]. In our experiment, the single SVM (or NN) classifier was constructed using the training set (obtained using 5 × 2-Fold cv), the same as that used to generate our proposed multiple classifier. Also note that for the sake of fair comparison, the parameters of the single classifiers were optimized, especially in order to prevent over-fitting. Specifically, we employed cross-validation approach [48] to decide an optimal number of hidden nodes of the single NN classifier. We started with a small number of hidden units (the starting number of hidden nodes was set to 2 in our approach) in the network. We then selected the optimal number of hidden nodes that lead to the best classification accuracy on the cross-validation samples. For the optimization of SVM classifier, we performed a so-called "grid-search" [44] on the associated parameters of kernel function and the regularization parameter using cross-validation to achieve the optimal generalization performance. After we found the best set of parameters, the whole training set was applied to generate the single SVM classifier with the best set of parameters. In our approach, tenfold cross-validation was used for the optimization of the single NN and SVM classifiers.

For comparative purposes, the best single SVM (or NN) classifier was used. For this, we first generated the K single classifiers (each trained with a particular feature representation in Table 1) and the best single classifier was then selected based on testing performances obtained using all the K single SVM (or NN) classifiers. Note that the best single classifier is trained with only one of the K feature representations. Additionally, we have compared the proposed multiple classifier with the single classifier approach using all the available features. For this purpose, we adopted "feature-

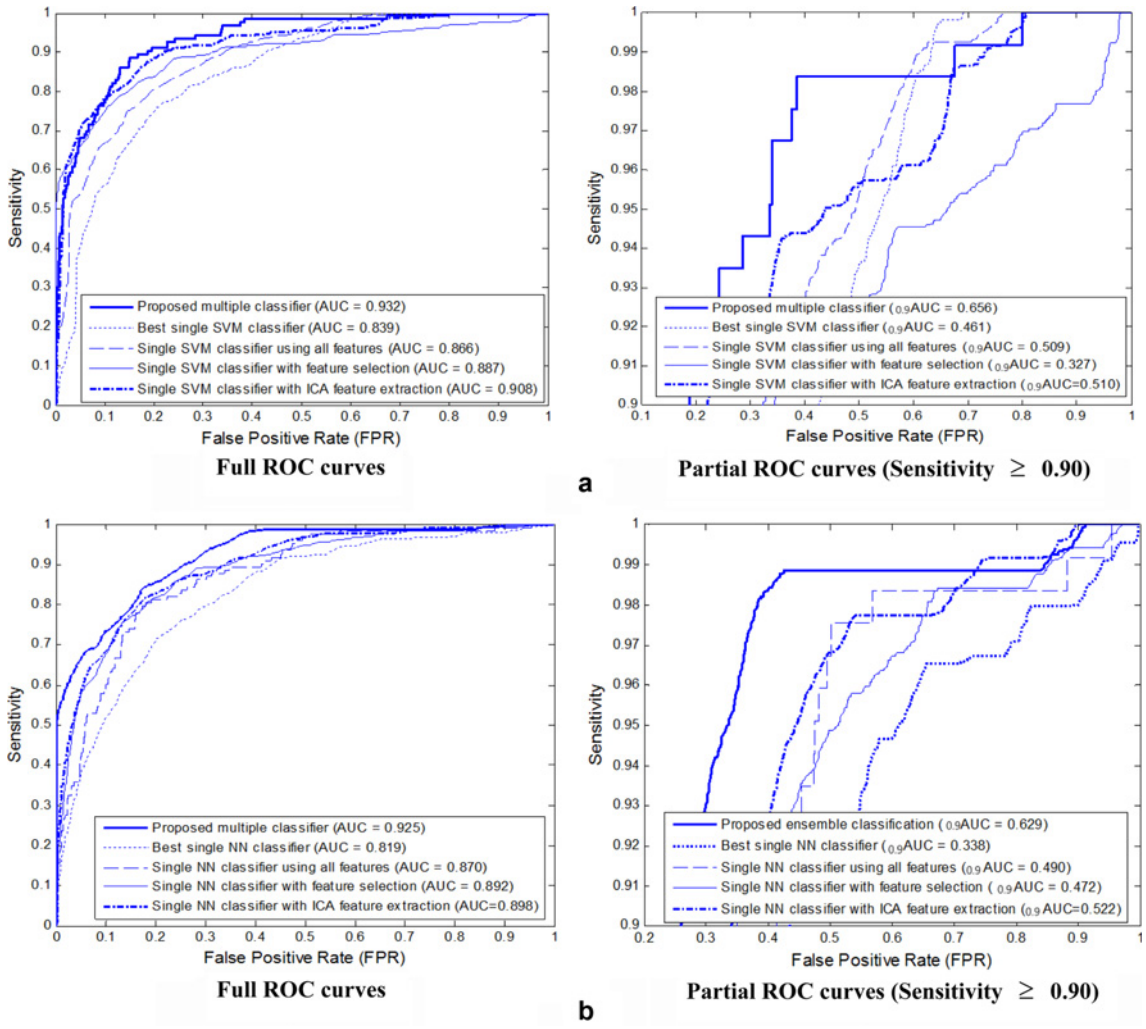

Fig. 4. Comparisons of ROC curves and AUC values. For the proposed multi-classifier, the parameter '*r*' was set to 0.4 and the number of boosting rounds '*T*' was set to 50. After finishing 50 boosting rounds, 49 SVM and 49 NN base classifiers were retained as the multiple classifier members. Note that AUC value of "1" represents a perfect classification. (a) SVM base classifiers. (b) NN base classifiers.

level" fusion strategy [50, 51] for combining complementary information resided in different feature spaces. Specifically, all the available feature representation vectors were concatenated at the feature-level in the standard column order, yielding a concatenated feature vector. These resulting concatenated feature vectors were then applied to construct a single classifier.

In addition, the single classifier with feature selection [8, 21] was compared with our proposed multiple classifier method. We have implemented the single classifiers using the stepwise feature selection (SFS) [13], extensively used for mammographic lesion classification. Note that all the available feature representations (in the form of concatenated feature vector) were applied to the SFS algorithm. At each step of the SFS procedure, one feature (or variable) is entered into or removed from the selected feature pool by analyzing its effect on a selection criterion [23]. Furthermore, the single classifier with Independent Component Analysis (ICA) feature extraction [55] was compared. To that end, following the feature fusion using ICA proposed in [54], each of the available feature representations described in Table 1 was individually applied to the so-called "unmixing matrix" [54, 55] (obtained from ICA algorithm) so as to extract *independent feature* (i.e., independent components). All independent features (extracted from all of the feature representations in Table 1) were then fused at the feature level, yielding "combined features". These combined features were used during the construction of single classifier models (such as SVM and NN) for classification purpose.

The results are given in Fig. 4. It can be seen that the proposed multiple classifier greatly outperforms the best single SVM and NN classifiers, in terms of both AUC and $_{0.9}$AUC (sensitivity ≥ 0.9). In particular, comparing to the best single classifier, the values of $_{0.9}$AUC significantly

increase with about 17.5% and 29.1%, in the order of SVM and NN, respectively. Also we can see that the results of the proposed method are much better than those obtained for the single classifier approach using all available the features for both SVM and NN. In addition, looking into results in Fig. 4, the proposed multiple classifier method achieves a better classification performance than single classifier in conjunction with ICA feature extraction, with around 14.6% and 10.7% improvement in $_{0.9}$AUC for SVM and NN, respectively. Moreover, as can be seen from Fig. 4, the proposed method outperforms the single classifier with feature selection in terms of AUC and $_{0.9}$AUC for both SVM and NN. In particular, using our multiple classifier solution, classification performance with regard to $_{0.9}$AUC can be substantially improved with around 32.9% and 15.7% for SVM and NN, respectively. This result indicates that our approach to find the best feature during the generation of base classifiers is more effective than the conventional approach (in mammography) based on a single classifier combined with feature selection.

The results shown in Fig. 4 clearly demonstrate that the proposed multiple classifier approach can be much more effective than the previous single classifier approaches, in terms of designing clinically relevant CAD systems that achieve high specificity performance at high sensitivity.

FP reduction performance on easy versus difficult mammograms

A high proportion of dense tissues may hamper cancer detection and recognition on mammograms by increasing the subtlety of a lesion [1, 4], causing high recall rates in mammography screening [1]. Thus, classifiers designed for FP reduction in CAD are likely to face a great challenge in differentiating mass and dense normal tissue, mainly due to poor image contrast [52]. In this sense, we investigated the feasibility of the proposed ensemble classifier against dense breast cases. To this end, we organized two smaller DBs by choosing subsets of mammograms from a total of 303 mammograms. The first subset contained the 171 mammograms with density rating of '1' or '2' (i.e., fatty or scattered fibroglandular densities) according to BIRADS categories; it will be referred to as the "easy" mammogram cases. The second subset consisted of 132 mammograms with density rating of '3' or '4' (i.e., heterogeneously dense or extremely dense). We called this DB the "difficult" mammogram cases.

To eliminate any potential variation (as much as possible) in measured performance of the classifier method, we randomly selected 2/3 mammogram cases from the full database (i.e., all 303 mammograms) and then divided these chosen cases into training and validation halves (each used for ensemble generation and ensemble selection) using

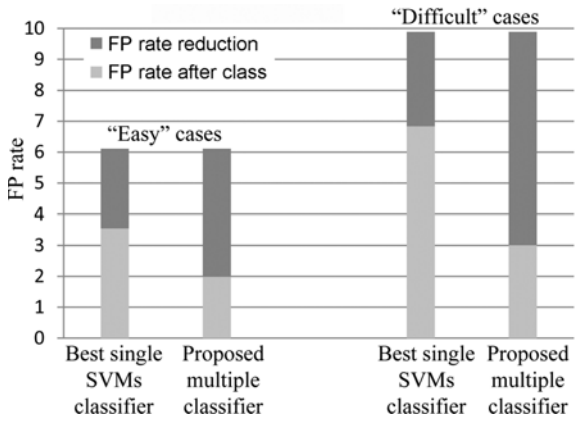

Fig. 5. Distribution of average FP rate and its FP rate reduction on "easy" and "difficult" case groups each with loss of 6.09% and 5.91% sensitivity rate induced by performing classification for the purpose of FP reduction.

5 × 2-Fold cv. The aforementioned process (including both random partition and 5 × 2-Fold cv) is repeated 20 times to guarantee stable experimental results. On the other hand, two testing sets (so-called "easy" and "difficult" mammogram cases) were always fixed for all cross-validation runs, which allows evaluating the proposed method with respect to the type of "easy" and "difficult" cases used for testing. Also, during the generation of ROIs via the computer segmentation, we chose operating thresholds which led to an average number of 6.11 FPs and 9.89 FPs per image at initial detection sensitivity of about 83.63% and 78.03%, for "easy" and "difficult" mammogram cases, respectively.

Fig. 5 shows the distribution of the FP rate (number of FPs per image) and its FP reduction on respective "easy" and "difficult" case groups with different number of FPs. As can be seen in Fig. 5, compared to the best single classifier, our multiple classifier approach allows for achieving better FP reduction performance. In particular, a much bigger FP reduction can be made on "difficult" cases; the overall average FP rate is reduced as much as 69.57% (from 9.89 to 3.01 per image) by using the proposed method at only the cost of 5.91% sensitivity loss (from 78.03% to 72.12%), while it can only be reduced for 30.84% (from 9.89 to 6.84 per image) by using the best single classifier with the same sensitivity loss. This observation is a very encouraging result, considering that since the dense breast usually has higher risk of cancer incidences, and higher FP detections, as well as higher false-negative detections [1, 4], alleviating the FP problem of dense mammogram cases may have a bigger impact on cancer screening with CAD.

Fig. 6 shows some mammography examples of FP reduction results. It can be observed that compared to using the best single classifier, the number of FPs in both heterogeneously and extremely dense cases (i.e., "difficult" cases) is greatly

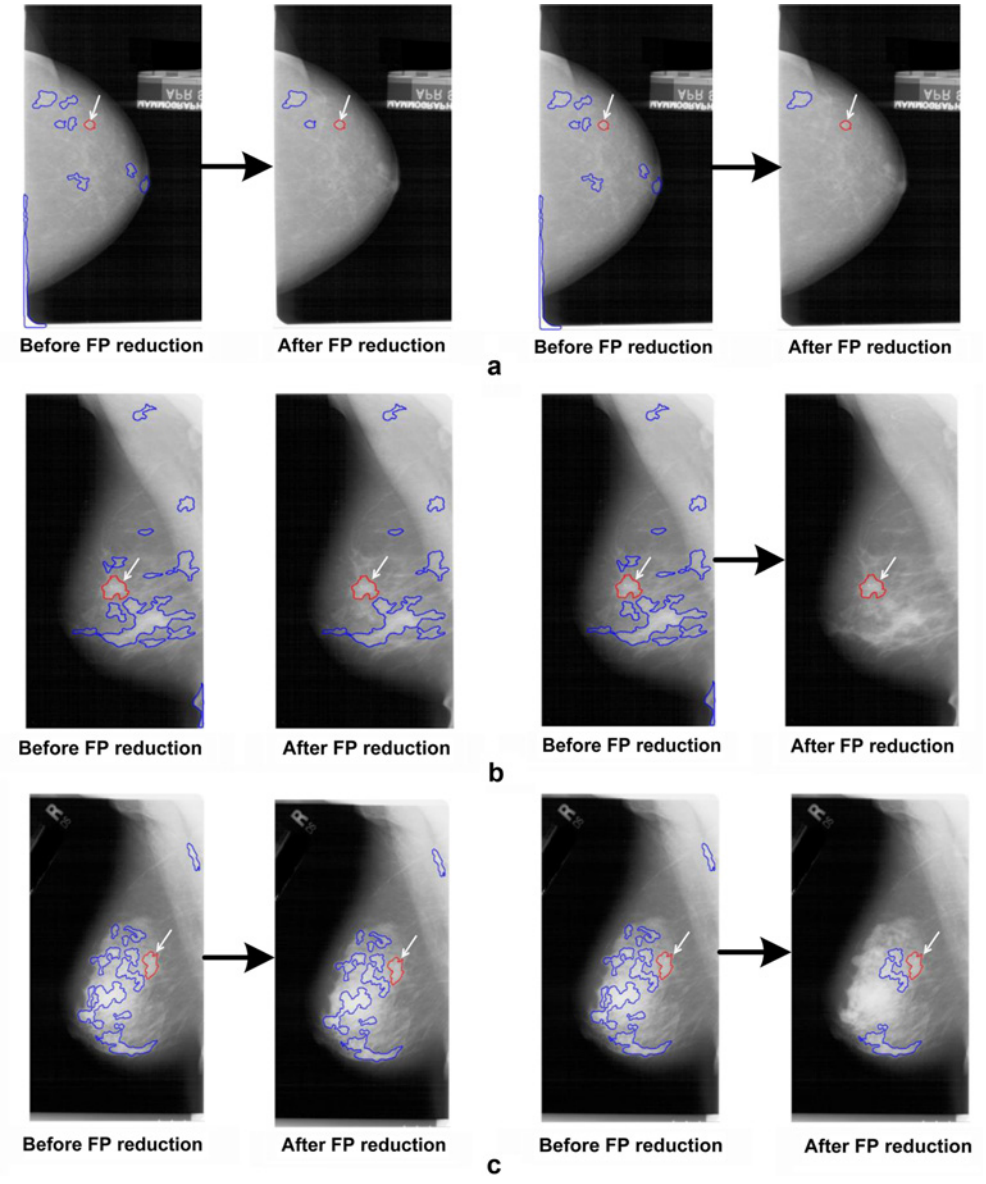

Fig. 6. Examples of mammograms for demonstrating the effectiveness of our proposed multiple classifier approach in terms of reducing the number of FPs. Note that the examples on the left side are shown for the case of using the best single SVM classifier, while the examples on the right side are for the proposed multiple classifier approach. Also note that true masses are indicated by white colored arrows in all mammograms, while the others are FP regions. (a) Fatty. (b) Heterogeneously dense. (c) Extremely dense.

reduced by using the proposed method while true masses are well preserved (kept).

Based on all results presented in this section, the effectiveness of the proposed approach is much more significant for challenging mammogram cases. The main reason for much-improved performance can be explained as follows. The proposed multiple classifier generation can be viewed as a process of finding the best feature representation to train a *local learner* that specializes in correctly classifying mass instances (patterns) resided in a *particular local decision space*. For this reason, the proposed method allows *hard-to-classify mass instances* coming from dense mammogram cases to *be handled by some base classifier members that specialize for correct classification on those difficult instances*. This means that our multiple classifier approach could provide *more expressive power* in correctly determining the complex and arbitrary decision boundaries (induced by hard-to-classify mass instances), which would be difficult to reach with only a single classifier.

Effect of parameters

Note that there are two important parameters affecting classification performance of the proposed multiple classifier scheme: 1) data resampling parameter "*r*" described in Figs. 3

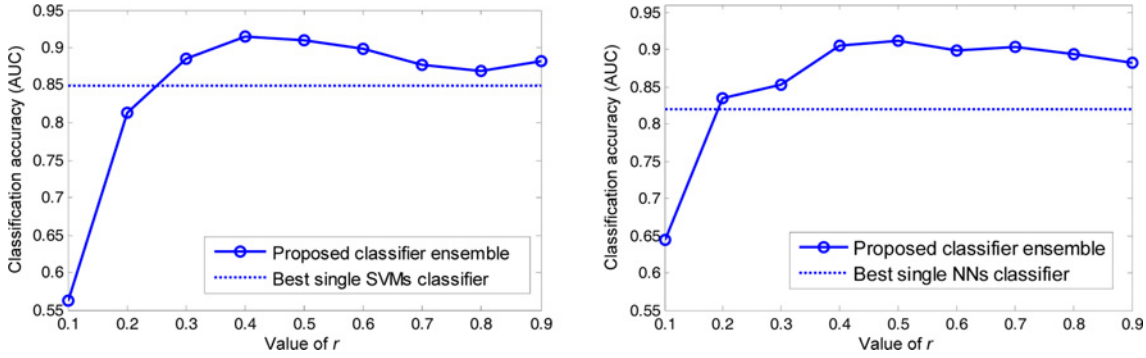

Fig. 7. Classification performances of the proposed multiple classifier as a function of parameter '*r*'. The graphs on the left side correspond to SVM base classifier, while those on the right side to NN base classifier.

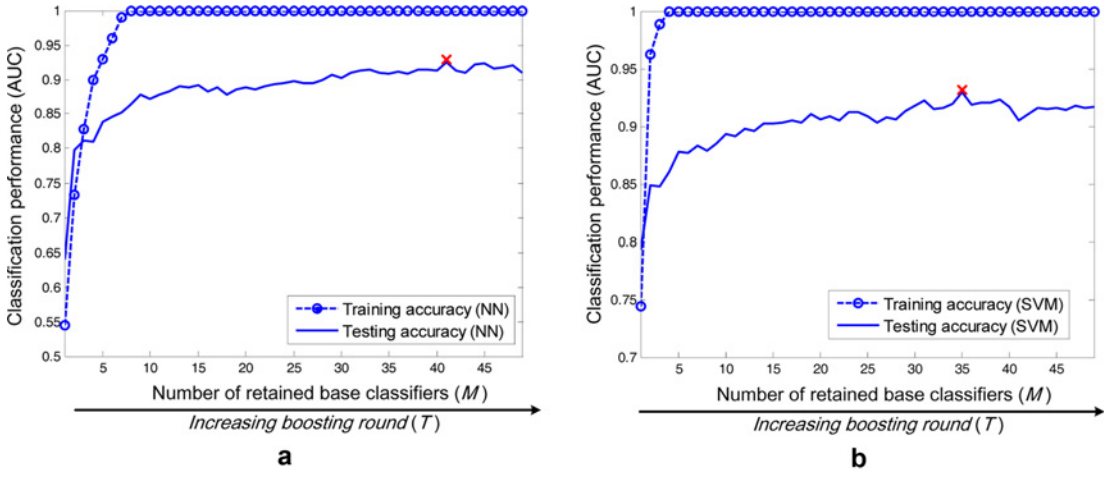

Fig. 8. Training and testing classification performance as a function of *M*. In each plot, "cross mark" represents the maximum testing performance, as obtained for a particular value of *M* less than the total number of retained base classifiers. (a) SVM. (b) NN.

and 2) the number of boosting rounds denoted as "*T*" (or the number of retained base classifiers denoted as *M*). In this section, experimental analysis has been performed to examine how the classification performance is influenced by these two factors.

Fig. 7 shows the variation in testing classification performance with respect to changes in value *r*. Note that in Fig. 7, the performance of the best single SVM (or NN) classifier is referred to as baseline performance. Also, in Fig. 7, the weakest and strongest base classifiers are assumed to be produced when $r = 0.1$ and $r = 1$, respectively. From Fig. 7, three common observations can be made as follows: (1) classification performance is optimal (or nearly optimal) for values of *r* in the range of [0.4, 0.6] for both SVM and NN; this result justifies the advantage of using *r* for adjusting weakness of base classifiers; (2) classification performance can be significantly deteriorated when base classifiers are considered to be too weak (i.e., when $r = 0.1$ or $r = 0.2$); (3) the results also validate the robustness (tolerance) to a certain extent, against variations in value of *r*, since the classification performance for proposed multiple classifier is always better than the classification performance for baseline method unless the value of *r* falls below 0.3.

Figs. 8a and 8b show the training and testing classification performance as a function of *M* for SVM and NN, respectively. Training classification performance for both SVM and NN continues to increase as *M* becomes large, and quickly levels off. Considering testing performance, it seems to generally improve as *M* increase up to a particular number and repeat increasing and decreasing, and finally converges to nearly same constant value. In Fig. 8, testing performance for SVM and NN is maximized at corresponding $M = 35$ and $M = 41$, each of which is smaller than the total number of retained base classifiers. This result indicates that in practical application, just sequentially adding the classifier to current multiple classifier system may not always guarantee improvement in testing performance [53]. To resolve this issue, classifier selection solutions [53] could be used in our multiple classifier framework to effectively choose base classifiers for further improving testing performance. This work should be

considered out of scope for the current paper and, therefore, has been left as future research.

CONCLUSIONS

We developed a new multiple classifier framework with an application to improve classification of breast masses on mammograms in Computer-aided detection (CAD) systems. Differing from the existing multiple classifier methods, we proposed the combined use of different feature representations (of the same instance) and data resampling to generate more diverse and accurate base classifiers. Another distinct characteristic is to select the best base classifier (and/or best feature representation) at each boosting round − yielding the most accurate results on a weighted training set. In addition, to overcome a weak learner limitation of boosting-like multi-classifier learning, we developed a simple but effective mechanism that regulates the degree of weakness of the base classifiers by adjusting the size of a resampled set. This allows our proposed multiple classifier framework to be generalized to work with strong classifiers extensively used in mammography CAD systems.

In this paper, the generation of base classifiers together with different kinds of features is restricted to using classification accuracy on a weighted training set. This may lead to some of the selected base classifiers that do not make a contribution (to a certain extent) to the multiple classifier system in terms of maximizing *generalized* classification performance. This may be mainly attributed to the fact that the classification outputs of some base classifiers (chosen) are highly correlated with those obtained from other base classifiers. To resolve this problem, for future work, we will extend our work by incorporating *base classifier selection scheme* into our proposed multiple classifier system, aiming to emphasize interaction and cooperation among individual base classifiers. The goal of base classifier selection is to find the best subset from a whole set of generated base classifiers, not just to maximize classification accuracy, but also to *maximize diversity* [14, 53] (i.e., *minimize co-linearity*) between individual base classifiers built using different kinds of features. To this end, we will develop a novel selection criterion, which is designed for making optimal balance between the classification accuracy and diversity [14, 53] during the generation of base classifiers. This is expected to yield better *generalized* classification performance.

ACKNOWLEDMENTS

This work was supported by the Jungwon University Research Grant(2015-017).

CONFLICT OF INTEREST STATMENTS

Choi JY declares that he has no conflict of interest in relation to the work in this article.

REFERENCES

[1] Kopans DB. Breast imaging. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2007.
[2] Kwon HJ, Shin BH, Gopaul D, Fienberg S. Strain ratio vs. modulus ratio for the diagnosis of breast cancer using elastography. Biomed Eng Lett. 2014; 4(3):292-300.
[3] Mert A, Kilic N, Akan A. An improved hybrid feature reduction for increased breast cancer diagnostic performance. Biomed Eng Lett. 2014; 4(3):285-91.
[4] Suri JS, Rangayyan RM. Recent advances in breast imaging, mammography, and computer-aided diagnosis of breast cancer. Washington: SPIE PRESS; 2006.
[5] Sampat MP. Computer-aided detection and diagnosis in mammography. In: Bovik AC. Handbook of image and video processing. 2nd ed. New York: Academic; 2005. pp. 1195-217.
[6] Tang J, Rangayyan RM, Xu J, El Naqa I, Yang Y. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Trans Inf Technol Biomed. 2009; 13(2):236-51.
[7] Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. Med Phys. 1999; 26(12):2654-88.
[8] Cheng HD, Shi XJ, Min R, Hu LM, Cai XP, Du HN. Approaches for automated detection and classification of masses in mammograms. Pattern Recognit. 2006; 39(4):646-68.
[9] Berber T, Alpkocak A, Balci P, Dicle O. Breast mass contour segmentation algorithm in digital mammograms. Comput Methods Programs Biomed. 2013; 110(2):150-9.
[10] Bilska-Wolak AO, Floyd CE Jr. Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer. Phys Med Biol. 2004; 49(18):4219-37.
[11] Mudigonda NR, Rangayyan RM, Desautels JEL. Gradient and texture analysis for the classification of mammographic masses. IEEE Trans Med Imaging. 2000; 19(10):1032-43.
[12] Jesneck JL, Nolte LW, Baker JA, Floyd CE, Lo JY. Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis. Med Phys. 2006; 33(8):2945-54.
[13] Wei D, Chan HP, Petrick N, Sahiner B, Helvie MA, Adler DD, Goodsitt MM. False-positive reduction technique for detection of masses on digital mammograms: global and local multiresolution texture analysis. Med Phys. 1997; 24(6):903-14.
[14] Kuncheva LI. Combining pattern classifiers: methods and algorithms. New York: Wiley; 2004.
[15] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997; 55(1):119-39.
[16] Constantinidis AS, Fairhurst MC, Rahman AFR. A new multi-expert decision combination algorithms and its application to the detection of circumscribed masses in digital mammograms. Pattern Recognit. 2001; 34(8):1527-37.
[17] Yoon SJ, Kim SJ. AdaBoost-based multiple SVM-RFE for classification of mammograms in DDSM. IEEE Int Conf Bioinf Biomed Workshops. 2008; 75-82.
[18] Breiman L. Random forests. Mach Learning. 2001; 45(1):5-32.
[19] Lu J, Plataniotis KN, Venetsanopoulos AN, Li SZ. Ensembled-

based discriminant learning with boosting for face recognition. IEEE Trans Neural Netw. 2006; 17(1):166-78.
[20] Murua A. Upper bounds for error rates of linear combinations of classifiers. IEEE Trans Pattern Anal Mach Intell. 2002; 24(5):591-602.
[21] Way TW, Sahiner B, Hadjiiski LM, Chan HP. Effect of finite sample size on feature selection and classification: a simulation study. Med Phys. 2010; 37(2):907-20.
[22] Wei L, Yang Y, Nishikawa RM, Jiang Y. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. IEEE Trans Med Imaging. 2005; 24(3):371-80.
[23] Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, Goodsitt MM. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging. 1996; 15(5):598-609.
[24] Wei D, Chan HP, Helvie MA, Sahiner B, Petrick N, Adler DD, Goodsitt MM. Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis. Med Phys. 1995; 22(9):1501-13.
[25] Mudigonda NR, Rangayyan RM, Desautels JEL. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. IEEE Trans Med Imaging. 2001; 20(12):1215-27.
[26] Kupinski MA, Giger ML. Investigation of regularized neural networks for the computerized detection of mass lesions in digital mammograms. IEEE Int Conf Eng Med Biol Soc (EMBS). 1997; 3:1336-9.
[27] Santo MD, Molinara M, Tortorella F, Vento M. Automatic classification of clustered microcalcifications by a multiple expert system. Pattern Recognit. 2003; 36(7):1467-77.
[28] Oliver A, Torrent A, Llado X, Tortajada M, Tortajada L, Sentis M, Freixenet J, Zwiggelaar R. Automatic microcalcification and cluster detection for digital and digitised mammograms. Knowledge-Based Syst. 2012; 28:68-75.
[29] Arodź T, Kurdziel M, Sevre EO, Yuen DA. Pattern recognition techniques for automatic detection of suspicious-looking anomalies in mammograms. Comput Methods Programs Biomed. 2005; 79(2):135-49.
[30] Fung G, Krishnapuram B, Merlet N, Ratner E, Bamberger P, Stoeckel J, Rao RB. Addressing image variability while learning classifiers for detecting clusters of micro-calcifications. Int Conf Digit Mammogr. 2006; 4046:84-91.
[31] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Annal Stat. 2000; 28(2):337-407.
[32] Dominguez AR, Nandi AK. Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection. Comput Med Imaging Graph. 2008; 32(4):304-15.
[33] Hong BW, Sohn BS. Segmentation of regions of interest in mammograms in a topographic approach. IEEE Trans Inf Technol Biomed. 2010; 14(1):129-39.
[34] Choi JY, Ro YM. Multiresolution local binary pattern texture analysis combined with variable selection for application to false positive reduction in computer-aided detection of breast masses on mammograms. Phys Med Biol. 2012; 57(21):7029-52.
[35] Sahiner B, Chan HP, Petrick N, Helvie MA, Hadjiiski LM. Improvement of mammographic mass characterization using speculation measures and morphological features. Med Phys. 2001; 28(7):1455-65.
[36] Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. Med Phys. 1998; 25(4):516-26.
[37] Shavlik JW, Mooney RJ, Towell GG. Symbolic and neural learning algorithms: an experimental comparison. Mach Learn. 1991; 6(2):111-43.
[38] Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell. 1991; 13(3):252-64.
[39] Alimoglu F, Alpaydin E. Combining multiple representations and classifiers for pen-based handwritten digit recognition. Turk J Electr Eng. 2001; 9(1):1-12.
[40] Rokach L. Ensemble-based classifiers. Artif Intell Rev. 2010; 33(1):1-39.
[41] Ranawana R, Palade V. Multi-classifier systems: Review and a roadmap for developers. Int J Hybrid Intell Syst. 2006; 3(1):35-61.
[42] Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer PJ. The digital database for screening mammography. Int Conf Digit Mammography. 2000; 212-8.
[43] Catarious DM Jr, Baydush AH, Floyd CE Jr. Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system. Med Phys. 2004; 31(6):1512-20.
[44] Eltonsy NH, Tourassi GD, Elmaghraby AS. A concentric morphology model for the detection of masses in mammography. IEEE Trans Med Imaging. 2007; 26(6):880-9.
[45] Zhou XH, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine. New York: Wiley-Interscience; 2002.
[46] Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis. Phys Med Biol. 1998; 43(10):2853-71.
[47] Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
[48] Setiono R. Feedforward neural network construction using cross validation. Neural Comput. 2001; 13(12):2865-77.
[49] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011; 2(3):1-27.
[50] Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems. Pattern Recognit. 2005; 38(12):2270-85.
[51] Mangai UG, Samanta S, Das S, Chowdhury PR. A survey of decision fusion and feature fusion strategies for pattern classification. IETE Tech Rev. 2010; 27(4):293-307.
[52] Nishikawa RM. Current status and future directions of computer-aided diagnosis in mammography. Comput Med Imaging Graph. 2007; 31(4-5):224-35.
[53] Ruta D, Gabrys B. Classifier selection for majority voting. Inf Fusion. 2005; 6(1):63-81.
[54] Wei X, Zhou C, Zhang Q. ICA-based feature fusoin for face recognition. Int J Innov Comput Inf Control. 2010; 6(10):4651-61.
[55] Naik GR, Kumar DK. An overview of independent component analysis and its applications. Inform. 2011; 35(1):63-81.