Published in final edited form as:

Phys Med Biol. 2014 August 7; 59(15): 4357–4373. doi:10.1088/0031-9155/59/15/4357.

Reduction of false-positive recalls using a computerized mammographic image feature analysis scheme

Maxine Tan^{1,3}, Jiantao Pu², and Bin Zheng^{1,2}

¹ School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK 73019

² Department of Radiology, University of Pittsburgh, Pittsburgh, PA 15213

Abstract

The high false-positive recall rate is one of the major dilemmas that significantly reduce the efficacy of screening mammography, which harms a large fraction of women and increases healthcare cost. This study aims to investigate the feasibility of helping reduce false-positive recalls by developing a new computer-aided diagnosis (CAD) scheme based on the analysis of global mammographic texture and density features computed from four-view images. Our database includes full-field digital mammography (FFDM) images acquired from 1052 recalled women (669 positive for cancer and 383 benign). Each case has four images: two craniocaudal (CC) and two mediolateral oblique (MLO) views. Our CAD scheme first computed *global* texture features related to the mammographic density distribution on the segmented breast regions of four images. Second, the computed features were provided to two artificial neural network (ANN) classifiers that were separately trained and tested in a ten-fold cross-validation scheme on CC and MLO view images, respectively. Finally, two ANN classification scores were combined using a new adaptive scoring fusion method that automatically determined the optimal weights to assign to both views. CAD performance was tested using the area under a receiver operating characteristic curve (AUC). The $AUC=0.793\pm0.026$ was obtained for this four-view CAD scheme, which was significantly higher at the 5% significance level than the AUCs achieved when using only CC ($p = 0.025$) or MLO ($p = 0.0004$) view images, respectively. This study demonstrates that a quantitative assessment of global mammographic image texture and density features could provide useful and/or supplementary information to classify between malignant and benign cases among the recalled cases, which may eventually help reduce the false-positive recall rate in screening mammography.

Keywords

Breast cancer screening; Full-field digital mammography (FFDM); Computer-aided diagnosis (CAD); Quantitative mammographic density measures

³ Corresponding author, Maxine.Y.Tan-1@ou.edu.

1. Introduction

Breast cancer is the most prevalent cancer and the second leading cause of cancer death in women, according to the latest statistics from the American Cancer Society (American Cancer Society, 2014), which estimates that about 1 in 8 (12%) of women in the United States will develop breast cancer during their lifetime. Since breast cancer is a very heterogeneous disease and most of the cancers are detected among women without currently known risk factors, a uniform and population-based mammography screening protocol is recommended for all women who are older than the qualifying age. Scientific evidence has shown that early detection of breast cancer combined with improved treatment strategies has incrementally and significantly reduced patients' mortality and morbidity rates over the past four decades (Tabar *et al.*, 2001; Smith *et al.*, 2011). However, the task of visually interpreting mammograms and detecting cancer is difficult due to the high variability of breast abnormalities and overlapping dense fibroglandular tissue (FGT). This issue is further compounded by the low breast cancer prevalence rate (~3-5 cancers detected per 1000 non-baseline screening examinations) in a screening environment (Sickles *et al.*, 2002).

Such difficulty generates very high recall rates (e.g., 6% to 12%) in which the majority (90% to 95%) are false-positive recalls. For example, one study (Hubbard *et al.*, 2011) reported that after 10 years of annual screening, more than half of women will receive at least one false positive (FP) recall, and 7% to 9% will receive a FP biopsy. The negative consequences of FP recalls are many and varied; they include long-term psychosocial consequences (of up to three years after a FP finding) (Brodersen and Siersma, 2013; Consedine, 2013), cumulative radiation exposure (Linnet *et al.*, 2012; Yaffe and Mainprize, 2011) as well as the effects on limited healthcare resources and associated high costs (Buist *et al.*, 2011; Haneuse *et al.*, 2012). As a result, the efficacy of screening mammography remains a very controversial issue to date. In order to improve efficacy of screening mammography, a number of new research ideas have been proposed aiming to establish an optimal personalized breast cancer screening paradigm (Brawley, 2012).

In the last two decades, various computer-aided detection (CAD) schemes of mammograms, which function as a "second reader" to radiologists, have also been developed and tested (Wei *et al.*, 2011b; Zheng *et al.*, 2006; Zheng *et al.*, 2012; Eltonsy *et al.*, 2007; Varela *et al.*, 2007; Kom *et al.*, 2007; Catarious *et al.*, 2006; Velikova *et al.*, 2009; Qian *et al.*, 2007). The majority of these schemes including the commercialized CAD schemes are single image based schemes that focus on detecting more subtle cancers with high false-positive rates. A number of studies have indicated that using current CAD schemes as the "second reader" reduced radiologists' specificity in reading mammograms (Gur *et al.*, 2004b; Fenton *et al.*, 2011).

Since in each screening mammography examination, two view images are taken of each breast, namely a craniocaudal (CC) and a mediolateral oblique (MLO) view, reading both images is the current standard clinical practice that enables to detect more cancers and have a lower FP rate than reading single view images alone (Ikeda and Sickles, 1988; Sickles *et al.*, 1986; Wald *et al.*, 1995). Thus, in order to improve CAD performance, many investigators have also developed and tested different methods of combining multiple view

images into the classification process to improve CAD performance in detection and classification of mammographic abnormalities. These efforts include more accurately locating the breast masses and/or FP detections on two ipsilateral (CC and MLO) views (Zheng *et al.*, 2006; Zheng *et al.*, 2009; van Engeland and Karssemeijer, 2007; Wei *et al.*, 2009; Yuan *et al.*, 2008) and more effectively combining the information detected from different image views (Liu *et al.*, 2005; Wang *et al.*, 2011; Wei *et al.*, 2011b). Despite these research efforts, multi-view based CAD schemes have not been accepted and used in the clinical practice to date.

In this study, we proposed a new four-view full-field digital mammography (FFDM) image based scheme with a different prospect. Specifically, unlike other existing CAD schemes that classify the detected individual suspicious mammographic abnormality being positive or negative, our CAD scheme does not detect or classify individual suspicious breast lesions. The scheme only extracts and fuses the global mammographic texture and density features computed on *all four images* of the MLO and CC views. This new approach is developed based on the factor that mammographic tissue density distribution could be more consistently detected using CAD-type schemes (Glide-Hurst *et al.*, 2007; Wang *et al.*, 2003; Chang *et al.*, 2002; Zhou *et al.*, 2001; Li *et al.*, 2005; Byng *et al.*, 1997; Wei *et al.*, 2011a) and the computed mammographic density features are useful to predict the risk of developing or having breast cancer (Glide-Hurst *et al.*, 2007; Wang *et al.*, 2003; Chang *et al.*, 2002; Zhou *et al.*, 2001; Li *et al.*, 2005; Byng *et al.*, 1997; Wei *et al.*, 2011a). In this study, we further hypothesized that using the global mammographic density and tissue distribution patterns among the four-view FFDM images could also help improve performance of classification between the malignant and benign recalled cases. To test this hypothesis, our new CAD scheme computes global image texture based features from the whole breast region and the dense breast regions to objectively assess mammographic tissue density distribution and difference. The scheme also includes an adaptive scoring fusion method to combine information from the CC and MLO views by automatically optimizing the allocation of weights/feature importance attributed to both views to yield a final classification score. We then tested the feasibility and performance of using this classification score to assess the likelihood of the recalled cases being malignant.

2. Materials and methods

2.1. The FFDM Image Dataset

From an existing large and diverse FFDM image database as reported in our previous studies to develop CAD schemes of detecting breast masses and predicting cancer risk (Tan *et al.*, 2013; Zheng *et al.*, 2012), we selected all available recalled and biopsied cases to assemble a testing dataset including 1052 fully anonymized FFDM cases for this study. Specifically, the dataset includes (1) 625 verified cancer cases depicting masses that were detected during FFDM screening and confirmed by diagnostic work-up; (2) 44 “interval” cancer cases depicting masses that were detected in the interval between two screening examinations; and (3) 383 recalled cases due to suspicious findings on FFDM images and were ultimately determined to be benign by biopsy. Each case has four FFDM images representing the CC and MLO views of the left and right breasts. In summary, altogether

4208 FFDM images were acquired from 1052 women who underwent FFDM screening examinations. Of these, 669 were positive for cancer and 383 were benign.

Figure 1 displays an example, namely CC and MLO view FFDM images of the left and right breasts of a malignant case from our dataset. In addition to the FFDM images, the subjective breast density rating using BI-RADS was collected for each case. Figure 2 displays the distribution of all the cases in our dataset into four categories of mammographic density based on the BI-RADS ratings of the radiologists. The data analysis result based on an unpaired *t*-test shows that there is no statistically significant difference of the mammographic density distribution (BI-RADS) in the two case subgroups of recalled benign and positive for cancer ($p = 0.53$).

2.2. Automated breast segmentation for the CC and MLO view images

In the first step of our CAD development, a breast region segmentation scheme reported in our previous studies (Zheng *et al.*, 2012; Gur *et al.*, 2004a) was modified and applied in this study. A flowchart of the segmentation scheme is displayed in figure 3. In brief, the scheme first detects the image orientation of the left or right breast, and automatically flips the image of the right breast. In this way, the sides of the chest walls of the breasts always align on the left edge of the images. Following that, the gray level histogram of the image is plotted, and an iterative searching method is applied to detect the smoothest curvature between the breast tissue and the background (air). The pixels depicting the background are discarded from the segmented breast region. Also, we removed the skin region by performing a morphological erosion operation on the segmented breast so that the features are computed only on the whole and dense breast regions excluding the skin region.

For the MLO view images, an additional step has to be performed to detect the chest wall (or pectoral muscle). The scheme first horizontally scans the breast image from left to right of a certain number of predefined rows to search for the location of the pixel with the maximum gradient in each row. Then, the pectoral muscle is detected by employing linear regression to fit all the identified pixels with the maximum gradient on a straight line/fit. Finally, all the pixels within the pectoral muscle are discarded from the segmented breast area as these pixels do not add useful/beneficial information for the classification task (Law *et al.*, 2014).

2.3. Computation of global texture and mammographic density based features

In the second step of our CAD development, we computed 91 mammographic image features depicted on the whole segmented breast region and 50% of breast region with denser tissues as defined by Chang et al (Chang *et al.*, 2002) and one ratio of fibroglandular tissue (FGT) region to the whole breast region. These features cover a wide range of different image characteristics as summarized in table 1. In brief, the first feature group comprises of pixel value (intensity)-based statistical features, namely, average, standard deviation, skewness, and kurtosis values, which were computed using the same methods as previously reported (Wang *et al.*, 2011). The second group includes two fractal dimension based features as described in Ref. (Chang *et al.*, 2002).

The third group consists of 11 gray level run length based texture features (Tang, 1998), computed using the toolbox provided by Wei (Wei, 2007). These features consist of: short run emphasis, long run emphasis, low gray level run emphasis, high gray level run emphasis, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis, long run high gray level emphasis, gray level non-uniformity, run length non-uniformity, and run percentage. To compute the run length based features, we first reduced the gray level range of the images at a subsampling ratio of 16 from 4096 to 256 gray levels. Then, we computed four run length matrices along the following directions: 0, 45, 90, and 135 degrees. We computed the average and maximum values of the four run length features along the four different directions to obtain the final feature values.

The fourth group consists of the well-known gray level co-occurrence matrix (GLCM) based features (Haralick *et al.*, 1973), namely contrast, correlation, energy, and homogeneity computed at distance, $d = 1$. We only computed the GLCM features at $d = 1$ as it was reported in Refs. (Varela *et al.*, 2006; Mudigonda *et al.*, 2000) since at $d > 1$, the GLCM based features are strongly correlated. We also computed the GLCM features in four directions; in doing so, we hope to detect the radiating lines (spicules) that frequently characterize malignant masses within the breast regions. The gray level range of the images was reduced from 4096 to 256 levels in calculating the GLCM matrix as performed in Refs. (Mudigonda *et al.*, 2000; Varela *et al.*, 2006). As we computed the average and maximum values of the four features in the four directions, we obtained altogether $4 \times 2 = 8$ features per image. We computed the average value of the four GLCM features to obtain rotation invariance (Mudigonda *et al.*, 2001) as the spicules of the masses were likely to be oriented in many different directions. However, we also calculated the maximum value of the features as the possibility exists that the spicules may have a dominant orientation in any given direction.

The fifth group includes the x -axis and y -axis histogram (cumulative projection) based features proposed by Tzikopoulos *et al.* (Tzikopoulos *et al.*, 2011). In brief, we computed the mean, standard deviation, skewness, kurtosis, and median of the x -axis and y -axis cumulative projection histograms of each image. First, to find the x -axis cumulative projection histogram, we computed the sum of “on” pixels in every row of the image. We then obtained the cumulative projection of the y -axis histogram by computing the sum of “on” pixels in every column of the image. Then, the first-order statistics (mean, standard deviation, skewness, kurtosis, and median) were computed for both histograms. The sixth group includes a regional descriptor, namely the size (or area) of the segmented breast region in all four images. The last feature (in table 1) is used to estimate the mammographic density that mimics the BI-RADS standard, which rates mammographic density into four categories, as follows: (1) almost entirely fatty ($\text{FGT} \leq 25\%$); (2) scattered fibroglandular density ($25\% < \text{FGT} \leq 50\%$); (3) heterogeneously dense ($50\% < \text{FGT} \leq 75\%$); and (4) extremely dense ($\text{FGT} > 75\%$) (American College of Radiology, 2003). In this study, we computed a different and simple index to estimate the ratio between FGT region and whole breast region by modifying a well-established method used in a semi-automated software (Cumulus) (Byng *et al.*, 1994), which has gained wide acceptance as “a gold standard” for mammographic density assessment in many epidemiology based studies (Bertrand *et al.*,

2013; Li *et al.*, 2012; Noh *et al.*, 2006; Maskarinec *et al.*, 2009; Heine *et al.*, 2012; Manduca *et al.*, 2009). To use this software, two global gray level thresholds are identified by the operator to segment relevant regions of the breast required to estimate percentage density (*PD*). The first gray level threshold defines the global threshold to outline the edge of the breast. The second threshold is the global threshold to segment the dense region of the image subject to the discretion of the operator. Finally, the software computes *PD* as the ratio between the areas (i.e., total number of pixels) of the dense region to the segmented breast region. Our method uses the same concept as the Cumulus software, but it is a fully-automated and objective based measure to estimate mammographic density. We defined a new *PD* measure computed as the ratio of the areas of the region within the segmented breast with intensity values exceeding the average (mean) intensity value of the segmented breast to the whole segmented breast region as follows:

$$PD = \frac{N(DeT)}{N(BrT)} \quad (1)$$

where $N(R)$ is the number of pixels in region R , BrT is the segmented breast (tissue) region, and DeT is the dense tissue region estimated as the region within the segmented breast with intensity values exceeding the average intensity value of the segmented breast region.

In summary, our CAD scheme computed a total of 92 features (as shown in table 1) on the left and right breast images of the CC and MLO views, respectively. Although the absolute values of different features are different, the values of all features were first linearly normalized to the range from 0 to 1. These normalized global image features were then used to train the multi-feature based classifiers in the next step of the CAD scheme.

2.4. Optimization of a new four-view based classifier including a scoring fusion ANN

In this study, we tested a two-stage classification scheme, which was implemented using the Matlab® Neural Network Toolbox (as shown in Figure 4). The first stage has two artificial neural networks (ANN). One only uses image features computed from CC views of the left and right breast, while another uses the features computed from MLO view images. The second stage includes another ANN that creates a new scoring fusion approach to combine the classification scores generated by two CC and MLO view image feature based ANNs. In this way, the weights applied to two original ANN scores computed based on the CC and MLO view images will be adaptively adjusted.

Each ANN was trained and tested using a 10-fold cross-validation method in which 669 malignant cases and 383 benign cases in our database were randomly divided into 10 exclusive partitions/subsets. At each validation (training and testing) cycle, nine subsets were used to train the ANN classifier, and the trained classifier was then applied to the remaining testing subset. When using a linear activation function at the output layer of the ANN, the classification scores of testing cases were adjusted in the range from 0 to 1.2, whereby a higher score indicates a higher probability of the case in question being malignant. This process was iteratively repeated 10 times using the 10 different combinations of training and testing subsets. In this way, each of the 1052 malignant and benign cases was tested once with a corresponding classifier-generated classification score.

The ANN classifiers were trained with a gradient descent with momentum and adaptive learning rate back-propagation algorithm (Rumelhart *et al.*, 1986). To find the optimal number of hidden nodes to use in the hidden layers of the ANNs, we analyzed the networks' performance for different numbers of hidden nodes in the hidden layer, which was always initialized with random weights. For the CC and MLO based ANNs in the first stage, the number of hidden nodes was varied between 2-40 nodes; for the sequential scoring fusion ANN in the second stage, this range was 2-10. We employed a higher range for the CC and MLO ANNs as they had many more input features (92) compared with the scoring fusion ANN (2). We trained 150 different ANNs on the training subsets, and found the network that maximized the AUC on the training subsets. We also used the hyperbolic tangent activation function at the hidden nodes, and a linear activation function at the output node. Other parameters related to the ANN training are the number of training iterations (500), the training momentum (0.9) and the learning rate (0.01). In particular, a limited number of iterations, as well as a high ratio of the training momentum to the learning rate was implemented to reduce "over-fitting" and maintain classifier robustness.

We applied a receiver operating characteristic (ROC) type data analytic method to assess and compare the classification performance levels yielded using the CC and MLO view feature based ANNs, as well as the optimized scoring-fusion based ANN. To plot the ROC curves, we computed the average and standard deviation intervals over the ten folds of the cross-validation scheme. In the data analysis, we computed the areas under ROC curves (AUC values) of all three ANN classifiers, and performed statistical significance tests at the 5% significance level. We also tabulated the confusion matrix of the optimized scoring-fusion based ANN. In addition, we analyzed the performance of our classification scheme on different cancer subgroups within the testing image dataset, namely, verified cancer cases, interval cancer cases and recalled benign cases. We also stratified the images according to their density BIRADS ratings, and analyzed our scheme's classification sensitivity levels at different density ratings at the 90% specificity level.

3. Results

The ROC curves of the CC, MLO and scoring-fusion based ANNs are displayed in figure 5, while the computed average and standard deviation results of the AUC values of each ANN computed over the ten folds are summarized in table 2. The results show that the scoring fusion ANN yielded the highest AUC values as compared to the AUC values yielded separately by two CC and MLO view image feature based ANNs. The statistical significance at the 5% significance level using DeLong's test (Robin *et al.*, 2011; DeLong *et al.*, 1988) was demonstrated when comparing the AUC values of the scoring fusion ANN to both the ANN trained using CC view image features ($p = 0.025$) and the ANN trained by features computed from the MLO view images ($p = 0.0004$). However, the performance of the ANN trained by features computed from the CC view was not significantly different with that of the ANN trained by features computed from the MLO view images ($p = 0.710$). The results demonstrate that our optimized scoring fusion method enabled to extract and derive useful information from both views, and thus helped improve CAD performance.

Since the classification score generated by the scoring fusion ANN ranges from 0 to 1.2, we used a score of 0.6 as a threshold and computed a confusion matrix as displayed in table 3. It indicates that 71.8% (755 of 1052) of cases are correctly classified by the scoring fusion method whereas 28.2% (297 of 1052) of cases are misclassified at this classification threshold. In the cancer case subgroup, the classification accuracy is 73.4% (491 of 669). In the benign case subgroup, 68.9% (264 of 383) of cases are correctly classified at this classification threshold.

Figure 6 displays the ROC curves that were obtained for different cancer subgroups within our dataset using the scoring fusion ANN to combine the classification scores of both views, namely (1) all 669 positive cases; (2) 625 cancer cases and (3) 44 interval cancer cases. Table 4 displays the average and standard deviation of AUC results computed for each cancer subgroup using the 10-fold cross-validation experiment.

The results in table 4 and figure 6 indicate that our scheme performed higher in classifying screening mammography detected cancer cases than in classifying the interval cancer cases. We also analyzed performance of our scheme by stratifying our testing cases in four mammographic density BIRADS ratings. At 90% specificity, the classification sensitivity levels achieved at each of the four BIRADS categories (1-4) were 46.2% (12/26), 63.2% (144/228), 47.3% (187/395), and 55.0% (11/20), respectively. The results did not indicate a clear trend that performance of our classification scheme is mammographic density dependent.

In addition, to analyze the effectiveness of computing the relevant global features for the case classification purpose, we divided the features into two separate sets, namely: (1) features computed on the dense breast regions only, and (2) features computed on the whole breast regions, and repeated the experiments. For the features computed on the dense breast regions, $AUC = 0.787 \pm 0.038$, while for features computed on the whole breast regions, $AUC = 0.792 \pm 0.040$. Although the feature set computed on the whole breast regions slightly outperforms the features computed on the dense breast regions, the difference is not significant ($p = 0.653$). Performance differences between the combined whole and dense breast feature set are not significantly different with the dense breast feature set ($p = 0.483$), and the whole breast feature set ($p = 0.921$), respectively. It indicates that using the features computed from the whole breast regions is a sufficient and more computationally efficient approach.

4. Discussion

This study has a number of unique characteristics. First, the previous CAD efforts focused on improving lesion segmentation and then classifying the lesions into positive or negative using a set of local image features computed around the segmented lesions (Elter and Horsch, 2009). The performance levels of these lesion based schemes vary widely due to the use of different image processing algorithms, feature classifiers, and the training/testing datasets. For example, the reported performance levels ranged from $AUC=0.6$ (Heidt *et al.*, 2009), $AUC=0.81$ (Varela *et al.*, 2006), $AUC=0.84$ (Shi *et al.*, 2008), to $AUC=0.94$ (Guliatto *et al.*, 2008). However, our new CAD scheme is not another lesion-based classification

scheme. It is the first *global* mammographic image feature based scheme that incorporates information from *all four FFDM images* of a patient. In this study, the scheme yielded a classification performance level of AUC=0.79. Although our scheme has a lower performance level than many previously reported lesion based classification schemes and may not be used alone at its current performance level, our scheme does not compete with the existing lesion-based CAD schemes, but provides new and/or supplementary diagnostic information to the existing CAD schemes. Since the two types of CAD schemes are developed based on very different concepts and are unlikely to be highly correlated, similar to that demonstrated in our previous study to improve mass detection performance (Wang *et al.*, 2012), fusion of the classification scores generated by our scheme and the existing lesion based CAD schemes will have potential to improve recalled case classification results and thus help reduce false-positive recalls.

Second, although CAD of FFDM images has been reported in many studies, the testing datasets of FFDM are typically small. For example, one previous study assessing the performance of a commercialized CAD scheme of FFDM images (SecondLook, v.7.2; iCAD, Inc.) used 123 cancer cases (The *et al.*, 2009) and another study used a dataset involving 203 cases to assess another commercialized CAD scheme (ImageChecker M1000-DM system, v.3.1; R2 Technology, Sunnyvale, Calif) (Yang *et al.*, 2007). In this study, we assessed performance of our CAD scheme using a relatively large and diverse FFDM image dataset involving 1052 cases, which to our knowledge, is the largest study performed for CAD on FFDM images to classify between recalled malignant and benign cases to date. The testing results of our new CAD scheme indicate that useful and relatively higher discriminatory information can be derived from the quantitative analysis of global mammographic image features, in particular the fusion of features from four view FFDM images for the classification task using our approach. The results showed that the classification scores after information fusion from both CC and MLO views produced a higher AUC result that significantly outperformed the individual scores of the single views at the 5% significance level.

Third, although a semi-automated approach based on Cumulus software (Byng *et al.*, 1994) has been widely used as a standard tool to assess mammographic *PD*, it has a number of disadvantages. (1) The breast region is segmented simply by applying a global threshold on the whole image. This approach is too simplistic, and can result in the inclusion of the skin and other unwanted regions in the final estimation of *PD*, thus decreasing its final accuracy. (2) It is operator-dependent, and its final measurement is also subject to intra- and inter-observer variability (Heine *et al.*, 2012; Li *et al.*, 2012). (3) It cannot be implemented into a fully-automated CAD scheme. Hence, this semi-automated approach can incur substantial amounts of time and costs for a large-scale deployment in a clinical setting/environment. In this study, we investigated and tested a number of new features to estimate the global mammographic features including a new fully-automated measure that enables to more consistently and efficiently estimate mammographic *PD* that is defined in a similar way as the Cumulus software. In addition, our data analysis result also showed that the weighting factor of the computed mammographic *PD* feature is highest among other tissue texture

features (as shown in Table 1), which indicates that mammographic *PD* plays an important role in developing CAD schemes of FFDM case classification.

Fourth, the fixed weighting method has been used in previous studies to combine the CAD scores of different view images (Wang *et al.*, 2011; Liu *et al.*, 2005), in this study we also tested the fixed weighting method to combine two ANN-generated classification scores computed from the CC and MLO view images. Comparing with the result of our previous study (Wang *et al.*, 2011), we found that the optimal weighting factors on the two detection or classification scores computed from the CC and MLO view images were dataset dependent. Hence, we applied and tested a new scoring fusion approach to combine information from the CC and MLO views in this study (as shown in Figure 4). This is an adaptive approach in which the number of hidden nodes and the connection weights of a scoring fusion ANN are automatically determined and fully optimized on the training subsets whereby the testing subsets are kept completely independent and unseen with our approach. The results indicated that combining image features from four view FFDM images enabled to improve CAD performance.

Despite a new approach and encouraging results, this is a very preliminary study with a number of limitations. First, the testing cases were directly selected from an existing image database for developing conventional CAD schemes of detecting breast masses in screening FFDM images, which emphasizes on malignant cases and screening negative (not recalled) cases. Hence, the ratio between malignant and benign cases does not represent the cancer prevalence ratio of recalled cases in the clinical practice. Although in this study the classification performance levels between the malignant and benign case subgroups are not significantly skewed (as shown in Table 3), we will continue to collect more false-positive recalled cases to increase the size of benign cases and rebalance our dataset in future studies. Second, a large number of global mammographic density and tissue distribution features (as shown in Table 1) was applied in this CAD scheme, which is not an optimal approach in CAD development. The selection of a smaller and more effective feature set from this relatively large initial feature pool will be one of the major tasks in the next step of our study using the rebalanced dataset to improve the robustness of the scheme. Third, this is an approval of concept and technology development study. Whether a current classification performance (e.g., AUC=0.79) is sufficient to help radiologists reduce false-positive recalls has not been tested. Fourth, the study results indicate that our CAD scheme had lower performance in detecting interval cancer than malignant cancer cases. Although this is consistent with the clinical practice because the interval cancers are more difficult to be detected, how to use CAD to help radiologists detect more early cancer signs to reduce the interval cancer cases will be an interesting topic in our future studies. Finally, as this is the first laboratory-based retrospective study, the performance and robustness of our CAD scheme needs to be further examined and validated using other independent databases in the future.

In summary, this study addressed an important clinical challenge in screening mammography to reduce the false-positive recalls using a unique CAD approach. Comparing to other more advanced imaging approaches, such as a recently developed new imaging modality, digital breast tomosynthesis (DBT), which enables to significantly reduce

false-positive recall rates (e.g., up to 30%) with slightly lower cancer detection sensitivity (Gur *et al.*, 2009), a CAD of mammogram based approach is more cost-effective, if successful. Unfortunately, although great research efforts have been made in developing CAD schemes to classify between malignant and benign lesions, current CAD performance remains lower and not robust. To help solve this issue, this study presented the first CAD scheme that incorporates global mammographic image features computed from four-view FFDM images and directly applies to a relatively large and diverse recalled FFDM image dataset to classify between malignant and benign cases. The study demonstrated the feasibility of using this new CAD approach to provide a new quantitative classification index that has relatively higher discriminatory power. As a result, it may have potential to provide useful supplementary information to other existing classification methods and/or biomarkers to help improve the efficacy of screening mammography by significantly reducing false-positive recall rates.

Acknowledgments

This work is support in part by Grant of R01 CA160205 from the National Cancer Institute, National Institutes of Health.

References

- American Cancer Society. What are the key statistics about breast cancer?. 2014. <http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics>
- American College of Radiology. Bi-RADS Breast Imaging Reporting and Data System Breast Imaging Atlas. Reston, VA: 2003.
- Bertrand KA, Tamimi RM, Scott CG, Jensen MR, Pankratz VS, Visscher D, Norman A, Couch F, Shepherd J, Fan B, Chen YY, Ma L, Beck AH, Cummings SR, Kerlikowske K, Vachon CM. Mammographic density and risk of breast cancer by age and tumor characteristics. *Breast cancer research* : BCR 15 R104. 2013 (Epub ahead of print).
- Brawley OW. Risk-based mammography screening: an effort to maximize the benefits and minimize the harms. *Ann Intern Med*. 2012; 156:662–3. [PubMed: 22547477]
- Brodersen J, Siersma VD. Long-term psychosocial consequences of false-positive screening mammography. *Ann Fam Med*. 2013; 11:106–15. [PubMed: 23508596]
- Buist DSM, Anderson ML, Haneuse SJPA, Sickles EA, Smith RA, Carney PA, Taplin SH, Rosenberg RD, Geller BM, Onega TL, Monsees BS, Bassett LW, Yankaskas BC, Elmore JG, Kerlikowske K, Miglioretti DL. Influence of annual interpretive volume on screening mammography performance in the United States. *Radiology*. 2011; 259:72–84. [PubMed: 21343539]
- Byng JW, Boyd NF, Fishell E, Jong RA, Yaffe MJ. The quantitative analysis of mammographic densities. *Phys Med Biol*. 1994; 39:1629–38. [PubMed: 15551535]
- Byng JW, Yaffe MJ, Lockwood GA, Little LE, Tritchler DL, Boyd NF. Automated analysis of mammographic densities and breast carcinoma risk. *Cancer*. 1997; 80:66–74. [PubMed: 9210710]
- Catarious DM Jr. Baydush AH, Floyd CE Jr. Characterization of difference of Gaussian filters in the detection of mammographic regions. *Med Phys*. 2006; 33:4104–14. [PubMed: 17153390]
- Chang Y-H, Wang X-H, Hardesty LA, Chang TS, Poller WR, Good WF, Gur D. Computerized assessment of tissue composition on digitized mammograms. *Acad. Radiol*. 2002; 9:899–905. [PubMed: 12186438]
- Consedine NS. A false-positive on screening mammography has a negative psychosocial impact up to 3 years after receiving the all clear. *Evidence Based Mental Health*. 2013; 16:115. [PubMed: 23852830]

- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 11:837–45. [PubMed: 3203132]
- Elter M, Horsch A. CADx of mammographic masses and clustered microcalcifications: A review. *Med. Phys.* 2009; 36:2052–68. [PubMed: 19610294]
- Eltonsy NH, Tourassi GD, Elmaghraby AS. A concentric morphology model for the detection of masses in mammography. *IEEE Transactions on Medical Imaging*. 2007; 26:880–9. [PubMed: 17679338]
- Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D'Orsi C, Elmore JG, Barlow WE, Consortium f t B C S. Effectiveness of Computer-Aided Detection in Community Mammography Practice. *J Natl Cancer Inst*. 2011; 103:1152–61. [PubMed: 21795668]
- Glide-Hurst CK, Duric N, Littrup P. A new method for quantitative analysis of mammographic density. *Med Phys*. 2007; 34:4491–8. [PubMed: 18072514]
- Guliatto D, Rangayyan RM, Carvalho JD, Santiago SA. Polygonal modeling of contours of breast tumors with the preservation of spicules. *IEEE Trans Biomed Eng*. 2008; 55:14–20. [PubMed: 18232342]
- Gur D, Abrams GS, Chough DM, Ganott MA, Hakim CM, Perrin RL, Rathfon GY, Sumkin JH, Zuley ML, Bandos AI. Digital breast tomosynthesis: observer performance study. *AJR Am J Roentgenol*. 2009; 193:586–91. [PubMed: 19620460]
- Gur D, Stalder JS, Hardesty LA, Zheng B, Sumkin JH, Chough DM, Shindel BE, Rockette HE. Computer-aided detection performance in mammographic examination of masses: assessment. *Radiology*. 2004a; 233:418–23. [PubMed: 15358846]
- Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim C, Hardesty L, Poller WR, Shah R, Wallace L. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *J Natl Cancer Inst*. 2004b; 96:185–90. [PubMed: 14759985]
- Haneuse S, Buist DS, Miglioretti DL, Anderson ML, Carney PA, Onega T, Geller BM, Kerlikowske K, Rosenberg RD, Yankaskas BC, Elmore JG, Taplin SH, Smith RA, Sickles EA. Mammographic interpretive volume and diagnostic mammogram interpretation performance in community practice. *Radiology*. 2012; 262:69–79. [PubMed: 22106351]
- Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. *IEEE Trans. Syst. Man, Cybern*. 1973; 3:610–21.
- Heidt, S-R.; Elter, M.; Wittenberg, T.; Paulus, D. *Bildverarbeitung für die Medizin 2009*. Meinzer, H-P., et al., editors. Springer Berlin Heidelberg; 2009. p. 287-91.
- Heine JJ, Scott CG, Sellers TA, Brandt KR, Serie DJ, Wu FF, Morton MJ, Schueler BA, Couch FJ, Olson JE, Pankratz VS, Vachon CM. A novel automated mammographic density measure and breast cancer risk. *J Natl Cancer Inst*. 2012; 104:1028–37. [PubMed: 22761274]
- Hubbard RA, Kerlikowske K, Flowers CI, et al. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: A cohort study. *Ann Intern Med*. 2011; 155:481–92. [PubMed: 22007042]
- Ikeda DM, Sickles EA. Second-screening mammography: one versus two views per breast. *Radiology*. 1988; 168:651–6. [PubMed: 3406393]
- Kom G, Tiedeu A, Kom M. Automated detection of masses in mammograms by local adaptive thresholding. *Comput Biol Med*. 2007; 37:37–48. [PubMed: 16487954]
- Law YN, Lieng MK, Li J, Khoo DA-A. Automated Breast Tissue Density Assessment using High Order Regional Texture Descriptors in Mammography. *Proc.SPIE 9035, Medical Imaging 2014: Computer-Aided Diagnosis*. 2014 90351Q.
- Li H, Giger ML, Olopade OI, Margolis A, Chinander MR. Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms. *Acad. Radiol*. 2005; 12:863–73. [PubMed: 16039540]
- Li J, Szekely L, Eriksson L, Heddson B, Sundbom A, Czene K, Hall P, Humphreys K. High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer. *Breast cancer research : BCR* 14 R114. 2012 (Epub ahead of print).

- Linet MS, Slovis TL, Miller DL, Kleinerman R, Lee C, Rajaraman P, Berrington de Gonzalez A. Cancer risks associated with external radiation from diagnostic imaging procedures. *CA: A Cancer Journal for Clinicians*. 2012; 62:75–100.
- Liu B, Metz CE, Jiang Y. Effect of correlation on combining diagnostic information from two images of the same patient. *Med Phys*. 2005; 32:3329–38. [PubMed: 16372412]
- Manduca A, Carston MJ, Heine JJ, Scott CG, Pankratz VS, Brandt KR, Sellers TA, Vachon CM, Cerhan JR. Texture Features from Mammographic Images and Risk of Breast Cancer. *Cancer Epidemiol Biomarkers Prev*. 2009; 18:837–45. [PubMed: 19258482]
- Maskarinec G, Verheus M, Steinberg FM, Amato P, Cramer MK, Lewis RD, Murray MJ, Young RL, Wong WW. Various doses of soy isoflavones do not modify mammographic density in postmenopausal women. *The Journal of nutrition*. 2009; 139:981–6. [PubMed: 19321587]
- Mudigonda NR, Rangayyan RM, Desautels JE. Gradient and texture analysis for the classification of mammographic masses. *IEEE Trans. Med. Imaging*. 2000; 19:1032–43. [PubMed: 11131493]
- Mudigonda NR, Rangayyan RM, Desautels JE. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. *IEEE Trans Med Imaging*. 2001; 20:1215–27. [PubMed: 11811822]
- Noh JJ, Maskarinec G, Pagano I, Cheung LWK, Stanczyk FZ. Mammographic densities and circulating hormones: A cross-sectional study in premenopausal women. *The Breast*. 2006; 15:20–8. [PubMed: 16000251]
- Qian W, Song D, Lei M, Sankar R, Eikman E. Computer-aided mass detection based on ipsilateral multiview mammograms. *Acad Radiol*. 2007; 14:530–8. [PubMed: 17434066]
- Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*. 2011; 12:77. [PubMed: 21414208]
- Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986; 323:533–6.
- Shi J, Sahiner B, Chan HP, Ge J, Hadjiiski L, Helvie MA, Nees A, Wu YT, Wei J, Zhou C, Zhang Y, Cui J. Characterization of mammographic masses based on level set segmentation with new image features and patient information. *Med. Phys*. 2008; 35:280–90. [PubMed: 18293583]
- Sickles EA, Weber WN, Galvin HB, Ominsky SH, Solitto RA. Baseline screening mammography: one vs two views per breast. *AJR Am J Roentgenol*. 1986; 147:1149–53. [PubMed: 3490749]
- Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology*. 2002; 224:861–9. [PubMed: 12202726]
- Smith RA, Cokkinides V, Brooks D, Saslow D, Shah M, Brawley OW. Cancer screening in the United States, 2011: A review of current American Cancer Society guidelines and issues in cancer screening. *CA Cancer J Clin*. 2011; 61:8–30. [PubMed: 21205832]
- Tabar L, Vitak B, Chen HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer*. 2001; 91:1724–31. [PubMed: 11335897]
- Tan M, Zheng B, Ramalingam P, Gur D. Prediction of near-term breast cancer risk based on bilateral mammographic feature asymmetry. *Acad Radiol*. 2013; 20:1542–50. [PubMed: 24200481]
- Tang X. Texture information in run-length matrices. *IEEE Trans. Image Proc*. 1998; 7:1602–9.
- The JS, Schilling KJ, Hoffmeister JW, Friedmann E, McGinnis R, Holcomb RG. Detection of breast cancer with full-field digital mammography and computer-aided detection. *AJR Am J Roentgenol*. 2009; 192:337–40. [PubMed: 19155392]
- Tzikopoulos S, Mavroforakis ME, Georgiou HV, Dimitropoulos N, Theodoridis S. A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry. *Comput Methods Programs Biomed*. 2011; 102:47–63. [PubMed: 21306782]
- van Engeland S, Karssemeijer N. Combining two mammographic projections in a computer aided mass detection method. *Med Phys*. 2007; 34:898–905. [PubMed: 17441235]
- Varela C, Tahoces PG, Mendez AJ, Souto M, Vidal JJ. Computerized detection of breast masses in digitized mammograms. *Comput Biol Med*. 2007; 37:214–26. [PubMed: 16620805]

- Varela C, Timp S, Karssemeijer N. Use of border information in the classification of mammographic masses. *Phys. Med. Biol.* 2006; 51:425–41. [PubMed: 16394348]
- Velikova M, Samulski M, Lucas PJ, Karssemeijer N. Improved mammographic CAD performance using multi-view information: a Bayesian network framework. *Phys Med Biol.* 2009; 54:1131–47. [PubMed: 19174596]
- Wald NJ, Murphy P, Major P, Parkes C, Townsend J, Frost C. UKCCCR multicentre randomised controlled trial of one and two view mammography in breast cancer screening. *BMJ (Clinical research ed.)*. 1995; 311:1189–93.
- Wang X, Lederman D, Tan J, Wang XH, Zheng B. Computerized prediction of risk for developing breast cancer based on bilateral mammographic breast tissue asymmetry. *Med Eng Phys.* 2011; 33:934–42. [PubMed: 21482168]
- Wang X, Li L, Xu W, Liu W, Lederman D, Zheng B. Improving the performance of computer-aided detection of subtle breast masses using an adaptive cueing method. *Physics In Medicine And Biology.* 2012; 57:561–75. [PubMed: 22218075]
- Wang XH, Good WF, Chapman BE, Chang Y-H, Poller WR, Chang TS, Hardesty LA. Automated assessment of the composition of breast tissue revealed on tissue-thickness-corrected mammography. *AJR Am J Roentgenol.* 2003; 180:257–62. [PubMed: 12490516]
- Wei J, Chan HP, Sahiner B, Zhou C, Hadjiiski LM, Roubidoux MA, Helvie MA. Computer-aided detection of breast masses on mammograms: dual system approach with two-view analysis. *Med Phys.* 2009; 36:4451–60. [PubMed: 19928076]
- Wei J, Chan HP, Wu YT, et al. Association of computerized mammographic parenchymal pattern measure with breast cancer risk: a pilot case-control study. *Radiology.* 2011a; 260:42–9. [PubMed: 21406634]
- Wei J, Chan HP, Zhou C, Wu YT, Sahiner B, Hadjiiski LM, Roubidoux MA, Helvie MA. Computer-aided detection of breast masses: four-view strategy for screening mammography. *Med Phys.* 2011b; 38:1867–76. [PubMed: 21626920]
- Wei, X. Gray Level Run Length Matrix Toolbox v1.0. Beijing Aeronautical Technology Research Center; 2007. <http://www.mathworks.com/matlabcentral/fileexchange/17482-gray-level-run-length-matrix-toolbox>. [12 December 2013]
- Yaffe MJ, Mainprize JG. Risk of radiation-induced breast cancer from mammographic screening. *Radiology.* 2011; 258:98–105. [PubMed: 21081671]
- Yang SK, Moon WK, Cho N, Park JS, Cha JH, Kim SM, Kim SJ, Im JG. Screening mammography-detected cancers: sensitivity of a computer-aided detection system applied to full-field digital mammograms. *Radiology.* 2007; 244:104–11. [PubMed: 17507722]
- Yuan Y, Giger ML, Li H, Sennett C. Correlative feature analysis on FFDM. *Med Phys.* 2008; 35:5490–500. [PubMed: 19175108]
- Zheng B, Leader JK, Abrams GS, Lu AH, Wallace LP, Maitz GS, Gur D. Multiview-based computer-aided detection scheme for breast masses. *Med. Phys.* 2006; 33:3135–43. [PubMed: 17022205]
- Zheng B, Sumkin JH, Zuley ML, Lederman D, Wang X, Gur D. Computer-aided detection of breast masses depicted on full-field digital mammograms: a performance assessment. *Br J Radiol.* 2012; 85:e153–e61. [PubMed: 21343322]
- Zheng B, Tan J, Ganott MA, Chough DM, Gur D. Matching breast masses depicted on different views a comparison of three methods. *Acad Radiol.* 2009; 16:1338–47. [PubMed: 19632867]
- Zhou C, Chan HP, Petrick N, Helvie M, Goodsitt M, Sahiner B, Hadjiiski L. Computerized image analysis: estimation of breast density on mammograms. *Med Phys.* 2001; 28:1056–69. [PubMed: 11439475]

(a)

(b)

Figure 1.

An example of a case in the subgroup of verified cancer cases, which shows the left and right breasts of the (a) craniocaudal (CC) and (b) mediolateral oblique (MLO) view images (cancer pointed by an arrow) from our dataset. The cancer was detected by a radiologist, and was later confirmed in biopsy and pathology examinations.

Figure 2.

Histogram distribution of two case subgroups in our dataset (benign and cancer) in the four categories of mammographic density (Breast Imaging Reporting and Data System [BI-RADS]) ratings.

Figure 3.
Flowchart of our breast segmentation scheme.

Figure 4.

A diagram illustrates a two-stage ANN classification scheme in which the final classification score is generated by adaptive fusion of information derived from two ANNs trained and optimized using the craniocaudal (CC) and mediolateral oblique (MLO) view image features.

Figure 5. Comparisons of three receiver operating characteristic (ROC) curves generated by three ANNs trained using the features computed from the CC and MLO views alone, and the scoring fusion approach of both views.

Figure 6. Comparison of three receiver operating characteristic (ROC) curves generated by ANNs trained by our scoring fusion approach for different cancer subgroups within our dataset, namely all 669 positive cases, 625 cancer cases and 44 interval cancer cases.

Table 1

Computed features according to their relevant grouping, and their corresponding notes/descriptions.

Feature group	ID	Notes/Description
Generic statistical (gray value) features *	1-8	Mean, standard deviation, skewness, and kurtosis of gray values.
Fractal dimension feature (Chang <i>et al.</i> , 2002) *	9-10	Computation of fractal dimension feature originally described in (Chang <i>et al.</i> , 2002)
Mean and maximum values of gray level run length-based texture features, computed along 0, 45, 90, and 135 degrees (Tang, 1998) *	11-54	Short run emphasis, long run emphasis, low gray level run emphasis, high gray level run emphasis, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis, long run high gray level emphasis, gray level non-uniformity, run length non-uniformity, and run percentage
X-axis and y-axis histogram (or cumulative projection) features (Tzikopoulos <i>et al.</i> , 2011) *	55-74	Mean, standard deviation, skewness, kurtosis, and median values of the cumulative projection histograms
Mean and maximum values of gray level co-occurrence matrix (GLCM)-based features (Haralick <i>et al.</i> , 1973) *	75-90	Statistics computed from a gray level co-occurrence matrix along four different directions (0, 45, 90, and 135 degrees), as follows (Haralick <i>et al.</i> , 1973): $Contrast = \sum_{i,j} i - j ^2 p(i, j)$ $Correlation = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j)p(i, j)}{\sigma_i \sigma_j}$ $Energy = \sum_{i,j} p(i, j)^2$ $Homogeneity = \sum_{i,j} \frac{p(i, j)}{1 + i - j }$
Regional descriptor	91	Area (number of pixels) of the segmented breast region
Mammographic density estimation/quantification	92	We defined a new percentage density (PD) measure computed as: $PD = \frac{N(DeT)}{N(BrT)}$ where $N(R)$ is the number of pixels in region R , BrT is the segmented breast tissue region, and DeT is the estimated dense tissue region

* Features computed on both the whole and 50% dense breast regions

Table 2

Average AUC results and corresponding standard deviation intervals of the three ANNs optimized on the (1) CC and (2) MLO views, and (3) the optimized scoring fusion method derived from both views, computed over the 10-fold cross-validation scheme.

No.	ANN	AUC
1	Features computed from the CC view	0.770±0.022
2	Features computed from the MLO view	0.761±0.037
3	Optimized scoring fusion method from both views	0.793±0.026

Table 3

Confusion matrix when applying a threshold of 0.6 on the output of the scoring fusion ANN.

Actual ↓	Benign cases	Malignant cases
Benign cases	264	119
Malignant cases	178	491

Table 4

Average AUC results and corresponding standard deviation intervals of the ANNs optimized using our scoring fusion approach for different cancer subgroups ((1) all 669 positive cases, (2) 625 cancer cases and (3) 44 interval cancer cases) within our database computed over the 10-fold cross-validation experiment.

No.	Cancer subgroup/subtype	AUC
1	All positive cases	0.793±0.026
2	Cancer cases	0.802±0.028
3	Interval cancer cases	0.705±0.171