HHS Public Access

Author manuscript

Phys Med Biol. Author manuscript; available in PMC 2016 June 07.

Published in final edited form as:

Phys Med Biol. 2015 June 7; 60(11): 4413–4427. doi:10.1088/0031-9155/60/11/4413.

A new approach to develop computer-aided detection schemes of digital mammograms

Maxine Tan[1,4], **Wei Qian**[2], **Jiantao Pu**[3], **Hong Liu**[1], and **Bin Zheng**[1,3]

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

[2] Department of Electrical and Computer Engineering, University of Texas, El Paso, TX 79968

[3] Department of Radiology, University of Pittsburgh, Pittsburgh, PA 15213

Abstract

The purpose of this study is to develop a new global mammographic image feature analysis based computer-aided detection (CAD) scheme and evaluate its performance in detecting positive screening mammography examinations. A dataset that includes images acquired from 1896 full-field digital mammography (FFDM) screening examinations was used in this study. Among them, 812 cases were positive for cancer and 1084 were negative or benign. After segmenting the breast area, a computerized scheme was applied to compute 92 global mammographic tissue density based features on each of four mammograms of the craniocaudal (CC) and mediolateral oblique (MLO) views. After adding three existing popular risk factors (woman's age, subjectively rated mammographic density, and family breast cancer history) into the initial feature pool, we applied a Sequential Forward Floating Selection (SFFS) feature selection algorithm to select relevant features from the bilateral CC and MLO view images separately. The selected CC and MLO view image features were used to train two artificial neural networks (ANNs). The results were then fused by a third ANN to build a two-stage classifier to predict the likelihood of the FFDM screening examination being positive. CAD performance was tested using a ten-fold cross-validation method. The computed area under the receiver operating characteristic curve was AUC=0.779±0.025 and the odds ratio monotonically increased from 1 to 31.55 as CAD-generated detection scores increased. The study demonstrated that this new global image feature based CAD scheme had a relatively higher discriminatory power to cue the FFDM examinations with high risk of being positive, which may provide a new CAD-cueing method to assist radiologists in reading and interpreting screening mammograms.

Keywords

Computer-aided detection (CAD) of mammograms; Full-field digital mammography; Mammographic image feature analysis; Improvement of screening mammography efficacy

[4] Corresponding author, Maxine.Y.Tan-1@ou.edu.

1. Introduction

Breast cancer is one of the most prevalent cancers in women and has a high mortality rate (Siegel *et al.*, 2013). Due to the heterogeneity of malignant tumors, cancer screening is widely considered an effective approach to detect breast cancer at an early stage. In the last four decades, the promotion of breast cancer screening along with the advancement of cancer treatment methods significantly reduced mortality rates of breast cancer patients (Berry *et al.*, 2005). Among all available breast cancer screening methods and/or imaging modalities, mammography is the only imaging modality that is accepted in conducting population-based breast cancer screening to date (Smith *et al.*, 2011). However, due to the large variability of breast lesions and overlapping dense fibro-glandular tissue in 2D projection images, mammography performance is not satisfactory in both cancer detection sensitivity and specificity (Fenton *et al.*, 2006). Studies have shown that the sensitivity of screening mammography is lower among women who are younger (e.g., < 50 years old) (Peer *et al.*, 1996), have dense breasts (Mandelson *et al.*, 2000), use hormone replacement therapy (Laya *et al.*, 1996), and carry certain breast cancer susceptibility genes (Kriege *et al.*, 2004). One recent multi-institutional prospective study reported that mammography detected only 53.2% (59 of 111) cancers among a group of 2098 women with elevated breast cancer risk (Berg *et al.*, 2012). In addition, mammography has lower specificity with high false-positive recalls that lead to generating a large number of benign biopsies (Hubbard *et al.*, 2011), which results in long-term psychosocial consequences or harms to many women who routinely participate in screening mammography examinations (Brodersen and Siersma, 2013). As a result, improving the efficacy of screening mammography remains an important clinical issue in breast cancer screening (Brawley, 2012).

As previous studies have shown that a high percentage of missed or overlooked breast cancers in prospective clinical practice were visually detectable in retrospective reviews (Birdwell *et al.*, 2001), to increase breast cancer detection sensitivity of screening mammography, a large number of computer-aided detection (CAD) schemes of mammograms have been developed and used as “a second reader” to assist radiologists in reading and interpreting mammograms. The current commercialized CAD schemes are single-image and lesion-based schemes. Despite the relatively-high positive lesion detection sensitivity, current CAD schemes generate substantially high false-positive detection rates and also have high correlation in positive lesion detection with radiologists (Gur *et al.*, 2004a). Studies have shown that using current CAD as “a second reader” is unable to help improve radiologists’ performance in prospective clinical practice (Gur *et al.*, 2004b; Fenton *et al.*, 2011). Hence, research efforts continue to attempt to improve the clinical utility of using CAD schemes to assist radiologists, which include the use of (1) an interactive approach that only shows CAD-cued marks with detection scores that match the suspicious regions queried by radiologists on the mammograms (Samulski *et al.*, 2010) and (2) a new computer-aided reading protocol that determines the display order of regions of interest based on the findings of a CAD scheme aiming to improve efficiency of radiologists’ work flow while maintaining their detection performance (Moin *et al.*, 2011).

In this study, we propose a different approach to improve the efficacy of applying CAD to assist radiologists in reading and interpreting screening mammograms. Our hypothesis is that since in a population-based annual screening environment the cancer prevalence rate is quite low (< 1%), this is probably one of the primary reasons that cause radiologists to miss or overlook subtle positive cases from the overwhelmingly negative cases as well as to generate high false-positive recalls. To help solve this problem, we need to develop a new CAD approach and/or scheme that can cue the warning sign on the cases with high risk of being positive. Hence, these risk cueing signs can warn radiologists to pay more attention in reading and detecting the suspicious lesions or signs depicted on these high-risk images. This may help radiologists reduce the number of missed or overlooked malignant lesions that are "visually detectable" in the retrospective review (Birdwell *et al.*, 2001). Meanwhile, since this is not a region based cueing as used in existing CAD schemes (Zheng *et al.*, 2012a; Kallenberg and Karssemeijer, 2008; The *et al.*, 2009), there is no need for radiologists to rule-out a large number of false-positive cues. As a result, this new case based CAD cueing method is different from the lesion-based cueing made by conventional CAD schemes, which may have different impact on radiologists' decisions in recalling suspicious cases. In order to test our hypothesis, in this study we applied a four-view global image feature analysis concept to optimize a new CAD scheme using a relatively large and diverse image dataset. Unlike existing lesion-based CAD schemes, our CAD scheme does not detect and segment suspicious lesions depicted on each image. It analyzes the bilateral global mammographic image features and their difference to generate a case or four-view image examination based likelihood score of being positive for cancer. We then assessed performance of this CAD scheme in classifying between the groups of positive (cancer) and negative (cancer-free) cases. The detailed description of the technical development of our new approach and experimental results is presented in the following sections.

2. Materials and methods

2.1. A dataset of digital mammography images

Under an institutional review board (IRB) approved image data collection protocol, an IRB-certified research staff randomly selected screening mammography cases based on the screening outcome (positive or negative for cancer) recorded in the existing clinical database of University of Pittsburgh Medical Center without viewing the mammograms. From the selected case ID number, the corresponding clinical information (including age, family history of breast cancer, subjectively-rated breast density based on BI-RADS categories, and verified diagnostic result) were collated. All FFDM examinations were acquired using Hologic Selenia FFDM systems after 2006. After a de-identification process, the fully anonymized FFDM images and the corresponding clinical information were transferred and stored in the research database for our studies. The details of our image data collection protocol has been previously reported (Zheng *et al.*, 2012a). Since this image data collection remains active to date, new cases are continually being added to our study database. Different subsets of this database have been used in a number of our previous studies (e.g., (Wang *et al.*, 2012; Zheng *et al.*, 2012b; Tan *et al.*, 2013; Tan *et al.*, 2014b)). In this study, we assembled a dataset that includes 1896 cases. Among them, 812 cases have been verified as positive, which are divided into three subgroups, namely (1) 746 verified cancer cases

that were detected during FFDM screening and confirmed by biopsy results; (2) 39 "interval" cancer cases in which cancers were detected in the interval between two screening examinations; and (3) 27 high-risk precancer cases (e.g., lobular carcinoma in situ) in which lesions were surgically removed.

The remaining 1084 cases are negative or benign. Based on Breast Imaging Reporting and Data System (BI-RADS) (Sickles *et al.*, 2013), 618 were screening negative cases (BI-RADS 1), which were not recalled by the radiologists in the original image reading and interpretation) and 466 were either not-recalled benign cases (BI-RADS 2) or recalled cases (i.e., BI-RADS 3 or 4) but later proven as benign through the additional imaging workup and/or biopsy. All of these cases remain cancer-free for at least two years (in next two sequential FFDM screening examinations).

Each case in this dataset has four FFDM images representing the craniocaudal (CC) and mediolateral oblique (MLO) views of the left and right breasts. The average age and standard deviation of the women in the positive and negative case groups are 59.6±12.8 and 50.4±9.2 years old, respectively. Figure 1 shows the distribution of mammographic density subjectively rated by radiologists using BI-RADS categories in the three groups of positive (with verified malignant lesions), screening negative (not recalled) and recalled/benign cases. The data analysis using an unpaired *t*-test showed that the BI-RADS based breast density distribution among these three case groups had no statistically significant difference (i.e., p =0.517 between cancer and benign case groups, p=0.725 between cancer and negative case groups, as well as p=0.719 between benign and negative case groups).

2.2. A new four-view CAD scheme of global mammographic image feature analysis

We recently developed and preliminarily tested a new CAD scheme to detect and analyze global bilateral mammographic image features of both CC and MLO views of the left and right breasts (Tan *et al.*, 2014b). We applied this scheme to test its feasibility of helping distinguish between true-positive (cancer) and false-positive recalled (benign) cases. In this study, the same concept was applied to optimize a new CAD scheme for a different application purpose namely, classifying between positive (cancer) and negative (non-cancer) FFDM examinations with much smaller number of image features in order to increase robustness of the CAD scheme. In brief, we used the following steps to optimize the new CAD scheme in this study.

First, we modified and applied a computerized scheme to segment the breast area depicted on each image (Zheng *et al.*, 2006). Based on the gray level histogram of the whole image, the scheme used an iterative searching method to detect the smoothest curvature between the breast tissue and background (air). After discarding the background region, a morphological erosion operation was performed to remove the skin region from the segmented breast. For the MLO view images, an additional step was applied to detect the chest wall or pectoral muscle line. This was performed by detecting the pixel with the maximum gradient iteratively in each row of the image, and employing a linear regression method to fit all identified pixels on a straight line to represent the chest wall boundary. Finally, all pixels within the pectoral muscle were discarded from the segmented breast region.

Second, we applied the computerized scheme to initially compute 92 global mammographic density and texture based image features from each view of four FFDM images. The detailed description of each feature definition and computational method has been reported in our previous study (Tan *et al.*, 2014b). Basically, these features are divided into the following categories: (1) the statistical pixel value (or intensity) based features (i.e., mean, standard deviation, skewness, and kurtosis of the pixel gray values) (Wang *et al.*, 2011), (2) fractal dimension based features to quantitatively assess breast tissue composition and mammographic density (Chang *et al.*, 2002), (3) gray level run length based texture features computed from the gray level resolution reduced images (from 4095 to 256 gray levels) along four different directions (Tang, 1998), (4) the first-order statistics of the *x*-axis and *y*-axis histogram (cumulative projection) based features presented by Tzikopoulos et al (Tzikopoulos *et al.*, 2011), (5) the gray level co-occurrence matrix (GLCM) based texture features proposed by Haralick et al (Haralick *et al.*, 1973), (6) segmented breast area size, and (7) a percentage density (PD) measure (Byng *et al.*, 1994). Since the absolute values of the different features vary greatly in their respective ranges, we normalized the values of each feature to fall between 0 and 1 based on ±2σ(standard deviation) of their original values.

Third, although these image features have been previously investigated and used in different computerized schemes to detect and/or represent mammographic tissue density features by different research groups, to improve classifier robustness and efficiency, we need to reduce the "curse of dimensionality" of our CAD scheme. Hence, we applied a new fast and accurate sequential floating forward selection (SFFS) method (Ververidis and Kotropoulos, 2008) to search for the optimal or effective features by eliminating redundant and irrelevant features, and thus avoid or minimize "overfitting" of the classifier during the training stages. The detailed description of applying this SFFS method to select effective features for developing CAD schemes of mammograms have been reported in our previous study (Tan *et al.*, 2014a). In this study, we established two initial image feature pools. One pool includes bilateral image features computed from CC view images and one includes the features computed from MLO view images. Each pool included 184 mammographic image features (92 computed from the left breast image and 92 computed from the right breast image) and 3 non-image features (namely women's age, family history of breast cancer, and subjectively-rated breast density based on BI-RADS). The SFFS method was applied separately to each feature pool to select small sets of optimal features for the next step of training the artificial neural network (ANN) based classifiers.

Fourth, since mammograms are two-dimensional projection images, the overlapping breast fibro-glandular tissue patterns in the CC and MLO views are often different, which results in image feature differences computed from the two views (Zheng *et al.*, 2006; Wang *et al.*, 2011). Hence, we built a two-stage classification scheme to analyze the bilateral global image features and classify between positive and negative cases. The first stage has two ANNs that were trained separately using the features computed from the bilateral CC and MLO view images. The second stage consists of a subsequent "scoring fusion" ANN that adaptively and optimally combines the classification scores generated by the two CC and MLO based ANNs. Each ANN was trained using a gradient descent based back-propagation

algorithm (Rumelhart *et al.*, 1986). To find the optimal number of hidden nodes to use in the hidden layer of the ANNs, we trained 150 different ANNs (with different numbers of hidden nodes), and selected the network that maximized the AUC on the training subsets. For the CC and MLO view image feature based ANNs that were optimized using the features selected by SFFS, the number of hidden nodes was varied between 2-40; for the "scoring fusion" ANN at the second stage, this range was 2-10. Namely, we opted for a lower range at the second stage of the classifier since it only had 2 input features compared with the features selected by SFFS at the first stage of the classifier. Other parameters used to train the ANN include the number of training iterations (500), training momentum (0.9) and learning rate (0.01). We used a large ratio of the training momentum to the learning rate and a limited number of iterations to maintain classifier robustness and reduce "overfitting." We also used the hyperbolic tangent activation function at the hidden nodes and the linear activation function at the output nodes of all three ANNs, which are the default parameters used in the Matlab-® Neural Network Toolbox.

2.3. Data analysis and CAD performance assessment

We trained and tested each ANN using a ten-fold cross-validation method, whereby the sum of positive (cancer) cases and negative (cancer-free) cases in our dataset were randomly divided into 10 exclusive partitions. Nine partitions were used to train the classifier including three ANNs in each validation cycle using the bilateral image features computed from the CC and MLO view images, respectively. The trained classifier was then applied to the remaining testing partition. For each testing case, the third "scoring fusion" ANN generated an output classification score, whereby a higher score indicates a higher risk or probability of the FFDM examination of interest being positive (depicting malignant lesions on the images). We iteratively repeated this process 10 times using the 10 different combinations of partitions. Thus, each of the positive and negative cases was tested once with a corresponding "scoring fusion" ANN-generated classification score.

To assess the performance of our new CAD scheme to identify the FFDM examinations with high risk of being positive, we used a number of performance assessment indices in this study. First, we computed the areas under a receiver operating characteristic curve (AUC) including the mean and standard deviation of AUC values over the ten folds of the cross-validation experiments. Second, we sorted the classification scores of all testing cases (including both positive and negative cases) in an ascending order and selected five threshold values to segment all cases into five subgroups (or bins) with an approximately equal number of cases within each subgroup. We then computed the adjusted odds ratios (ORs) for all subgroups using a multivariate statistical model. We computed and analyzed an OR increasing trend using a publically-available statistical software package, *R* (*R* version 2.1.1, http://www.r-project.org). Third, we assessed an absolute classification accuracy, as well as a positive predictive value (PPV) and a negative predictive value (NPV), using a confusion matrix that was computed using a threshold of 0.5 on the classification scores. This threshold is a middle point of the classification score range from 0 to 1. All testing results were tabulated and compared.

In addition, we analyzed CAD performance including the non-image/ epidemiology based features and on the different case subgroups within our image dataset, which includes (1) three positive subgroups namely the verified cancer cases, interval cancer cases and high-risk cases, (2) four mammographic density subgroups based on BI-RADS categories. We then also examined our CAD scheme performance (sensitivity levels) at various specificity levels (from 80% to 95%).

3. Results

In the ten-fold cross-validation procedure, the average number of image features selected by the SFFS method was 12.4±4.1 and 9.0±6.3 from the bilateral CC and MLO view image feature pools, respectively. The results also showed that among the different feature categories as discussed in Section II.B, the bilateral differences of breast region size, pixel value based statistical features, and fractal dimension were the commonly-selected input features for ANNs. Figure 2 displays the three corresponding ROC curves obtained using only the image features. The AUC values for classifying between 812 cancer case group and 3 non-cancer case groups including (1) all 1084 non-cancer cases, (2) 618 not-recalled negative cases, and (3) 466 benign cases are (1) 0.707±0.031, (2) 0.682±0.040 and (3) 0.727±0.031, respectively. Only the AUC results of the not-recalled negative and benign cases were significantly different from each other at the 5% significance level ($p = 0.02$).

Among the 3 non-computed image features (or epidemiology based risk factors), only woman's age was a popular feature selected by the SFFS algorithm and added to the ANN input features, while the family breast cancer history and subjectively-rated mammographic density (BI-RADS) were eliminated. Similar to Figure 2, Figure 3 shows three ROC curves after adding women's age as a feature into the ANN classifiers. The corresponding AUC values increased to 0.779±0.025, (2) 0.769±0.024 and (3) 0.793±0.033, respectively. Using the Wilcoxon rank sum test (or Mann-Whitney *U* test), these AUC results are not significantly different from each other at the 5% significance level with *p*-values ranging from 0.121 to 0.345. However, comparing to the use of only image features (Figure 2), the AUC values obtained after adding feature of "age" significantly increased ($p < 1e^{-5}$ using DeLong's test (DeLong *et al.*, 1988) for paired samples) indicating women's age is a strong risk factor, which is consistent with the existing epidemiology based breast cancer risk prediction models (Amir *et al.*, 2010).

Table 1 summarizes the odds ratios (ORs) and corresponding 95% confidence intervals (CIs) computed for the five subgroups (bins) of FFDM examinations. An increasing trend was observed in the CAD-generated classification scores from subgroups 1 to 5. When using the cases in subgroup 1 as a baseline, the computed ORs monotonically increased from 1.0 to 31.55 in subgroups 1 to 5. The slope of a regression trend using the data pairs between the CAD-generated classification scores and the adjusted ORs is significantly different from zero ($p = 0.005$), which demonstrates a positive association of classification scores generated by this global image feature analysis based CAD scheme and an increasing risk probability trend of the FFDM examinations of interest being positive. By excluding woman's age, ORs monotonically increased from 1.0 to 7.31 in subgroups 1 to 5 also with a significantly increasing risk slope ($p = 0.004$).

Table 2 displays a confusion matrix when applying a threshold of 0.5 on the CAD-generated classification scores of all testing. The FFDM examinations with CAD-generated classification scores greater than 0.5 were assigned to the positive ("high risk") case group; otherwise, the examinations were assigned to the negative ("low risk") case group. The result shows that using this criterion the new CAD scheme correctly classified 72.3% (1370 of 1896) of cases in our testing dataset of this study. The classification accuracy was higher in the negative case group, which was 81.4% (882 of 1084), than in the positive case group, which was 60.0% (488 of 812). The computed positive predictive value (PPV) was 0.71 (488 of 690) and the negative predictive value (NPV) was 0.73 (882 of 1206).

Figure 4 displays the ROC curves for classifying the different cancer subgroups within our image dataset, namely (1) 746 verified cancer cases; (2) 39 interval cancer cases; and (3) 27 high-risk precancer tumors. The corresponding AUC results are (1) 0.787±0.011, (2) 0.739±0.035 and (3) 0.542±0.046, respectively. Similar to our previous study (Tan *et al.*, 2014b), the results indicate that our scheme performs better at classifying verified cancers than interval cancers. The classification performance is poorest for high-risk cases which also corresponds with the results of our other previous study (Wang *et al.*, 2011). A possible explanation for this result could be that the high-risk cases are suspect precancer cases that have been detected at an early stage, and thus have not developed fully. Another explanation is the small size of the high-risk cases in our dataset (27), whereby due to the limited number of high-risk cases in our current dataset, we need to conduct more comprehensive studies in the future on this high risk precancer subgroup in order to derive more conclusive results.

Figure 5 shows and compares four ROC curves of applying our CAD scheme to four subgroups of cases divided by BIRADS categories of mammographic density. AUC values are 0.732±0.058, 0.825±0.016, 0.743±0.015, and 0.851±0.046 for four subgroups of BIRADS 1 to 4, respectively. Table 3 displays the sensitivity levels of our CAD scheme at the 80%, 85%, 90%, and 95% specificity levels on the images stratified according to BI-RADS ratings of mammographic density. The results showed that performance of our scheme does not heavily depend on mammographic density or substantially deteriorate as density increases from BIRADS category of 1 to 4. In this limited dataset, sensitivity levels obtained on the density BI-RADS 4 cases exceeded all other cases except at the 95% specificity level, whereby the density BI-RADS 2 cases yielded a slightly better result. The sensitivity levels of the density BI-RADS 3 cases also exceeded that of the density BI-RADS 1 cases for all specificity levels except the 80% specificity level.

4. Discussion

How to develop highly-performing CAD schemes of mammograms and optimally use CAD in clinical screening practice to help radiologists detect more breast cancers at an early stage has been extensively investigated in the last two decades. Despite the fact that a number of commercialized CAD schemes have been installed in many FFDM imaging systems and adopted in clinical practice in a large number of hospitals around the world, using current CAD of mammograms has been disappointing in terms of adding value to help improve the efficacy of screening mammography, in particular to help detect and classify soft tissue

based breast abnormalities. Hence, researchers believe that the task of continuously exploring new approaches to develop and use CAD is still needed (Nishikawa and Gur, 2014). In this study, we investigated a new CAD approach with a number of unique characteristics.

First, previous single-image and lesion-based CAD schemes of mammograms focused on detecting more positive lesions, which may yield detection sensitivity ranging from 60% to 95% based on mammographic density (BIRADS) categories and more than 2 false-positive cues per case (The *et al.*, 2009). The CAD scheme developed and presented in this study is not another lesion-based scheme. It is a new case-based scheme that combines and analyzes bilateral mammographic image features and their differences from all four-view FFDM images into the CAD decision-making process. Without detecting the specific or targeted lesions, the new CAD scheme yielded a case-based classification performance of AUC = 0.779±0.025 and a total classification accuracy rate of 72.3% (including PPV = 0.71 and NPV = 0.73) in this study. A positive association between CAD-generated classification scores and increase of the risk of the FFDM examinations being positive was also identified through the odds ratio analysis (Table 1). Although our previous study that applied a lesion-based CAD scheme to an early subset of the FFDM cases selected from our current FFDM database yielded 75.6% detection sensitivity with 0.32 false-positives per image (Zheng *et al.*, 2012a), the performance of our new case-based CAD scheme is not directly comparable to this and all other existing lesion-based CAD schemes. Our new CAD scheme does not compete with the lesion-based CAD scheme and may provide supplementary information with higher discriminatory information to help develop adaptive cueing method that can improve efficacy of lesion-based CAD cueing as we have demonstrated in the previous study (Wang *et al.*, 2012).

Second, ruling out CAD-cued false-positives without ignoring or discarding the CAD-cued subtle positive lesions that are likely to be missed or overlooked by radiologists has proven to be very difficult in a number of retrospective and prospective studies (Ko *et al.*, 2006; Nishikawa *et al.*, 2006). As a result, different research efforts have been tested to implement an optimal CAD cueing method (Samulski *et al.*, 2010; Moin *et al.*, 2011). The malignant lesions missed in screening mammography can be categorized into two types of lesions: The first type includes the mammography-occult lesions and the second type includes the visually-detectable lesions, but they had been overlooked by the radiologists in originally prospective reading of the screening mammograms. Based on the assumption that the purpose of developing CAD schemes of mammograms is not to detect the first type of mammography-occult lesions, which should or can only be detected using other imaging modalities, such as digital breast tomosynthesis (DBT) or breast magnetic resonance imaging (MRI), using CAD of mammograms should focus on helping radiologists detect more “visually-detectable” cancers. For this purpose, there is a significant advantage of developing a case-based CAD scheme and cueing method. The case-based cueing (with warning signs) can avoid attracting radiologists’ attention to rule out false-positive lesions with the risk of reducing sensitivity of detecting true-positive lesions (Zheng *et al.*, 2001) and thus may avoid increasing the false-positive detections (similar to the interactive cueing approach as proposed by other researchers (Samulski *et al.*, 2010)). In addition, by warning

the “high risk” cases using our case-based CAD cueing approach, the radiologists can pay more attention to read and interpret these cases. Similar to the retrospective review when the truth has been known, the majority of missed or overlooked, but actually visually-detectable positive lesions can be detected. This is the basic assumption and motivation for this study to develop a new case-based CAD scheme.

Third, although a large number of previous studies have been conducted to detect and quantify global mammographic image features (Chang *et al.*, 2002; Tzikopoulos *et al.*, 2011; Byng *et al.*, 1994; Glide-Hurst *et al.*, 2007; Wei *et al.*, 2011), these studies focused on analyzing mammographic image features computed from one image or computing an average value if multiple images were involved. In this study, we developed and optimized a four-view image based CAD scheme that integrated global image features from both CC and MLO views. Our scheme emphasizes on the bilateral image feature difference (not the average) between the left and right breasts, which is important to detect breast abnormalities and predict the risk of the FFDM examination being positive for breast cancer (Tan *et al.*, 2013). From an initial large feature pool containing 92 image features computed from one view image, we applied a fast and accurate SFFS based feature selection method to analyze the correlations and effectiveness of these features for detecting high risk cases. The results showed that although bilateral differences of the majority of these features can contribute in detecting high risk cases, the popular or more robust features selected by the SFFS algorithm are the simple features that correlate well with previous studies, which include bilateral difference of breast size (Scutt *et al.*, 1997) and statistical pixel value distributions (Zheng *et al.*, 2012b). As a result, instead of either using mammographic image features computed from one image or averaging the features computed from multiple images, using and fusing the image features computed from bilateral images and their differences is also a new contribution of this study to develop a four-view global image feature analysis based CAD scheme.

Fourth, we tested the feasibility of improving the case-based CAD performance by adding three popular epidemiology study based breast cancer risk factors namely, women’s age, family history of breast cancer, and breast density (rated by BI-RADS) (Amir *et al.*, 2010), into our initial feature pool. In the experiment, the family breast cancer history and BI-RADS breast density were eliminated by the SFFS algorithm. The results confirmed that compared to mammographic image feature difference, these two popular lifetime risk factors had significantly lower discriminatory power or contribution in predicting the individual’s risk of having breast cancer (Gail and Mai, 2010). However, women’s age was consistently selected as an effective feature by the SFFS algorithm and was fused with other mammographic image features in the ANN classifiers. This supports the fact that short-term breast cancer incidence rises as women’s age increases. The results of this study indicated that identifying other effective non-image feature based risk factors and fusing them with mammographic image features might also have potential to help improve the accuracy of CAD schemes in detecting high-risk positive cases.

Last, comparing to the majority of previously reported CAD studies, performance of our CAD scheme was assessed using a much larger and diverse FFDM image dataset that involves 1896 FFDM examinations (or 7584 images), which may yield more reliable study

results. In our data analysis, we also made a number of interesting observations. (1) Our scheme yielded higher performance at classifying between malignant and recalled negative cases, which indicates that the global background tissue distribution can also provide discriminatory information to reduce false-positive recalls. (2) Performance of our scheme does not depend on the mammographic density (BIRADS 1 to 4) because the scheme uses the bilateral tissue asymmetry information, which does not correlate to the mammographic density assessed from one image (Zheng *et al.*, 2012b).

In summary, we demonstrated a new concept and approach of developing CAD schemes of mammograms, which is based on the detection and analysis of bilateral global mammographic image features. The goal is to cue the FFDM examinations with high risk of being positive for breast cancer. Despite this new approach and encouraging results, this preliminary study also has a number of limitations. First, this is just a technology development study. Although we proposed a new case-based CAD cueing method to warn radiologists on which FFDM examinations have high risks of being positive, whether such a cueing method can actually help increase breast cancer detection yield and/or reduce false-positive recalls of screening FFDM examinations needs to be tested in future observer performance studies. Second, this is a laboratory based retrospective study using an image dataset with an enriched ratio of positive cases, which does not represent the cancer prevalence ratio of FFDM based screening examinations in clinical practice. Hence, the performance and robustness of our CAD scheme also needs to be further validated in future prospective studies. Third, our CAD scheme includes a number of empirically-determined parameters and thresholds, which may not be optimal. More effective optimization and validation methods (i.e., nested cross-validation) should be tested in our future studies. Fourth, this new CAD scheme only included global mammographic image features that have been investigated and used in previous studies. The local or region based bilateral image features and their differences have not been extracted and applied to this CAD scheme. Therefore, the performance of our CAD scheme may not be optimal and more development effort is needed to optimize this new CAD approach and performance in our future studies.

Acknowledgments

This work is supported in part by Grant R01 CA160205 from the National Cancer Institute, National Institutes of Health. The authors would like to acknowledge the support from the Peggy and Charles Stephenson Cancer Center, University of Oklahoma as well.

References

Amir E, Freedman OC, Seruga B, Evans DG. Assessing women at high risk of breast cancer: a review of risk assessment models. J. Natl. Cancer Inst. 2010; 102:680–91. [PubMed: 20427433]

Berg WA, Zhang Z, Lehrer D, Jong RA, Pisano ED, Barr RG, Bohm-Velez M, Mahoney MC, Evans WP 3rd, Larsen LH, Morton MJ, Mendelson EB, Farria DM, Cormack JB, Marques HS, Adams A, Yeh NM, Gabrielli G, Investigators A. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. JAMA : the journal of the American Medical Association. 2012; 307:1394–404.

Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, Mandelblatt JS, Yakovlev AY, Habbema JD, Feuer EJ, Cancer I and Surveillance Modeling Network C. Effect of screening and adjuvant therapy on mortality from breast cancer. The New England journal of medicine. 2005; 353:1784–92. [PubMed: 16251534]

Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology. 2001; 219:192–202. [PubMed: 11274556]

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]

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]

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]

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]

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]

Fenton JJ, Wheeler J, Carney PA, et al. Reality check: perceived versus actual performance of community mammographers. AJR Am J Roentgenol. 2006; 187:42–6. [PubMed: 16794153]

Gail MH, Mai PL. Comparing breast cancer risk assessment models. J. Natl. Cancer Inst. 2010; 102:665–8. [PubMed: 20427429]

Glide-Hurst CK, Duric N, Littrup P. A new method for quantitative analysis of mammographic density. Med Phys. 2007; 34:4491–8. [PubMed: 18072514]

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]

Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. IEEE Trans. Syst. Man, Cybern. 1973; 3:610–21.

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]

Kallenberg M, Karssemeijer N. Computer-aided detection of masses in full-field digital mammography using screen-film mammograms for training. Phys Med Biol. 2008; 53:6879–91. [PubMed: 19001703]

Ko JM, Nicholas MJ, Mendel JB, Slanetz PJ. Prospective assessment of computer-aided detection in interpretation of screening mammography. AJR Am J Roentgenol. 2006; 187:1483–91. [PubMed: 17114541]

Kriege M, Brekelmans CT, Boetes C, Besnard PE, Zonderland HM, Obdeijn IM, Manoliu RA, Kok T, Peterse H, Tilanus-Linthorst MM, Muller SH, Meijer S, Oosterwijk JC, Beex LV, Tollenaar RA, de Koning HJ, Rutgers EJ, Klijn JG, Magnetic Resonance Imaging Screening Study G. Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. The New England journal of medicine. 2004; 351:427–37. [PubMed: 15282350]

Laya MB, Larson EB, Taplin SH, White E. Effect of estrogen replacement therapy on the specificity and sensitivity of screening mammography. J Natl Cancer Inst. 1996; 88:643–9. [PubMed: 8627640]

Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, White E. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst. 2000; 92:1081–7. [PubMed: 10880551]

Moin P, Deshpande R, Sayre J, Messer E, Gupte S, Romsdahl H, Hasegawa A, Liu BJ. An observer study for a computer-aided reading protocol (CARP) in the screening environment for digital mammography. Acad Radiol. 2011; 18:1420–9. [PubMed: 21971259]

Nishikawa, RM.; Edwards, A.; Schmidt, RA.; Papaioannou, J.; Linver, MN. Can radiologists recognize that a computer has identified cancers that they have overlooked?. In: Jiang, Y.; Eckstein, MP., editors. SPIE Medical Imaging. SPIE; San Diego, CA: 2006. p. 614601-8.

Nishikawa RM, Gur D. CADe for Early Detection of Breast Cancer—Current Status and Why We Need to Continue to Explore New Approaches. Academic Radiology. 2014; 21:1320–1. [PubMed: 25086951]

Peer PG, Verbeek AL, Straatman H, Hendriks JH, Holland R. Age-specific sensitivities of mammographic screening for breast cancer. Breast cancer research and treatment. 1996; 38:153–60. [PubMed: 8861833]

Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323:533–6.

Samulski M, Hupse R, Boetes C, Mus RD, den Heeten GJ, Karssemeijer N. Using computer-aided detection in mammography as a decision support. Eur Radiol. 2010; 20:2323–30. [PubMed: 20532890]

Scutt D, Manning JT, Whitehouse GH, Leinster SJ, Massey CP. The relationship between breast asymmetry, breast size and the occurrence of breast cancer. Br J Radiol. 1997; 70:1017–21. [PubMed: 9404205]

Sickles, E.; D'Orsi, C.; Bassett, L., et al. ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. American College of Radiology; Reston, VA: 2013. ACR BI-RADS® Mammography.

Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA Cancer J. Clin. 2013; 63:11–30. [PubMed: 23335087]

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]

Tan M, Pu J, Zheng B. A new and fast image feature selection method for developing an optimal mammographic mass detection scheme. Med. Phys. 2014a; 41:081906. [PubMed: 25086537]

Tan M, Pu J, Zheng B. Reduction of false-positive recalls using a computerized mammographic image feature analysis scheme. Phys. Med. Biol. 2014b; 59:4357–73. [PubMed: 25029964]

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]

Ververidis D, Kotropoulos C. Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. Signal Process. 2008; 88:2956–70.

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]

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. 2011; 260:42–9. [PubMed: 21406634]

Zheng B, Ganott MA, Britton CA, Hakim CM, Hardesty LA, Chang TS, Rockette HE, Gur D. Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings. Radiology. 2001; 221:633–40. [PubMed: 11719657]

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. 2012a; 85:e153–e61. [PubMed: 21343322]

Zheng B, Sumkin JH, Zuley ML, Wang X, Klym AH, Gur D. Bilateral mammographic density asymmetry and breast cancer risk: a preliminary assessment. Eur. J. Radiol. 2012b; 81:3222–8. [PubMed: 22579527]

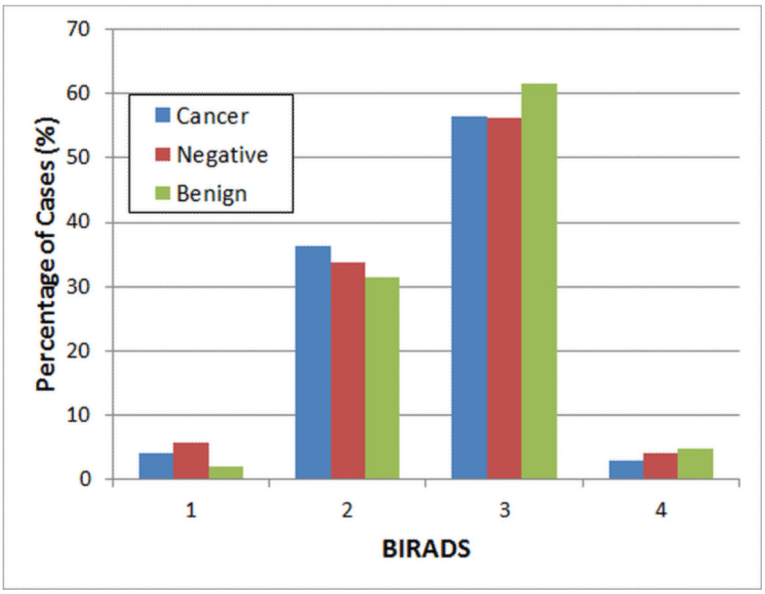

Figure 1.

Histogram distribution of mammographic density (BI-RADS) ratings in three groups of positive, negative and benign cases.

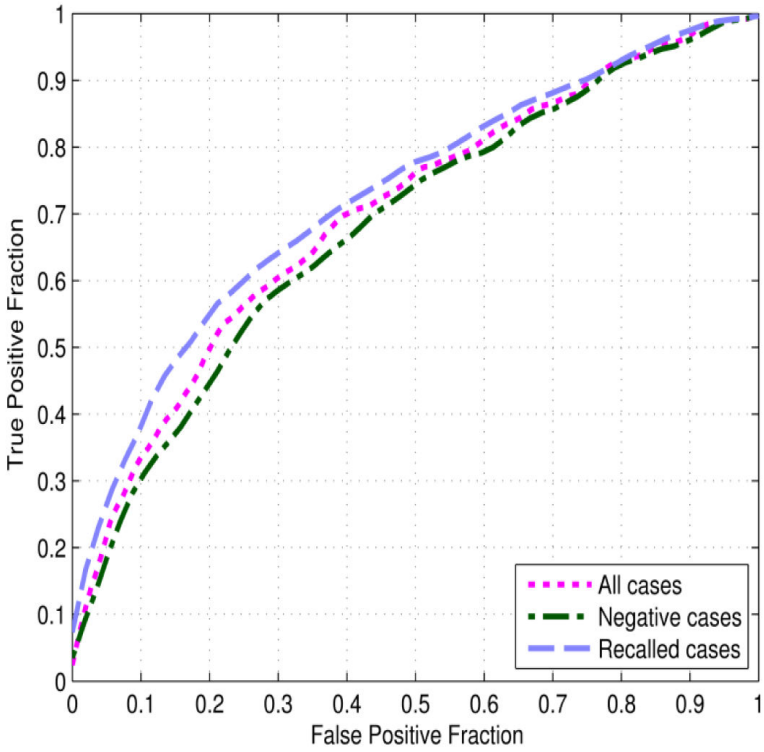

Figure 2.

Comparison of three ROC curves of applying our CAD scheme using image features only to classify between positive and three negative case subgroups including (1) all negative, recalled and benign cases, (2) only negative cases and (3) only benign and recalled negative/benign cases.

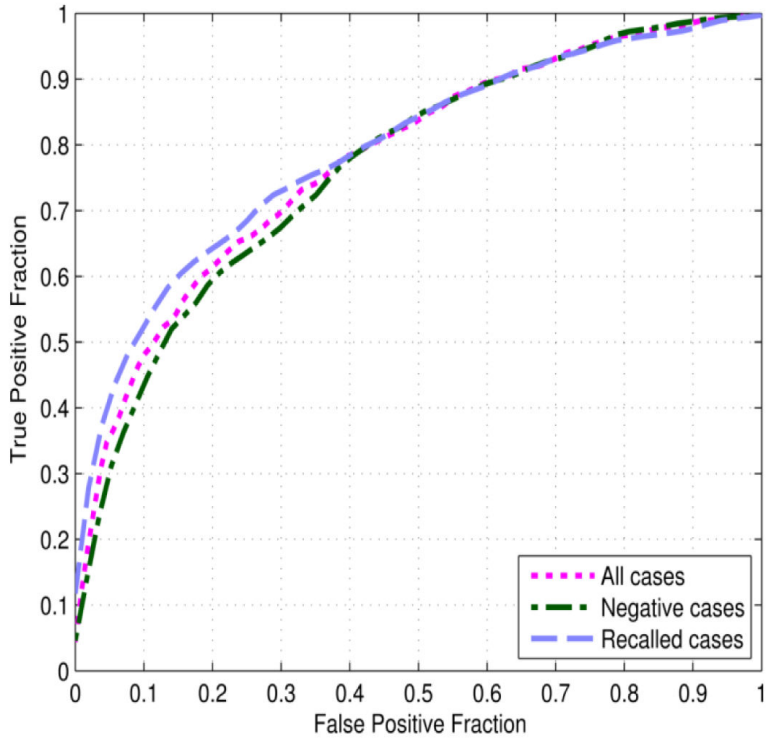

Figure 3.

Comparison of three ROC curves of applying our CAD scheme mixing image features and woman's age to classify between positive and three negative case subgroups including (1) all negative, recalled and benign cases, (2) only negative cases and (3) only benign and recalled negative/ benign cases.

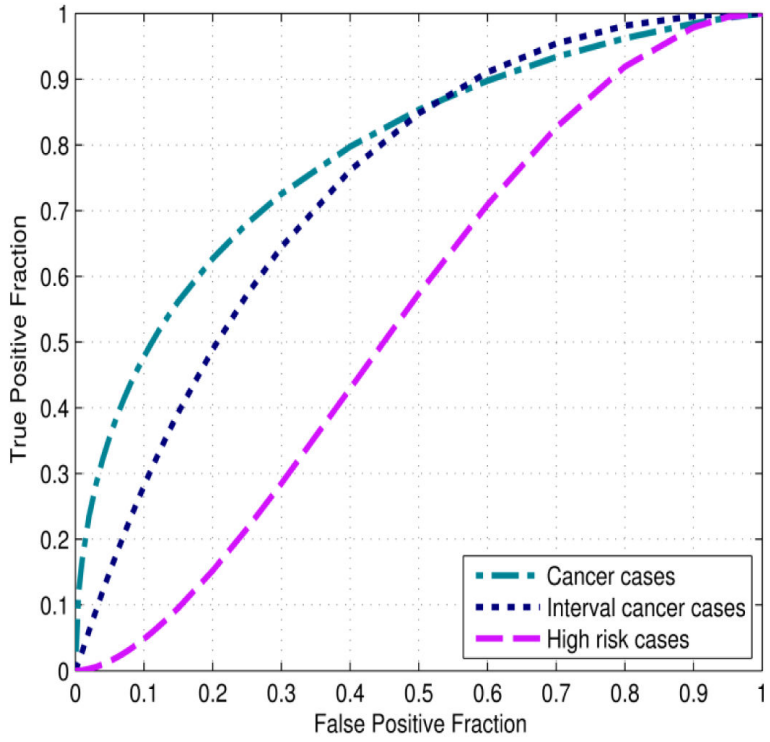

Figure 4.

Comparison of three ROC curves of applying our CAD scheme to classify different cancer subgroups within our image dataset, namely (1) 746 verified cancer cases, (2) 39 interval cancer cases and (3) 27 high-risk precancer tumors.

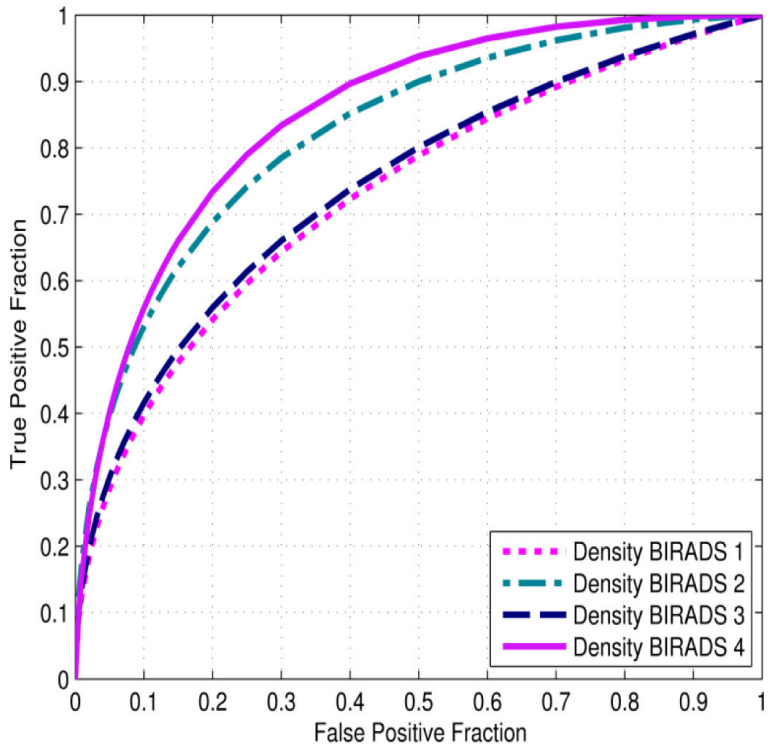

Figure 5.

Four ROC curves of applying our CAD scheme to four subgroups of cases rated in four different BIRADS categories of mammographic density.

Table 1

Summary of the adjusted odds ratios (ORs) and 95% confidence intervals (CIs) at five subgroups (bins) of the probability scores generated by our "scoring fusion" classifier.

Epidemiology based Risk Factors Included	Subgroup	Number of Cases (Positive – Negative)	Adjusted Odds Ratio (OR)	95% Confidence Interval (CI)
Yes	1	53 – 326	1.00	baseline
	2	103 – 276	2.30	[1.59, 3.32]
	3	137 – 242	3.48	[2.43, 4.98]
	4	201 – 178	6.95	[4.88, 9.89]
	5	318 – 62	31.55	[21.19, 46.96]
No	1	90 – 289	1.00	baseline
	2	117 – 262	1.43	[1.04, 1.98]
	3	136 – 243	1.80	[1.31, 2.47]
	4	205 – 174	3.78	[2.77, 5.16]
	5	264 – 116	7.31	[5.30, 10.08]

Table 2

A confusion matrix obtained when applying a threshold of 0.5 on CAD-generated classification scores.

Actual ↓	Negative cases	Positive cases
Negative cases	882	202
Positive (cancer) cases	324	488

Table 3

Sensitivity levels of our CAD scheme at four specificity levels by stratifying the testing cases according to the four mammographic density BIRADS rating categories.

Specificity	95%	90%	85%	80%
Density BI-RADS 1	28.7%	39.7%	47.7%	54.1%
Density BI-RADS 2	39.6%	53.1%	62.1%	68.8%
Density BI-RADS 3	30.4%	41.6%	49.6%	56.0%
Density BI-RADS 4	40.3%	55.9%	66.0%	73.4%