HHS Public Access

Author manuscript

IEEE Trans Med Imaging. Author manuscript; available in PMC 2018 May 01.

Published in final edited form as:

IEEE Trans Med Imaging. 2017 May ; 36(5): 1162–1171. doi:10.1109/TMI.2017.2654799.

Estimating the Accuracy Level Among Individual Detections in Clustered Microcalcifications

María V. Sainz de Cea[1], **Robert M. Nishikawa**[2], and **Yongyi Yang**[1]

[1]Medical Imaging Research Center, Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA

[2]Department of Radiology, University of Pittsburgh, Pittsburgh, PA 15231, USA

Abstract

Computerized detection of clustered microcalcifications (MCs) in mammograms often suffers from the occurrence of false positives (FPs), which can vary greatly from case to case. We investigate how to apply statistical estimation to determine the number of FPs that are present in a detected MC lesion. First, we describe the number of true positives (TPs) by a Poisson-Binomial probability distribution, wherein a logistic regression model is trained to determine the probability for an individual detected MC to be a TP based on its detector output. Afterward, we model the spatial occurrence of FPs in a lesion area by a spatial point process (SPP), of which the distribution parameters are estimated from the detections in the lesion and its surrounding region. Furthermore, to improve the estimation accuracy, we incorporate the Poisson-Binomial distribution of the number of TPs into the SPP model by using maximum *a posteriori* (MAP) estimation. In the experiments, we demonstrated the proposed approach on the detection results from a set of 188 full-field digital mammography images (95 cases) by three existing MC detectors. The results demonstrated that there was a strong consistency between the estimated and the actual number of TPs (or FPs) for these detectors. When the fraction of FPs in detection was varied from 20% to 50%, both the mean and median values of the estimation error were within 11% of the total number of detected MCs in a lesion. In particular, when the number of FPs increased to as high as 11.38 in a cluster on average, the error was 2.51 in the estimated number of FPs. In addition, lesions estimated to be more accurate in detection were shown to have better classification accuracy (for being malignant or benign) than those estimated to be less accurate.

Index Terms

Computer-aided diagnosis (CAD); clustered microcalcifications (MCs); false positives in detection; spatial point process; mammography

I. Introduction

Breast cancer is the most commonly diagnosed cancer, apart from skin cancer, among women in the US, accounting for nearly one in every three cancer cases. Breast cancer is

Correspondence to: Yongyi Yang.

also the second leading cause of cancer deaths among women after lung cancer [1]. Screening mammography is one of the most widely used methods for early breast cancer diagnosis. It is typically performed on asymptomatic women, and is reported to double the number of early-stage cancer cases that are diagnosed each year in the US [2].

Studies show that clustered microcalcifications (MCs) can be a sign of non-palpable breast cancer [3], and are present in 30–50% of patients diagnosed at early stage [4]. MCs are tiny calcium deposits that appear as bright spots in mammograms (Fig. 1). Although MCs are often seen, they are difficult to diagnose accurately. Only 10% to 40% of biopsies for evaluating MCs are ultimately malignant [5]. For this reason, there are a number of studies to determine the potential relationship between MC features and pathology [6]–[8].

In the literature, there have been great interests in developing computerized methods to aid the diagnosis of MC lesions, e.g. [9], [10]. These methods are collectively known as computer-aided diagnosis (CAD). Generally speaking, there are two distinct tasks in a CAD system for MC lesions. One task is to determine whether clustered MCs are present or not in a mammogram under consideration (called CADe). The purpose is to alert radiologists to potential lesions for further evaluation. The other task is to determine whether a detected MC lesion is malignant or benign (called CADx). In CADx, the individual MCs in a lesion region are first located either by a human or by a CADe detector; afterward, the detected MCs are quantified by a set of image features, which is subsequently classified as malignant or benign by a pattern classifier. Conceivably, the detection accuracy of the individual MCs can directly affect the accuracy of the classification outcome, because the features extracted from the detected MCs will be impacted [11].

In practice, the utility of a CAD system is often compromised by the occurrence of false-positives (FPs) in detection. MCs are typically very small, and can vary greatly in shape and size; they can be low contrast, and may even be hardly separable from their surrounding dense tissue [12]. Studies show that there can be many factors contributing to FPs in MC detection [13], which include MC-like noise patterns, imaging artifacts, linear structures such as milk ducts, etc. This has led to great efforts in developing CADe methods for improving the accuracy in MC detection (by increasing sensitivity and reducing FP rate) [12], [14]–[23].

In spite of these efforts, a major challenge facing MC detection algorithms is the great inter-patient variability in mammogram characteristics. For example, it is more difficult to accurately detect MCs in dense breasts [24] or in young women [12]. In the literature, the detection performance of an MC detector is typically reported in terms of how accurately the presence of an MC cluster (instead of individual MCs therein) is correctly detected in a mammogram [25], [23]. This is because the aim of a CADe system is mainly to alert radiologists of potential lesions (i.e., a detected MC cluster) for further evaluation. As a result, the accuracy of individual MCs in a detected cluster is not directly assessed in a CADe system. Conceivably, the latter is further subject to intra-patient variability, because individual MCs within a lesion can even vary greatly.

To illustrate this, in Fig. 2 we show a box-car plot of the sensitivity level achieved by an MC detector (the DoG detector in [16]) over a set of 200 MC lesions (Section III-D); a common decision threshold was used for the different lesions and the average sensitivity was 70%. We note that a wide range of variation in sensitivity occurred among the lesions, with 25% of the lesions having a sensitivity above 0.848 and 25% having a sensitivity below 0.563. Moreover, in Fig. 2 we also show a box-car plot of the fraction of FPs (FPF) among the individual detections in each detected cluster. Note that there was also a wide range of variation in the FPF among the lesions, with 25% of the lesions having an FPF above 0.558 and 25% with an FPF below 0.25.

Given the large variabilities among lesions in detecting individual MCs (as observed above), it is expected that the accuracy of a detected lesion by a CAD system is inevitably subject to large case-to-case variations. Thus, it is important to determine how accurate the detections are in a detected MC cluster. For example, in a CADe system, knowing the level of FPs among the detections can provide important information on the confidence of the alert generated by the system. Similarly, in a CADx system, the accuracy of a CADx classifier can be adversely affected by the presence of high levels of FPs (as to be seen in the results in Section IV-D). Since the purpose of both CADe and CADx systems is to assist human readers, providing such information on the accuracy of detections can potentially prompt the reader to more closely examine the validity of the CAD output when a detected lesion is in question.

In this work, we aim to develop a framework to assess the level of accuracy among individual MCs in a detected cluster. For this purpose, we estimate the number of FPs (or, equivalently, the number of true positives (TPs)) in a cluster. Specifically, suppose that there are n objects (i.e., both FPs and TPs) detected. We want to determine among them how many are FPs and how many are TPs. We propose two different approaches for this problem. In the first approach, we use a Poisson-Binomial probability model to describe the number of TPs based on the probability for each detected object to be a TP (or FP); for the latter, we use a logistic regression model, which is based on the characteristics of the detector output of a detected object. In the second approach, we use a spatial point process (SPP) to characterize the occurrence of FPs in a lesion region, and apply statistical estimation (maximum likelihood (ML) or maximum *a posteriori* (MAP)) to determine the number of FPs.

The proposed framework is general and expected to be applicable to different MC detectors. In this study we demonstrated its performance with three existing detectors. The first two have been well cited in the literature: the difference of Gaussian (DoG) in [16] detector, which is an example of image enhancement detectors, and the support vector machine (SVM) detector in [22] (an example of machine learning detectors). The third detector is a context-sensitive MC detector published recently in [26]. As to be seen in the results (Section IV), these three detectors differed in terms of their detection accuracy levels. Thus, they serve as a good test bed for the proposed framework.

To the best of our knowledge, there is no previous work reported in the literature which directly deals with how to determine the detection accuracy on a case by case basis. While

our application is for MC detection, the proposed framework is expected to be applicable for other similar detection problems, such as cell detection in microscopic images [27]–[29].

The rest of the paper is organized as follows: In Section II, we present the methodology for determining the accuracy of a detected MC cluster and explain the different methods used for this purpose. In Section III, we describe the approach used for evaluating the proposed framework. We present our results in Section IV, and provide discussions in Section V. Finally, we provide conclusions in Section VI.

II. Methodology

A. Problem formulation and overview

Consider a lesion region $\mathscr{A}$ which contains a number of clustered MCs in a mammogram image. Assume that, for identifying the individual MCs, an MC detector $f(\cdot)$ has been applied to the image region such that a set of detected objects (i.e., potentially MCs) is obtained in $\mathscr{A}$. We want to estimate the number of true MCs among the set of detections. For convenience, let's denote each detected object i by a feature vector $\mathbf{x}^{(i)}$ (to be defined subsequently), $i = 1, \ldots, n$, where n is the number of detected objects, and let $y^{(i)}$ denote its unknown label (1 for being a true MC, and 0 otherwise). Then, the number of true MCs among all the detections is given by

$$M = \sum_{i=1}^{n} y^{(i)}. \quad (1)$$

Our goal is to estimate the value of M.

Note that the number of false positives (FPs) among the detections is given by $n - M$. Thus, from this point on we refer to the estimation of the number of TPs indistinctly from that of FPs.

We considered two approaches for estimating M in (1), namely a probabilistic model approach, and a spatial point process (SPP) estimation approach, as outlined below:

1. *Probabilistic model approach.* We employed a logistic regression model [30] to estimate the probability for each detection $\mathbf{x}^{(i)}$ being a TP or a FP (i.e., $y^{(i)} = 1$ or 0) based on the characteristics of the detector output at $\mathbf{x}^{(i)}$. Afterward, the probability distribution (a Poisson-Binomial distribution) was derived for the number of TPs (i.e. M) among all detections.
2. *Spatial point process (SPP) modeling approach.* We modeled the occurrence of FPs in a lesion region by a 2D stochastic point process [31]. To account for the noise characteristics of each case, we made use of a reference region in the immediate vicinity of the lesion, in which we assumed only FPs can occur. We used this reference region together with the lesion region to estimate both the rate parameter of FPs and the number of TPs. We considered two different approaches for estimating these parameters: i) maximum likelihood (ML)

estimation, and ii) maximum *a posteriori* (MAP) estimation. For the latter, we used the Poisson-Binomial distribution approach derived above as a prior for the number of TPs.

B. Probabilistic model approach

Consider a detected object i in lesion region $\mathscr{A}$. We aim to estimate the probability for it to be a TP (i.e., $y^{(i)} = 1$) based on the characteristics in its detector output (described by feature vector $\mathbf{x}^{(i)}$). In this study, we estimated this probability via a logistic regression model [30] as

$$p_1(\mathbf{x}^{(i)}) \triangleq p(y^{(i)}=1|\mathbf{x}^{(i)})=\frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}^{(i)}+b)}} \quad (2)$$

where $\mathbf{w}$ and b are parameters to be determined through supervised learning (as described subsequently). Note that the probability for $\mathbf{x}^{(i)}$ to be an FP (i.e., $y^{(i)} = 0$) is given by $1 - p_1(\mathbf{x}^{(i)})$.

Now consider all the detections $\mathbf{x}^{(i)}$, $i = 1, \ldots, n$, in $\mathscr{A}$. With the probability specified for each detection as in (2), we can derive the probability distribution for the number of TPs among the detections, i.e., M in (1). Indeed, M follows a Poisson-Binomial distribution [32], which is given by

$$p(M=k)=\sum_{A\in F_k}\prod_{i\in A}p_1(\mathbf{x}^{(i)})\prod_{j\in A^c}(1-p_1(\mathbf{x}^{(j)}) \quad (3)$$

where F_k is the collection of all possible subsets formed from k distinct integers in $\{1, 2, \ldots, n\}$, and A^c is the complement of A, i.e. $A^c = \{1, 2, \ldots, n\}\backslash A$.

With the distribution defined in (3), we can derive either a confidence interval estimate or a point estimate for M. In this study, we considered the latter. In particular, we used the mode of the distribution as an estimator for M. That is,

$$\hat{M}=\arg\max_k p(M=k). \quad (4)$$

Alternatively, we may also estimate M by using its statistical mean, which is given by

$$\mu=\sum_{i=1}^{n}p_1(\mathbf{x}^{(i)}). \quad (5)$$

We note that the estimator in (5) is slightly simpler computationally than the mode estimator in (4), while the latter always yields an integer value. To simplify the numerical evaluation of the Poisson-Binomial distribution in (3), in our implementation, we used a well-known Poisson approximation [32] for its computation:

$$p(M=k) \approx \frac{\mu^k e^{-\mu}}{k!}, k=0,1,\ldots,n \quad (6)$$

where n is the total number of detections and μ is given by (5).

In our experiments we found that the above two estimators could yield similar results. We will report results mainly for the mode estimator in (4). We note that the distribution in (3) is also to be used as a prior for M in the MAP estimation approach later in Section II-C2.

Finally, for characterizing the properties in the detector output of a detected object i, we used a small window centered around the detected object. This is out of consideration that MCs are typically limited in size (0.1 to 1.0 mm in diameter) [12] and that their spatial extent is further reduced in the domain of detector output. In our experiment, we used a 3×3 window in the detector output to form a 9-dimensional vector $\mathbf{x}^{(i)}$ for each detected object (image resolution = 100 μm/pixel). Alternatively, one may consider using other features, such as image contrast, size, shape, etc. But such features will require segmenting the candidates, which will vary with the segmentation method used.

To determine the parameters $\mathbf{w}$, b for the logistic regression model in (2), we used a set of training samples $\mathcal{T} = \{(\mathbf{x}^{(j)}, y^{(j)}), j = 1, \cdots, m\}$. These training samples $\mathbf{x}^{(j)}$ were obtained from the detections by the MC detector in use on a set of training mammograms (Section III-D) in which the MCs were known (for defining the labels $y^{(j)}$). During the training step, the parameters $\mathbf{w}$, b are determined through minimizing the following cost function [30]:

$$J(\mathbf{w},b) = -\sum_{j=1}^{m} y^j \log p_1(\mathbf{x}^{(j)}) - (1-y^{(j)}) \log(1-p_1(\mathbf{x}^{(j)})) \quad (7)$$

C. Spatial point process (SPP) model approach

A spatial point process (SPP) is used to describe the random distribution pattern of a set of points in a d-dimensional space ($d = 2$ in our application) [33]. After applying an MC detector to a given lesion region $\mathcal{A}$, neither the locations of the detected FPs nor their number of occurrences are known. To accommodate their random nature, we modeled the spatial distribution of the FPs within $\mathcal{A}$ by a Poisson point process [34]. Such a process is characterized by the following two properties: i) for a bounded spatial region B, the number of points contained within B follows a Poisson distribution of which the mean is proportional to the area of B; and ii) for two disjoint regions B_1 and B_2, the number of points within B_1 is independent of that within B_2.

A challenge, however, is that for a given lesion region $\mathscr{A}$ both FPs and TPs can occur simultaneously, but neither of the two is known. To facilitate modeling the random occurrence of FPs in $\mathscr{A}$, we introduced a reference region $\mathscr{R}$, which is located in the immediate vicinity of $\mathscr{A}$ from the same mammogram, as illustrated in Fig. 3. This reference region is assumed to have the following properties: i) it does not contain any true MCs; hence, only FPs will be detected by the MC detector in $\mathscr{R}$; ii) the image noise characteristics in $\mathscr{R}$ are similar to those in the lesion region $\mathscr{A}$; thus, the occurrence of FPs follows a common random process in the two regions.

To quantify the random process of FPs, we applied the same MC detector to $\mathscr{A}$ and $\mathscr{R}$. Let N_r denote the number of resulting detections in $\mathscr{R}$. Then N_r obeys the following Poisson distribution:

$$p(N_r|\lambda)=\frac{(\lambda A_r)^{N_r}}{N_r!}e^{-(\lambda A_r)} \quad (8)$$

where A_r denotes the area of $\mathscr{R}$, and λ denotes the rate parameter which is the expected number of FPs per unit area.

On the other hand, for the lesion region $\mathscr{A}$, let M be the number of detected MCs as in (1), and N_l the total number of detections. Then $N_l - M$ is the number of FPs, and N_l obeys the following distribution:

$$p(N_l|\lambda, M)=\frac{(\lambda A_l)^{N_l-M}}{(N_l-M)!}e^{-(\lambda A_l)} \quad (9)$$

where A_l denotes the area of lesion region $\mathscr{A}$.

Given that the two regions $\mathscr{A}$ and $\mathscr{R}$ are non-overlapping, the joint distribution of the detections in the two regions can be written as

$$p(N_r, N_l|\lambda, M)=\frac{\lambda^{N_r+N_l-M}A_r^{N_r}A_l^{N_l-M}}{N_r!(N_l-M)!}e^{-\lambda(A_r+A_l)}. \quad (10)$$

Our goal then becomes to estimate the parameters λ (rate of FPs) and M (the number of TPs) in (10). Below we describe two different approaches for this estimation.

1) Maximum likelihood (ML) estimation—Given the observations N_r and N_l, we used maximum likelihood (ML) estimation [35] to estimate λ and M. That is,

$$[\hat{\lambda}, \hat{M}]=\arg\max_{\lambda,M}\log p(N_r, N_l|\lambda, M). \quad (11)$$

In our experiment, we used the interior-point algorithm [36] for the optimization problem in (11).

2) Maximum a posteriori (MAP) estimation—In this approach, we take advantage of the Poisson-Binomial distribution $p(M)$ derived earlier in (3) on the number of TPs. For this purpose, we seek a solution for M based on maximum *a posteriori* (MAP) estimation. That is,

$$[\hat{\lambda}, \hat{M}] = \arg\max_{\lambda, M} [\log p(N_r, N_l | \lambda, M) + \log p(M)]. \quad (12)$$

It is noted that a uniform prior is used for the rate parameter λ above.

The MAP estimate above takes into account both the spatial distribution properties of FPs (as in ML) and the detector output properties (as in the probabilistic model approach).

III. Performance evaluation

A. Estimation accuracy of TPs/FPs in detection

To evaluate the estimation performance, for a given lesion with n detected objects, we compared the estimated number of TPs (or FPs) among these n detections against the actual number (i.e., the marked MCs). For quantifying the estimation accuracy, we used the following relative error:

$$e_P = \frac{|M - \hat{M}|}{n} \quad (13)$$

where M is the actual number of TPs and $\hat{M}$ is the estimated number.

Note that the fraction of TPs (TPF) among the n detections is given by $\frac{M}{n}$. Thus, the relative error e_P simply corresponds to the difference between the actual TPF and the estimated TPF in the detection results of a given lesion.

Furthermore, while the relative error e_P is defined in terms of the number of TPs (i.e., M) in (13), it can be shown that it can also be used equally for measuring the accuracy of the estimated number of FPs. That is, the relative error e_P also corresponds to the difference between the actual fraction of FPs (FPF) and the estimated FPF among the detected objects.

To summarize the estimation accuracy over all the lesions in the dataset, we report several statistics on e_P, including the mean, median, first and third quartiles.

B. Classification of cases with different estimated accuracy levels in detection

As an indirect validation on the estimated accuracy on the detections in an MC lesion, we applied a CADx classifier to discriminate the lesion being malignant or benign based on the

detected MCs (which are subject to FPs). It is reasonable to expect that a lesion with fewer FPs (i.e., lower FPF) is more likely to be correctly classified than one with many FPs, and its performance will be closer to the human marked.

To demonstrate this, we divided the lesions in the dataset into two equal sized groups based on the estimated FPF values in their detected MCs, with group A having the lowest FPF values and group B having the highest FPF values. Afterward, the CADx classifier is applied to the cases within each group, and the classification accuracy is calculated accordingly. To summarize the classification accuracy, we used the area under the receiver operating characteristic (ROC) curve (AUC).

For the CADx classifier, we made use of a support vector machine (SVM) classifier previously developed in [37]. The decision function of this classifier is given by

$$f(\mathbf{x})=\sum_{k=1}^{N_s} w_k K(\mathbf{x},\mathbf{s}_k)+b \quad (14)$$

where $\mathbf{s}_k$, $k = 1, \cdots, N_s$ are the support vectors, w_k and b are the model parameters, all of which are determined from training.

In (14), the Gaussian radial basis function (RBF) kernel is used, i.e.,

$$K(\mathbf{x}_i,\mathbf{x}_j)=exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}\right) \quad (15)$$

where $\sigma > 0$ is the kernel width.

For characterizing each lesion, we used the same set of features as in [37]; these features were selected to have intuitive meanings that are consistent with image features interpreted by radiologists [38]. Specifically, they are as follows: 1) the number of MCs in the cluster; 2) the mean effective volume (area times effective thickness) of individual MCs; 3) the area of the cluster; 4) the circularity of the cluster; 5) the relative standard deviation of the effective thickness; 6) the relative standard deviation of the effective volume; 7) the mean area of MCs; and 8) the second highest MC-shape-irregularity measure. The specific details about these features can be found in [38].

The SVM classifier in (14) was pre-trained on a dataset of 376 screen-film mammogram images from 222 subjects (118 benign, 104 cancer). For the SVM classifier, the following parameters were used: $\sigma = 10$ and $C = 100$ (C is a regularization parameter used to control the trade-off between the model complexity and empirical risk in model training).

C. MC detectors for demonstration

To demonstrate the proposed approach for estimating the accuracy in MC detection, we considered three existing MC detectors, among which the first two have been well cited in the literature and the third one was published recently [26]. More specifically, the first detector is an SVM detector developed in [22], which is based on supervised learning. The second one is the DoG detector [16], which is of low computational complexity for MC detection. The third detector is a context-sensitive MC detector, which is designed to suppress FPs in MC detection. As an illustration, in Fig. 4 we show the output of these three detectors when applied to the two ROIs shown earlier in Fig. 1. As can be seen, the MCs are notably enhanced in the output of the detectors. However, the noise patterns are quite different among the detectors, and there are also numerous bright spots which would be falsely detected depending on the operating threshold used.

For MC detection in each lesion image, the output of the detector in use was compared against an operating threshold T. The detected pixels in the output were then grouped (with 8-neighbor connection) into objects. Those objects smaller than 3 pixels in size were discarded in order to reduce spurious detections. A detected object was treated as a TP when it was less than 0.3 mm away from a marked MC or at least 40% of its area overlaped with that of an MC; otherwise it was counted as an FP.

To fully assess the performance of the proposed approach at different FP levels, in our experiments the operating threshold T was adjusted such that the sensitivity level in detection was varied over a wide range (from 60% to 85%). Afterward, the proposed approach was applied to the detection results obtained by the three detectors at different sensitivity levels. In view that when FPF is above 50% there are more FPs than TPs in a detected lesion, we tested these detectors for FPF range up to approximately 50%.

D. Mammogram dataset

For this study we made use of a set of full-field digital mammography (FFDM) images collected by the Department of Radiology at the University of Chicago under IRB approval. The image set consisted of 188 images from 95 cases (43 malignant, 52 benign), all containing clustered MCs. They were acquired using a Senographe 2000D FFDM system (General Electric Medical Systems; Milwaukee, WI) with a spatial resolution of 100 μm/pixel. Most of the cases had both craniocaudal and mediolateral oblique views. The MCs in each mammogram were manually identified by a researcher with more than 15 years of experience in mammography research and with special training on interpreting mammograms. In total, there were 8,979 MCs marked in these 188 mammograms.

For the purpose of evaluating the accuracy in the detected MCs, the lesion regions of clustered MCs in these mammograms were marked out by a bounding circle with a diameter of 1 cm, 2 cm, or 3 cm (according to the lesion size), so that all the marked MCs were contained inside the circle; for those elongated lesions, an ellipse of equal area was used in place of the bounding circle. For the spatial point process model approach, a reference region is needed for each lesion. In our experiment, this region was set to be the immediate

annulus region outside the lesion of which the area is equivalent to that of a disk of 3 cm in diameter. For illustration, an example is shown in Fig. 3 for a lesion and its reference region.

In the probabilistic model approach, a training set is needed for the logistic regression model as in (2). For this purpose, we adopted a 2-fold cross validation procedure [39] in which the cases in the dataset were randomly divided into two subsets so that the cases were used independently either for training or for testing, but never both. The test results from each fold were aggregated together to obtain the estimation error e_P [40]. The model was trained separately for each detector at the highest sensitivity (75% for DoG and SVM, and 85% for the context-sensitive detector).

IV. Results

For convenience, below the probabilistic approach is referred to as PBD (for Poisson-Binomial distribution), and the spatial point process model approach is referred to as SPP-ML when ML estimate is used or as SPP-MAP when MAP estimate is used. To avoid confusion, we present the results separately for the three MC detectors. For each detector, we report the estimation results at four different operating points with the fraction of FPs in detection ranging approximately from 20% to 50%.

A. Estimation of the number of TPs/FPs in SVM detector

In Fig. 5(a) we show a scatter plot of the estimation results obtained by the PBD method for all the lesions in the dataset when the detector sensitivity level is at 70%. Similarly, the estimation results obtained by SPP-ML and SPP-MAP are shown in Figs. 5(b) and 5(c), respectively. In these plots, each point corresponds to a lesion, of which the y coordinate is the estimated number of TPs ($\hat{M}$) and the x coordinate is the actual number of TPs (M). Thus, a point on the 45° line represents a perfect match between $\hat{M}$ and M. Note that a logarithmic scale is used in the plot in order to accommodate the large range in the number of TPs.

From Fig. 5 it is observed that there is good agreement between the actual and the estimated results obtained by all three methods. Quantitatively, the correlation coefficients between $\hat{M}$ and M were 0.9684 for PBD, 0.9467 for SPP-ML, and 0.9707 for SPP-MAP. Furthermore, a paired t-test comparison revealed no statistically significant difference between the mean values of $\hat{M}$ and M, with p-value = 0.7731 for PBD, 0.8523 for SPP-ML, and 0.9731 for SPP-MAP. These results indicate that there is no systematic bias in the estimates by the three methods.

To further quantify the accuracy of the estimation results in Fig. 5, in Table I (2nd row) we show a summary of the estimation error e_P obtained by the three methods. For each method, the mean, median, first and third quartiles of e_P over all the lesions in the dataset are given. The median values of e_P were 0.1007 for PBD, 0.1429 for SPP-ML, and 0.0833 for SPP-MAP. The SPP-MAP approach is noted to be more accurate than SPP-ML (p-value $< 10^{-4}$, bootstrapping test with 20,000 samples) and PBD (p-value=0.0185).

We also tested the estimation methods when the MC detector was set at different sensitivity levels. For brevity, we summarize these results in Table I; as an indication of the FP levels, the corresponding mean FPF is also shown for each sensitivity level. For example, for sensitivity at level 65%, the mean values of e_P were 0.1216, 0.1627, and 0.1053 for PBD, SPP-ML, and SPP-MAP, respectively. As observed above, these results indicate that SSP-MAP is more accurate than both PDB and SPP-ML, and PDB is more accurate than SPP-ML.

Moreover, in all three methods, the median value of e_P was smaller than its corresponding mean value at all sensitivity levels. This indicates that the error distribution is skewed toward the left (i.e., lesions having smaller estimation errors).

B. Estimation of the number of TPs/FPs in DoG detector

In Table II we show a summary of the estimation results when the DoG detector was used for MC detection on the lesions in the dataset. As in Table I above, the mean, median, first and third quartiles of the estimation error e_P are given for each of the three methods, namely PBD, SPP-ML, and SPP-MAP. In this case, given the higher FPF, the detection sensitivity was varied at a sensitivity range from 60% to 75%, and the estimation results are given for each level.

From Table II, it can been seen that at sensitivity level 60% the mean value of e_P was 0.0822 for SPP-MAP, compared to 0.0983 for PBD (p-value=0.0174) and 0.1474 for and SPP-ML (p-value $< 10^{-4}$), respectively. Similarly, the median value of e_P was 0.0526 for SPP-MAP, compared to 0.0714 for PBD, and 0.0952 for SPP-ML. On the other hand, at sensitivity level 75%, the mean value of e_P was 0.0925 for PBD, compared to 0.1697 for SPP-ML and 0.0965 for SPP-MAP; a statistical comparison yielded no difference between PBD and SPP-MAP (p-value=0.6568).

C. Estimation of the number of TPs/FPs in context-sensitive MC detector

In Table III we summarize the estimation results obtained for the context-sensitive MC detector. Given the much lower FPF in this detector, the sensitivity levels are shown from 70% to 85%.

As in the results above, the results in Table III indicate that SSP-MAP is more accurate than both PDB and SPP-ML. For example, at sensitivity 80%, the median values of e_P were 0.1256 for PBD, 0.1408 for SPP-ML, and 0.1030 for SPP-MAP. The SPP-MAP approach is noted to be more accurate than SPP-ML (p-value $< 10^{-4}$) and PBD (p-value=0.0033).

D. CADx accuracy vs. estimated detection accuracy

As described in Section III-B, in this experiment we tested whether the estimated detection accuracy of the MCs in a lesion can correlate with its classification accuracy by a CADx classifier. For this purpose, we divided the lesions into two equal sized groups, with group A having the lesions with the lowest estimated FPF values, and group B having the rest. Below we show the results obtained with the SPP-MAP method when detection sensitivity was set at 70% for the SVM and DoG detectors and 80% for the context-sensitive detector. We

chose these sensitivity levels in order to achieve high sensitivity in detection while keeping FPF below 50%. For the SVM detector, the estimated average FPFs were 0.1927 for group A (actual 0.2263) and 0.4470 for group B (actual 0.4316). For the DoG detector, the average estimated FPFs were 0.2313 for group A (actual 0.2456) and 0.5296 for group B (actual 0.5173). For the context-sensitive detector, the average estimated FPFs were 0.1876 for group A (actual 0.2096) and 0.5723 for group B (actual 0.6283). While similar results could be obtained at other sensitivity levels and also for PBD and SPP-ML methods, they are not shown here in the interest of space.

In Table IV we show the classification results on the two groups when the SVM detector was used, where the AUC is given for each group. As reference, the AUC values are also given in Table IV when the human marked MCs were used by the CADx classifier. As can be seen, for group B the AUC of the detected MCs was notably lower than that of the human marked MCs, where the difference in AUC was 0.1881 (denoted by dAUC_B); in comparison, for group A the AUC value of the detected MCs was much closer to that of the human marked, where the difference in AUC was 0.0243 (denoted by dAUC_A). To quantify the statistical difference in AUC deviations between the two groups (i.e., dAUC_A vs. dAUC_B), we conducted a randomized permutation test [41], which yielded a p-value of 0.0257 for the observed AUC differences between the two groups. In the permutation test, all the lesions in group A_p were randomly permuted with those in group B_p (each lesion with probability 0.5), based on which the null distribution for $\mathrm{dAUC}_B - \mathrm{dAUC}_A$ was obtained. A total of 20,000 permutations were used.

Similarly, we show in Table V the classification results on the two groups when the DoG detector was used. In this case, the differences in AUC between human marked and detected were −0.0196 for group A and 0.2156 for group B. A permutation test yielded a p-value of 0.0027 on such differences.

In Table VI we show similar results for the context-sensitive detector. As can be seen, for group B the AUC of the detected MCs are notably lower than that of the human marked MCs, where the difference was AUC is 0.1433; in comparison, for group A the AUC value of the detected MCs was much closer to that of the human marked, where the difference in AUC was −0.0216 (p-value = 0.0236).

V. Discussions

A. Estimation accuracy vs level of FPs

From Fig. 5, the estimation accuracy is observed to vary with the number of TPs in a lesion in all three methods. In particular, the estimation error tends to be larger for lesions with fewer TPs. This is more related to the number of detections in the cluster. To further examine this, we divided the lesions in the dataset into three groups: group one having lesions with 10 or fewer detections (76 lesions), group two having lesions with between 11 and 30 detections (89 lesions), and group three having lesions with 31 or more detections (35 lesions). We then calculated the estimation error e_P for the lesions in each of the three groups. In Fig. 6 we show a box-car plot of the distribution of e_P within each group obtained by PBD, SPP-ML, and SPP-MAP, respectively. As can be seen, the error e_P decreased from

group 1 to group 3. For example, for PBD, the median e_P were 0.1429 in group one, 0.1062 in group two, and 0.0889 in group three. Moreover, the within-group variations also decreased from group one to group three.

We believe that the above observation can be explained as follows: in both PBD and SPP-ML we need to estimate the statistical distribution (i.e., the Poisson-Binomial distribution or the spatial point process) underlying each method based on the detected objects from a lesion area. Among the three groups above, the number of detections increased from group 1 to group 3, and so did the number of FPs, with the average being 2.13 in group 1, 6.5 in group 2, and 17.65 in group 3. Thus, the estimates of the distribution parameters in both PBD and SPP-ML are expected to become more accurate as more detected samples become available from group 1 to group 3. This in turn can help improve the estimation accuracy even though the number of FPs increased.

The same reasoning above can also be used to explain the estimation results at different sensitivity levels by the three methods. Take PBD in Table I for example. The average error e_P varied only slightly in the range between 0.1175 and 0.1284 as the sensitivity level was increased from 65% to 80% (FPF increased correspondingly from 0.2316 to 0.5441). For example, at sensitivity level of 75%, the average number of detections in a cluster is 25.51, among which 11.38 are FPs. The corresponding average estimation errors on the number of FPs (or TPs) were 2.685 by PBD and 2.515 by SPP-MAP.

B. Estimation accuracy of different methods

From the results in Tables I, and II and III it is observed that the three estimation methods achieved different results in performance. We believe this is due to the underlying differences in the information utilized by the different methods. Specifically, the PBD approach directly makes use of the detector output values on the individual detections. In contrast, the SPP-ML approach is based on only the spatial distribution of detections in a reference region aside from the lesion region. As a result, the latter is expected to be supplementary, as it does not directly make use of the detector output which reflects the strength of a detected signal. However, with SPP-MAP, it combines the information from both sources, and thus, the estimation accuracy is expected to improve over both PBD and SPP-ML. This improvement is noted in particular when the detection sensitivity is low (at which there are fewer detections). For example, in Table I, at sensitivity 65%, the average error e_P was improved to 0.1053 in SPP-MAP, compared to 0.1216 for PBD and 0.1627 for SPP-ML.

C. Ground truth in evaluation

In this study we quantified the estimation accuracy by the different methods based on the human marked MCs. In the literature, human marked MCs are routinely used as the ground truth in development of MC detection algorithms (e.g. [42]–[45]). However, human marked MCs are inevitably subject to errors associated with inter- and intra-observer variations. As a result, the obtained error level e_P could be affected by the errors associated with the marked MCs. Nevertheless, the classification results in Tables IV, V, and VI show that the human marked MCs consistently yielded higher AUC values than the detected in group B (which

were estimated to have more FPs). This indicates that the human marked MCs are more accurate than those detected for classifying a lesion being malignant or benign.

VI. Conclusion

In this study, we investigated how to determine the accuracy of a detected MC lesion by estimating the number of FPs (and TPs) present among the detected MCs. We developed two different approaches for this purpose: one is to derive a probability model for each detection to be a TP (or FP) based on the detector output; the number of TPs among the detections in a lesion is then described by a Poisson-Binomial distribution. The second approach is to model the occurrence of FPs in a lesion area by a spatial point process (SPP) for which the parameters are estimated based on the detections within both the lesion and its surrounding region. For the latter approach we also applied MAP estimation in which the Poisson-Binomial distribution from the first approach is incorporated into the SPP model so that both detector output and the spatial occurrence information can be utilized. We demonstrated these estimation methods with three different MC detectors, namely SVM detector, DoG detector and context-sensitive classification model detector, on a set of 188 FFDM images from 95 cases. The results showed that the estimated number of FPs (or TPs) can be fairly accurate when compared to their actual number in each lesion in the dataset. Moreover, as a possible application of the accuracy estimation method we also demonstrated that the classification on a case being malignant or benign can be more accurate when it is estimated to have lower FPs. This indicates that the proposed estimation methods can be useful not only for assessing the accuracy of the detected MCs in a lesion when human marking is not available, but also for providing a confidence measure on the output of a CADx system based on the level of FPs present. In future work, we plan to investigate how to improve the classification performance of a CADx system by exploiting the estimated detection accuracy in a lesion.

Acknowledgments

This work was supported in part by NIH/NIBIB under grant R01EB009905.

References

1. DeSantis CE, Fedewa SA, Goding Sauer A, Kramer JL, Smith RA, Jemal A. Breast cancer statistics, 2015: Convergence of incidence rates between black and white women. CA: A Cancer Journal for Clinicians. 2015
2. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. New England Journal of Medicine. 2012; 367(21):1998–2005. [PubMed: 23171096]
3. Sickles EA. Mammographic features of early breast cancer. American Journal of Roentgenology. 1984; 143(3):461–464. [PubMed: 6331721]
4. Chan HP, Doi K, Vyborny CJ, Lam KL, Schmidt RA. Computer-aided detection of microcalcifications in mammograms methodology and preliminary clinical study. Investigative Radiology. 1988; 23(9):664–671. [PubMed: 3182213]
5. Fondrinier E, Lorimier G, Guerin-Boblet V, Bertrand AF, Mayras C, Dauver N. Breast microcalcifications: Multivariate analysis of radiologicand clinical factors for carcinoma. World Journal of Surgery. 2002; 26(3):290–296. [PubMed: 11865363]

6. Kettritz U, Morack G, Decker T. Stereotactic vacuum-assisted breast biopsies in 500 women with microcalcifications: radiological and pathological correlations. European Journal of Radiology. 2005; 55(2):270–276. [PubMed: 16036159]
7. Karamouzis MV, Likaki-Karatza E, Ravazoula P, Badra FA, Koukouras D, Tzorakoleftherakis E, Papavassiliou AG, Kalofonos HP. Non-palpable breast carcinomas: Correlation of mammographically detected malignant-appearing microcalcifications and molecular prognostic factors. International Journal of Cancer. 2002; 102(1):86–90. [PubMed: 12353238]
8. Nakayama R, Uchiyama Y, Watanabe R, Katsuragawa S, Namba K, Doi K. Computer-aided diagnosis scheme for histological classification of clustered microcalcifications on magnification mammograms. Medical Physics. 2004; 31(4):789–799. [PubMed: 15124996]
9. Mousa R, Munib Q, Moussa A. Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural. Expert Systems with Applications. 2005; 28(4):713–723.
10. Wei L, Yang Y, Nishikawa RM. Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis. Pattern Recognition. 2009; 42(6):1126–1132. [PubMed: 20161326]
11. Kallergi M. Computer-aided diagnosis of mammographic microcalcification clusters. Medical physics. 2004; 31(2):314–326. [PubMed: 15000617]
12. Cheng HD, Cai X, Chen X, Hu L, Lou X. Computer-aided detection and classification of microcalcifications in mammograms: a survey. Pattern Recognition. 2003; 36(12):2967–2991.
13. Ema T, Doi K, Nishikawa RM, Jiang Y, Papaioannou J. Image feature analysis and computer-aided diagnosis in mammography: Reduction of false-positive clustered microcalcifications using local edge-gradient analysis. Medical Physics. 1995; 22(2):161–169. [PubMed: 7565347]
14. Rangayyan RM, Ayres FJ, Desautels JL. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. Journal of the Franklin Institute. 2007; 344(3):312–348.
15. 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. Journal of the National Cancer Institute. 2004; 96(3):185–190. [PubMed: 14759985]
16. Dengler J, Behrens S, Desaga JF. Segmentation of microcalcifications in mammograms. IEEE Transactions on Medical Imaging. 1993; 12(4):634–642. [PubMed: 18218457]
17. Chan HP, Vyborny CJ, MacMahon H, Metz CE, Doi K, Sickles EA. Digital mammography: ROC studies of the effects of pixel size and unsharp-mask filtering on the detection of subtle microcalcifications. Investigative Radiology. 1987; 22(7):581–589. [PubMed: 3623862]
18. Dhawan AP, Buelloni G, Gordon R. Enhancement of mammographic features by optimal adaptive neighborhood image processing. IEEE Transactions on Medical Imaging. 1986; 5(1):8–15. [PubMed: 18243977]
19. Chen CH, Lee GG. On digital mammogram segmentation and microcalcification detection using multiresolution wavelet analysis. Graphical Models and Image Processing. 1997; 59(5):349–364.
20. Jing H, Yang Y, Nishikawa RM. Detection of clustered microcalcifications using spatial point process modeling. Physics in Medicine and Biology. 2011; 56(1):1–17. [PubMed: 21119233]
21. Karssemeijer N. Stochastic model for automated detection of calcifications in digital mammograms. Image and Vision Computing. 1992; 10(6):369–375.
22. El-Naqa I, Yang Y, Wernick MN, Galatsanos NP, Nishikawa RM. A support vector machine approach for detection of micro-calcifications. IEEE Transactions on Medical Imaging. 2002; 21(12):1552–1563. [PubMed: 12588039]
23. Chan HP, Lo SCB, Sahiner B, Lam KL, Helvie MA. Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network. Medical Physics. 1995; 22(10):1555–1567. [PubMed: 8551980]
24. Morrow WM, Paranjape RB, Rangayyan RM, Desautels JEL. Region-based contrast enhancement of mammograms. IEEE Transactions on Medical Imaging. 1992; 11(3):392–406. [PubMed: 18222882]
25. Nishikawa RM, Giger M, Doi K, Vyborny C, Schmidt R. Computer-aided detection of clustered microcalcifications on digital mammograms. Medical and Biological Engineering and Computing. 1995; 33(2):174–178. [PubMed: 7643656]

26. Wang J, Nishikawa RM, Yang Y. Improving the accuracy in detection of clustered microcalcifications with a context-sensitive classification model. Medical Physics. 2016; 43(1): 159–170. [PubMed: 26745908]
27. Shitong W, Min W. A new detection algorithm (nda) based on fuzzy cellular neural networks for white blood cell detection. IEEE Transactions on information technology in biomedicine. 2006; 10(1):5–10. [PubMed: 16445244]
28. Maslov AY, Barone TA, Plunkett RJ, Pruitt SC. Neural stem cell detection, characterization, and age-related changes in the subventricular zone of mice. The Journal of neuroscience. 2004; 24(7): 1726–1733. [PubMed: 14973255]
29. Dong, B., Shao, L., Da Costa, M., Bandmann, O., Frangi, AF. Deep learning for automatic cell detection in wide-field microscopy zebrafish images. 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI); IEEE; 2015. p. 772-776.
30. Hosmer, DW., Lemeshow, S. Introduction to the logistic regression model. Applied Logistic Regression. 2. Wiley Online Library; 2000. p. 1-30.
31. Miles RE. On the homogeneous planar poisson point process. Mathematical Biosciences. 1970; 6:85–127.
32. Hong Y. On computing the distribution function for the poisson binomial distribution. Computational Statistics & Data Analysis. 2013; 59:41–51.
33. Møller J, Waagepetersen RP. Modern statistics for spatial point processes. Scandinavian Journal of Statistics. 2007; 34(4):643–684.
34. Baddeley, A., Bárány, I., Schneider, R. Spatial point processes and their applications. Stochastic Geometry: Lectures given at the CIME Summer School; Martina Franca, Italy. September 13–18, 2004; 2007. p. 1-75.
35. Scholz F. Maximum likelihood estimation. Encyclopedia of Statistical Sciences. 1985
36. Wright, SJ. Primal-dual interior-point methods. Siam; 1997.
37. Wei L, Yang Y, Nishikawa RM, Jiang Y. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. IEEE Transactions on Medical Imaging. 2005; 24(3):371–380. [PubMed: 15754987]
38. Jiang Y, Nishikawa RM, Wolverton DE, Metz CE, Giger ML, Schmidt RA, Vyborny CJ, Doi K. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology. 1996; 198(3):671–678. [PubMed: 8628853]
39. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician. 1983; 37(1):36–48.
40. Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter. 2010; 12(1):49–57.
41. Good, P. Permutation tests: a practical guide to resampling methods for testing hypotheses. Springer Science & Business Media; 2013.
42. Cheng HD, Lui YM, Freimanis RI. A novel approach to microcalcification detection using fuzzy logic technique. IEEE Transactions on Medical Imaging. 1998; 17(3):442–450. [PubMed: 9735907]
43. Freer TW, Ulissey MJ. Screening mammography with computeraided detection: prospective study of 12,860 patients in a community breast center 1. Radiology. 2001; 220(3):781–786. [PubMed: 11526282]
44. Balakumaran T, Vennila I, Shankar CG. Detection of microcalcification in mammograms using wavelet transform and fuzzy shell clustering. 2010 arXiv preprint arXiv:1002.2182.
45. Yao, C., Yang, Y., Chen, H., Jing, T., Hao, X., Bi, H. Adaptive kernel learning for detection of clustered microcalcifications in mammograms. IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI); IEEE; 2012. p. 5-8.

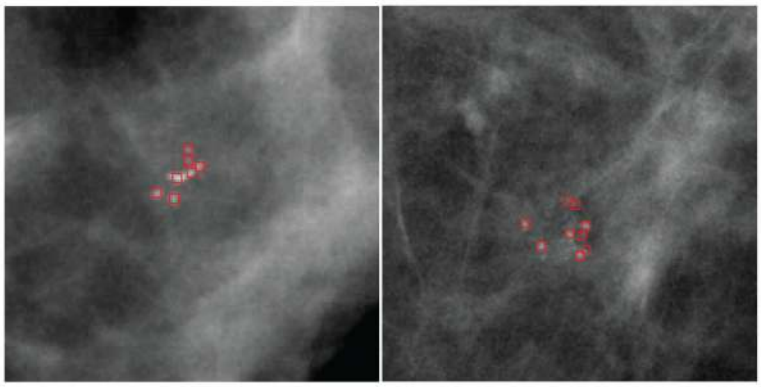

Fig. 1.
Two mammogram ROIs with clustered MCs (marked by red square symbols).

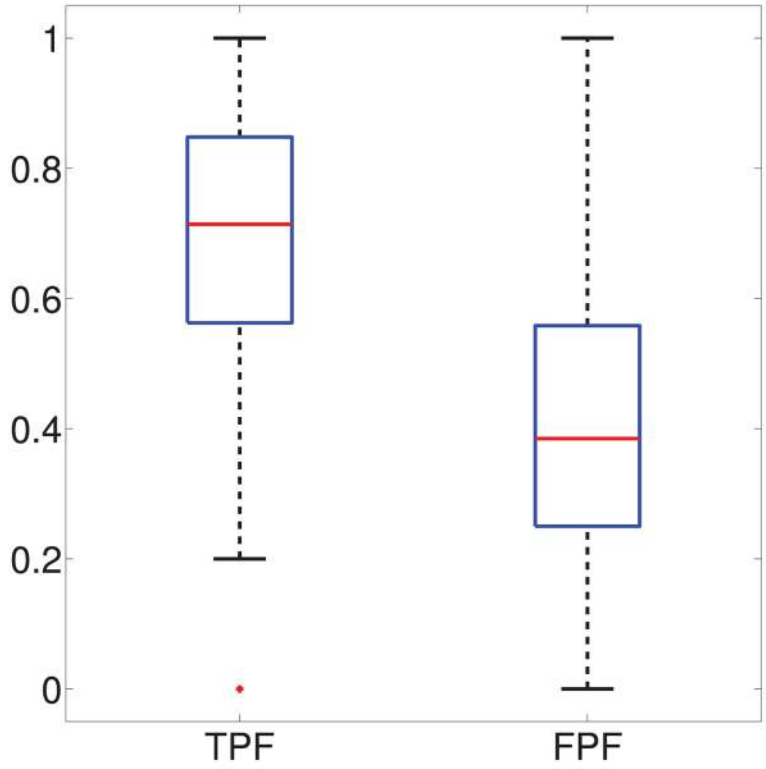

Fig. 2.

Box-car plots of true positive fraction (TPF) and false positive fraction (FPF) in detection over 200 MC lesions by the DoG detector with mean sensitivity 70%.

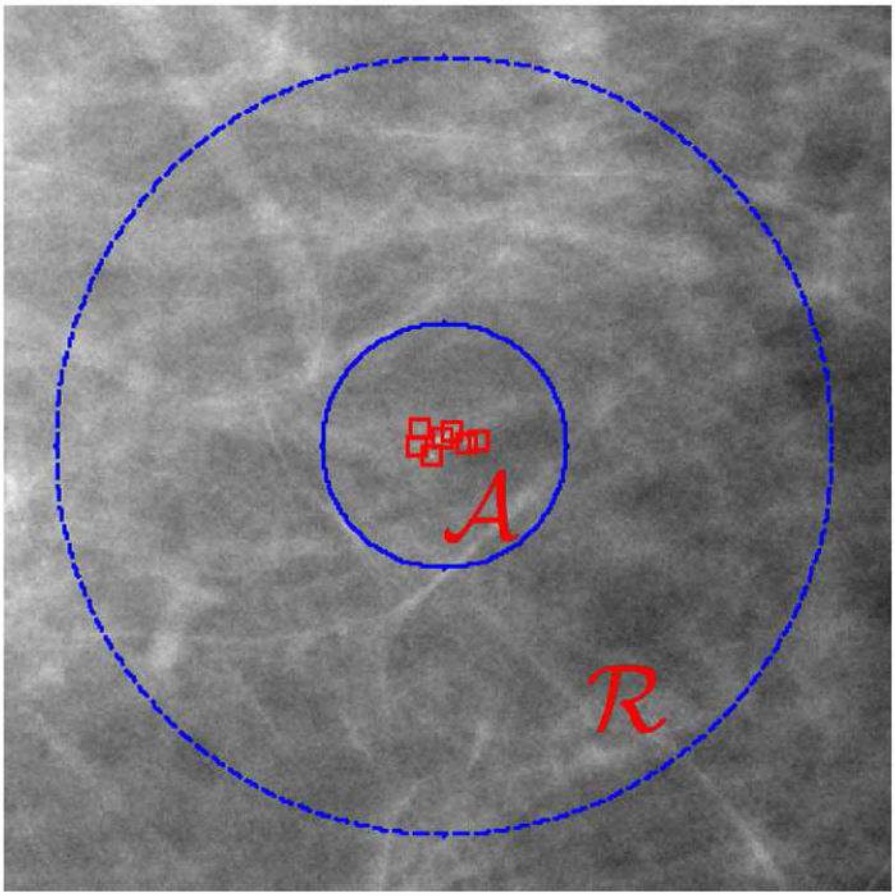

Fig. 3.
Illustration of a lesion region $\mathcal{A}$ (diameter 1 cm) and its reference region $\mathcal{R}$. There are 7 MCs in the lesion.

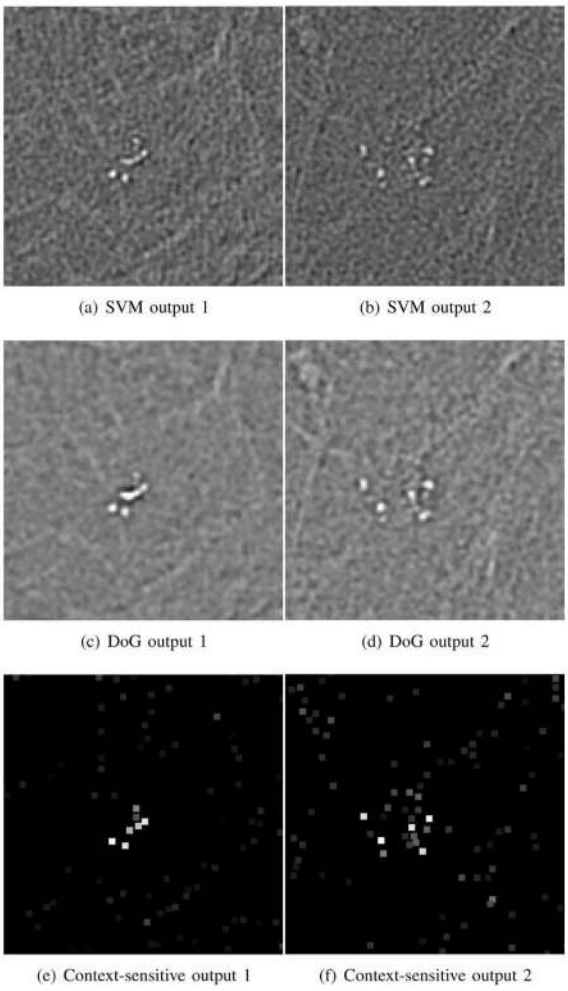

Fig. 4.

Output by SVM, DoG and contex-sensitive detectors for the two example ROIs (1 and 2) in Fig. 1.

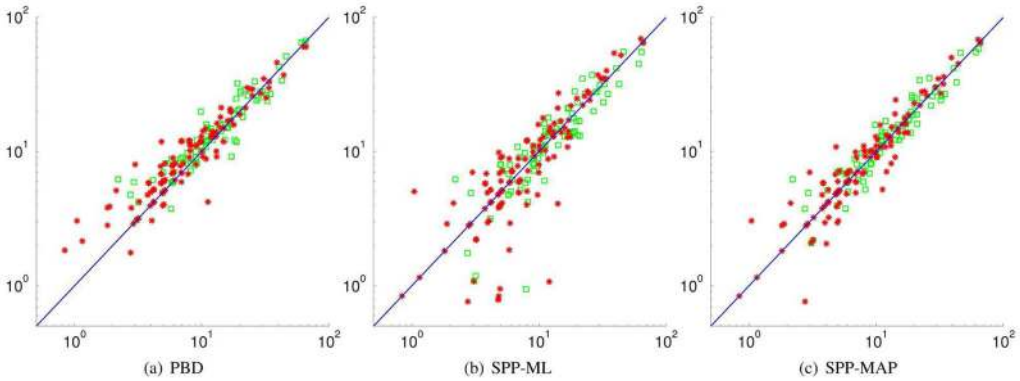

Fig. 5.

Scatter plots of actual (x-axis) vs estimated (y-axis) number of TPs by PBD, SPP-ML, and SPP-MAP for the SVM detector at TPF = 70%. Malignant cases are indicated by red * symbols and benign cases by green square symbols.

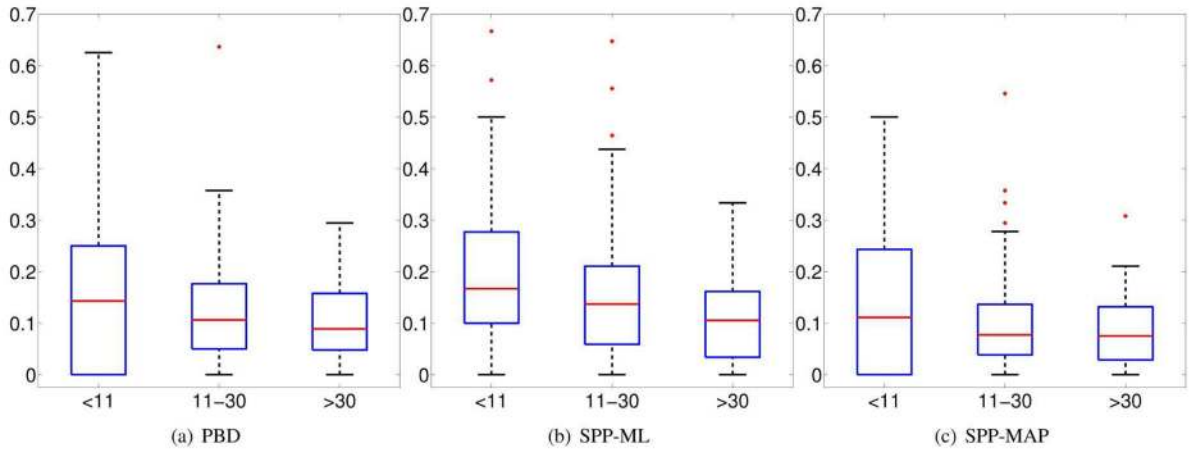

Fig. 6.

Box-car plots of the relative estimation error obtained by PBD, SPP-ML, and SPP-MAP for all the cases with detection sensitivity = 70%. Lesions in Group 1 have 10 or fewer detections, those in Group 2 have detections between 11 and 30, and those in Group 3 have more than 30 detections.

TABLE I

Relative estimation error e_P for SVM detector at different sensitivity levels

Sensitivity	Mean FPF	Error e_P	PBD	SPP-ML	SPP-MAP
65%	0.2316	Mean	0.1216	0.1627	0.1053
		Median	0.1044	0.1396	0.0880
		Quartile 1	0	0	0
		Quartile 3	0.1733	0.2436	0.1429
70%	0.3314	Mean	0.1175	0.1641	0.1103
		Median	0.1007	0.1429	0.0833
		Quartile 1	0.0149	0.0588	0.0177
		Quartile 3	0.1667	0.2222	0.1667
75%	0.4260	Mean	0.1176	0.1622	0.1064
		Median	0.1049	0.1429	0.0909
		Quartile 1	0.0513	0.0714	0.0432
		Quartile 3	0.1667	0.2308	0.1559
80%	0.5441	Mean	0.1284	0.1571	0.1170
		Median	0.11449	0.1250	0.1062
		Quartile 1	0.0541	0.0625	0.0509
		Quartile 3	0.1898	0.2290	0.1667

IEEE Trans Med Imaging. Author manuscript; available in PMC 2018 May 01.

TABLE II

Relative estimation error e_P for DOG detector at different sensitivity levels

Sensitivity	Mean FPF	Error e_P	PBD	SPP-ML	SPP-MAP
60%	0.2005	Mean	0.0983	0.1474	0.0822
		Median	0.0714	0.0952	0.0526
		Quartile 1	0	0	0
		Quartile 3	0.1504	0.2222	0.1280
65%	0.2779	Mean	0.1017	0.1741	0.0985
		Median	0.0845	0.1250	0.0678
		Quartile 1	0	0.0476	0
		Quartile 3	0.1450	0.2583	0.1455
70%	0.3937	Mean	0.0958	0.1724	0.1011
		Median	0.0817	0.1539	0.0909
		Quartile 1	0.0075	0.0588	0.0072
		Quartile 3	0.1429	0.2500	0.1519
75%	0.5094	Mean	0.0925	0.1697	0.0965
		Median	0.0833	0.1437	0.0833
		Quartile 1	0.0318	0.0679	0.0364
		Quartile 3	0.1250	0.2500	0.1333

TABLE III

Relative estimation error e_P for context-sensitive detector at different sensitivity levels

Sensitivity	Mean FPF	Error e_P	PBD	SPP-ML	SPP-MAP
70%	0.1991	Mean	0.1304	0.1217	0.0923
		Median	0.1111	0.0909	0.0714
		Quartile 1	0.0281	0	0
		Quartile 3	0.2000	0.2000	0.1429
75%	0.2666	Mean	0.1303	0.1362	0.09570
		Median	0.1194	0.1111	0.0833
		Quartile 1	0.0455	0	0
		Quartile 3	0.1818	0.2000	0.1474
80%	0.4110	Mean	0.1256	0.1408	0.1030
		Median	0.1111	0.1250	0.0909
		Quartile 1	0.0439	0.0625	0.0335
		Quartile 3	0.1836	0.2145	0.1428
85%	0.5510	Mean	0.1125	0.1297	0.0927
		Median	0.1052	0.1101	0.0833
		Quartile 1	0.0420	0.0469	0.0357
		Quartile 3	0.1528	0.1818	0.1295

IEEE Trans Med Imaging. Author manuscript; available in PMC 2018 May 01.

TABLE IV

AUC values for two groups with different estimated FPF levels with SVM detector

	A	***B***
Detected MCs	0.7270	0.5860
Marked MCs	0.7513	0.7741
dAUC	0.0243	0.1881

TABLE V

AUC values for two groups with different estimated FPF levels with DOG detector

	A	***B***
Detected MCs	0.7447	0.5626
Marked MCs	0.7251	0.7781
dAUC	−0.0196	0.2156

TABLE VI

AUC values for two groups with different estimated FPF levels with context-sensitive detector

	A	***B***
Detected MCs	0.7944	0.6364
Marked MCs	0.7728	0.7797
dAUC	−0.0216	0.1433