Breast Mass Segmentation based on Multi-scale Morphological Sifting and Self-grown Cascaded Random Forests

Hang Min, Shekhar S. Chandra, Stuart Crozier, Andrew P. Bradley

School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia

Abstract

In this paper, we present a mammographic breast mass detection and segmentation system. Breast mass detection and segmentation are challenging due to the fact that breast masses vary in size, shape, margin and contrast. Building classifiers for this task is also difficult since the normal tissue regions often greatly outnumber the abnormal regions, causing a class imbalance in the training set. To face the first challenge, we develop a multi-scale morphological sifting (MMS) approach using linear structuring elements mapped in radial patterns to extract the mass-like objects in mammograms. The MMS paired with a multi-level Otsu thresholding method can segment masses in various shapes, margins and sizes from the background tissue accurately. To tackle the class imbalance problem, we design a self-grown cascaded random forests (CasRFs) which adaptively assembles multiple layers of random forests (RFs) and adopts a probability-ranking based under-sampling method to balance the training set for each RF. Evaluated on two public available datasets, full-digital dataset INbreast and screen-film dataset DoD BCRP, the proposed method achieves an average sensitivity of 0.93 and 0.81 at 1.64 and 3.59 false positives per image respectively. The average Dice similarity index between the segmentation and the ground truth is calculated on INbreast and was found to be 0.85.

Keywords

Breast masses, Mammograms, Detection, Segmentation, Morphological sifting, Random forest

1 Introduction

Breast cancer is one of the leading causes of cancer death among females worldwide (Jemal et al., 2011). Early detection and diagnosis can increase the chances of survival and provides patients with more treatment options (Ganesan et al., 2013). Mammography is the primary modality in breast screening that has been proven to be effective in reducing breast cancer mortality (Andreea et al., 2011). The appearance of breast masses in mammograms is one of the most important signs of development of breast cancer (Schnabel et al., 2013). The detection of breast masses is generally regarded as more challenging compared with other breast abnormalities, such as calcifications, not only due to the large variation in size and shape of breast masses, but also because breast masses can appear in low contrast in mammograms (Oliver et al., 2010). With the development of image processing and machine learning technology, computer aided detection (CAD) has been introduced to breast image interpretation. Mammographic CAD has shown its potential in assisting radiologists to improve detection rate of breast cancer (Gromet, 2008).

Extensive studies have been done on mammographic CAD. Reviews of breast mass detection and segmentation can be found in articles (Oliver et al., 2010; Schnabel et al., 2013). Generally, the design of breast CAD consists a number of stages. Firstly, region candidates are extracted from the mammograms, representing either abnormal or normal regions. To generate the region candidates, both unsupervised and supervised methods can be employed. Unsupervised methods normally involve the process of partitioning mammograms into multiple regions/segments by analysing the characteristics of pixels such as intensity, contrast

and topographic representations (Schnabel et al., 2013). To generate the region candidates, techniques such as region-grow (Görgel et al., 2013), thresholding (Kozegar et al., 2013; Varela et al., 2007), frequency domain filtering (Zhang et al., 2016), clustering (Martins et al., 2009) have been explored. These region candidate generation methods mostly aim at locating the suspicious regions. The regions generated normally include both the lesions and also many normal dense regions. The unsupervised methods tend to have difficulties in capturing characteristic lesions, such as spiculated lesions with highly irregular margins and architectural distortions (Schnabel et al., 2013). An alternative way of generating candidates is to use supervised methods. Instead of unsupervised segmentation, supervised methods use machine learning techniques such as convolutional neural networks (CNN) (Dhungel et al., 2015a) to study the characteristic features that describe mammographic masses, and generate a signal at the location of breast masses. The supervised methods may have the potential to detect the characteristic lesions. However, these types of methods also face relatively high computational complexity caused by the fact that the search should be performed exhaustively through the pixels in the image and even on different scales to avoid missing small lesions (Schnabel et al., 2013). Moreover, it may only provide the location of the masses (e.g. using bounding boxes) without accurately contouring the masses in the detection process (Dhungel et al., 2015a; Liu et al., 2015).

After the region candidates are generated, machine learning methods are normally used to determine whether the candidates are masses or normal tissue. Various machine learning algorithms have been applied to solve this problem, such as neural networks (Varela et al., 2007), support vector machines (Görgel et al., 2013; Martins et al., 2009), deep learning and random forest (Dhungel et al., 2015a). However, data imbalance is a common issue in this step, and high class imbalance normally leads to poor classification performance for the minority class (He and Ma, 2013). Since the minority class in this case are the positive

samples (masses) and the true positive rate (TPR or sensitivity) is critical to the system, a learning algorithm that can deal with the severe class imbalance and achieve a satisfactory sensitivity is required. Various methods have been developed to address the data imbalance problem. Sampling, either under-sampling the majority class, or over-sampling the minority class (Chawla et al., 2002), is a commonly used approach to balance the data. However, the sampled data may not be able to represent the distribution of the original data (Bria et al., 2014; Kang and Cho, 2006). Another solution is to introduce evaluation metrics to the data by assigning different weights to each class in proportion to its frequency in the training set (Viola and Jones, 2002). However, it can be difficult to design a robust evaluation metric to properly value the minority class (He and Ma, 2013). This type of method may also not perform as well as the sampling methods when facing an extremely imbalanced data set (Kang and Cho, 2006). To improve the sampling method, ensemble learning has been proposed to cope with severe class imbalance. Ensemble methods use a set of learning machines to make decisions and have been proven to be successful in solving class imbalance problem (He and Ma, 2013). Ensemble learning can be paired with random under-sampling methods such as RUSBoost (Seiffert et al., 2010), or oversampling methods such as SMOTEBoost (Chawla et al., 2003). To date, there are only a few publications that have addressed the data imbalance problem in breast mass detection (Chu et al., 2015; Kozegar et al., 2013; Murthy et al., 2013).

In this work, we propose a novel algorithm that can both detect and segment breast masses with a simple and straightforward structure. Figure 1 shows the diagram of the mammographic breast mass detection and segmentation system. The system mainly consists of a multi-scale morphological sifting approach and a self-grown cascaded random forests. Firstly, region candidates are generated by using multi-scale morphological sifting (MMS) and multi-level Otsu thresholding (MLT). To extract mass-like patterns, the MMS analyses

images by using two sets of linear structuring elements mapped in multiple orientations at each scale. The novelty of the region candidate segmentation method is that it creates a rather general morphological model of breast densities that captures the masses accurately regardless of their size and shape variations. Although morphological based methods have been used in several studies (Kharel et al., 2017; Kimori, 2011; Tan et al., 2015), however, they are mainly morphological operations used with other filters as pre-processing, while in our work the MMS plays the key role in region candidate segmentation.

The region candidate generation method normally produces many more normal tissue regions than mass regions, leading to a highly skewed training set. To cope with the severe data imbalance problem, a probability ranking based, adaptively self-grown cascaded random forests (CasRFs) is designed to classify the region candidates as abnormal or normal regions. The CasRFs assemble a series of random forest (RF) classifiers in a cascade manner, and gradually discard the least suspicious region samples throughout the layers by using an under-sampling approach guided by probability ranking. Once the remaining negative samples can no longer be separated from the positive samples, the cascade ceases to grow. Unlike many ensemble learning methods that adopted random under-sampling to attain a balanced training subset for each individual base classifier (Seiffert et al., 2010; Wei et al., 2015), the CasRFs under-sample the majority class by choosing the less lesion-like samples based on their probability to be a lesion, which ensures that the training subset at the current RF layer is highly separable. This enables the cascade to grow stably. The CasRFs are constructed in a simple and straightforward way without the need of involving complex weight matrixes and cost functions (Bria et al., 2016; Viola and Jones, 2002). The CasRFs are capable of dealing with highly skewed training sets and adaptive to the imbalance problem without the need of setting the number of layers, as in previous studies (Baumann et al., 2013; Tang et al., 2012). This work is an extension of our preliminary work (Min et al., 2017). We

have replaced the region candidate generation method with a novel multi-scale morphological sifting algorithm, and adopted a different initialization approach for the self-grown CasRFs. The new MMS method performs significantly better than the previous region candidate generation method based on regular morphological structuring elements (Min et al., 2017). The redeveloped self-grown CasRFs perform the training process on all training samples from all scales rather than training at each individual scale (Min et al., 2017). This work also presents extended analysis on the mechanism and performance of the CasRFs. Overall, the proposed method combines mass detecting and segmenting functions in one relatively simple structure that can be established on a regular desktop. Evaluated on two publicly available mammographic datasets, the proposed method achieves competitive performance compared with state-of-the-art methods that have much higher complexity (Dhungel et al., 2015a; Dhungel et al., 2017).

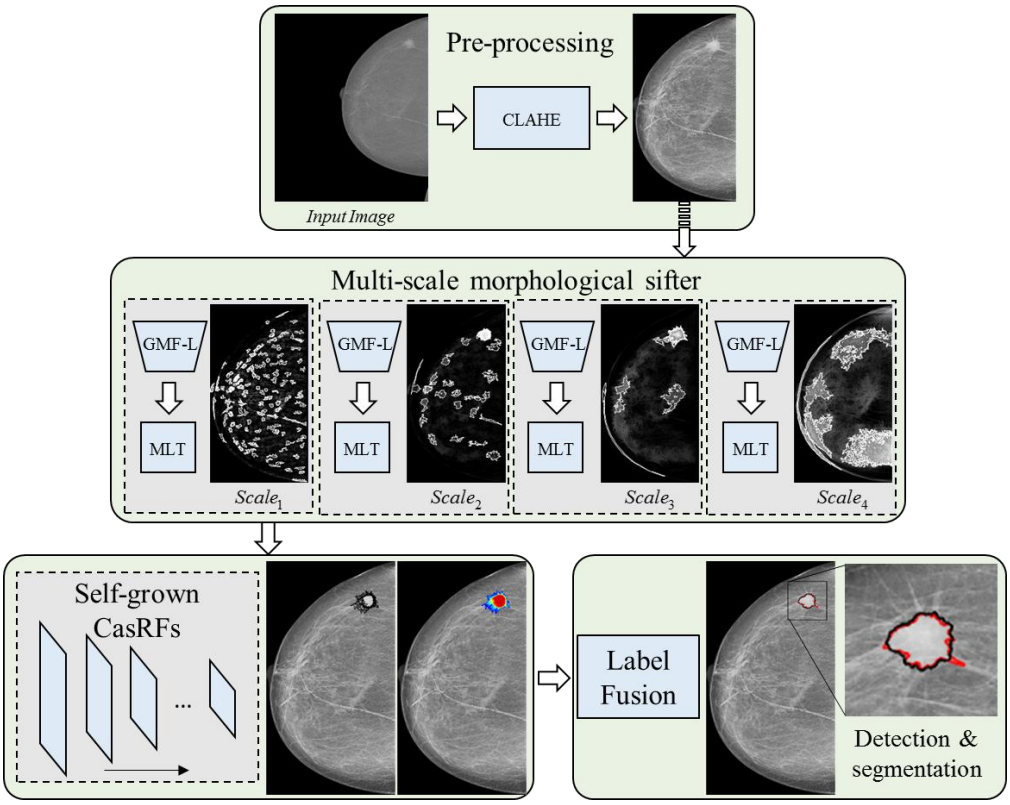


Figure 1. The diagram of the mammographic breast mass detection and segmentation system using multi-scale morphological sifting and self-grown cascaded random forests. CLAHE stands for contrast limited adaptive

histogram equalization. GMF-L stands for grayscale morphological filter using linear structuring elements. MLT stands for multi-level Otsu thresholding. The outlines of the detected patches are shown in black, and the ground truth is marked in red.

2 Material and methods

2.1 Mammographic datasets

The system is evaluated on two public mammographic datasets, INbreast (Moreira et al., 2012) and DoD BCRP (Bowyer et al., 1996). INbreast is a full-digital mammographic database containing 115 cases (410 images) with a 70 μm pixel size and a 14-bit contrast resolution. There are 116 masses within a size range between 15 mm^2 and 3689 mm^2 in this dataset. INbreast is currently the largest publicly available, annotated full-field digital mammographic dataset (Dhungel et al., 2017). The lesion locations and boundaries are outlined by an image specialist, which enables evaluation of segmentation performance of the CAD system. DoD BCRP is a screen-film mammographic dataset whereby a training set of 39 cases (80 annotated masses) and a testing set of 40 cases (81 annotated masses) are provided. The pixel size is 43.5 μm and the contrast resolution is 12-bit. DoD BCRP contains many characteristic lesions such as spiculated, ill-defined masses and architectural distortions, which is good for testing the performance of the proposed method on highly irregular masses. INbreast is more clinical relevant compared with DoD BCRP since most screening programmes have adopted full-digital mammography. Therefore, INbreast is used as the primary evaluation dataset for our algorithm, and the performance on DoD BCRP is only presented as secondary results. We believe it is important to use public available datasets for evaluation to attain unbiased performance comparison between studies.

2.2 Pre-processing

To speed up the process, the mammograms are often resized to a lower resolution in pre-processing (Chu et al., 2015; Dhungel et al., 2015a; Eltonsy et al., 2007; Min et al., 2017). Here, we reduce the size of the mammograms by a factor of 4 using bi-cubic interpolation. The breast profile is pre-segmented using a simple threshold and then the redundant background is cropped away in INbreast. The redundant background regions in DoD BCRP are manually cropped away due to the fact that the mammograms from DoD BCRP contain various image noises and artefacts in the background. The pixel values in the image are linearly rescaled to 16-bit and the contrast limited adaptive histogram equalization (CLAHE) (Pizer et al., 1987) is then applied. The number of tiles in CLAHE is set as {4, 4}, which is arbitrarily chosen, and the contrast enhancement limit is set as 0.01 (default).

2.3 Region candidate generation

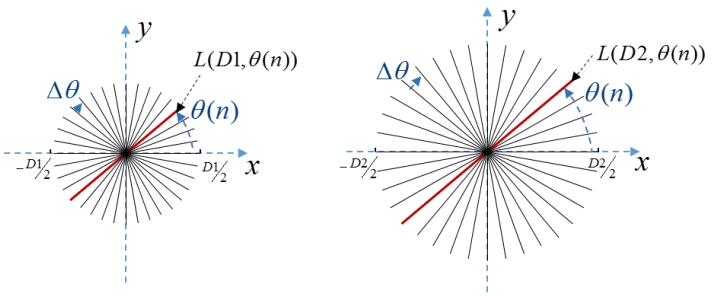
In order to generate the region candidates, a multi-scale morphological sifting (MMS) approach is applied to extract elements that are likely to be masses. Then multi-level Otsu thresholding (MLT) is employed to partition the image into meaningful patches that either represent mass regions or normal tissue regions.

2.3.1 Multi-scale morphological sifters

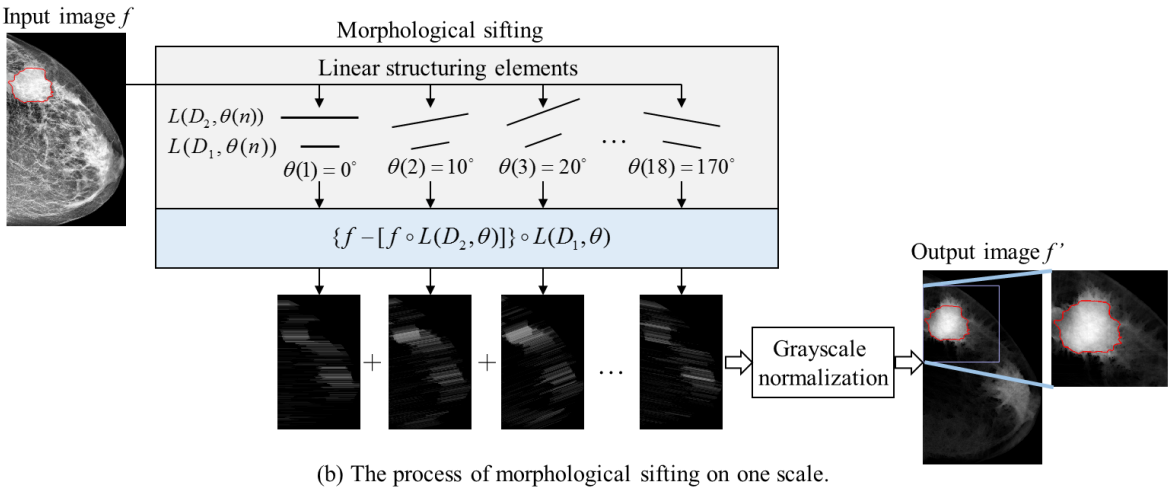
Morphological operations with structuring elements can be used to probe and analyse the images under the study for properties of interest (Gonzalez and Woods, 2008). Applying a dual-stage morphological filtering approach using two morphological operations (two top-hat operations or one top-hat followed by an opening) with different structuring elements can suppress the background tissue and extract the patterns of interest (Li et al., 2001; Wang, 2006). By altering the size and shape of structuring elements, this dual-stage morphological approach can extract elements that fit the size and shape priors introduced by the structuring

elements. Here, we propose a new set of multi-scale grayscale morphological filters by using linear structuring elements (MGMF-L) that cover a range of orientations, as malignant mammographic densities are often surrounded by a radial pattern of linear spicules (Karssemeijer and te Brake, 1996). Since the dual-stage morphological approach has the ability to sieve elements of interest from the background, we name the approach 'morphological sifting (MS)'. The algorithm is described as follows.

Firstly, two sets of linear structuring elements are defined as $\{L(D1,\theta(n))\,|\,n=0,1...N-1\}$ and $\{L(D2,\theta(n))\,|\,n=0,1...N-1\}$, where $D1$, $D2$ stands for the magnitudes, and $\theta(n)$ stands for the orientation of each individual line. $\theta(n)$ is equally spaced in $[0^\circ,180^\circ]$ as shown in Figure 2 (a). There are N elements in each set and the orientation difference between two adjacent lines is $\Delta\theta$ ($\Delta\theta=180^\circ/N$).



(a) Two sets of linear structuring elements. The lines marked in red represent a linear element pair $[L(D1,\theta(n)),L(D2,\theta(n))]$.



(b) The process of morphological sifting on one scale.

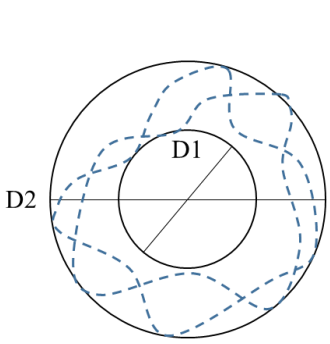
Figure 2. The process of morphological sifting using linear structuring element pairs in different orientations. (a) shows how the linear elements are mapped in space. (b) shows how the structuring element pairs

$\{[L(D1,\theta(n)), L(D2,\theta(n))] \mid n = 0,1...N-1\}$ are applied on the input image on each scale. This is an example on scale 4 and the lesion lands in scale 4 due to its size.

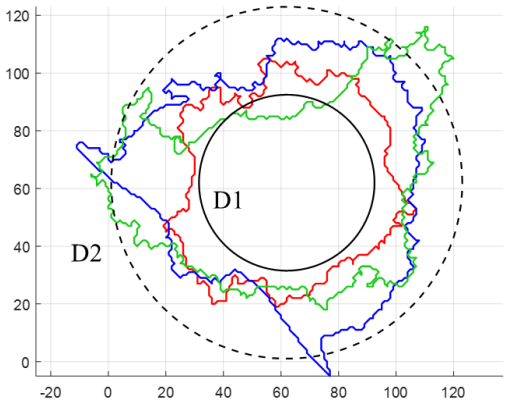
The grayscale morphological filter is described in the equation below.

$$f' = \sum_{n=0}^{N-1} \{ f - [f \circ L(D2,\theta(n))] \} \circ L(D1,\theta(n)) \ , \qquad 1$$

where f stands for the input image. '$\circ$' stands for morphological opening. f' is the processed image, defined as the summation of the outputs of morphological sifters over all the orientations. Figure 2 (b) shows the process of morphological sifting. Two linear structuring elements $[L(D1,\theta(n)), L(D2,\theta(n))]$, with the same orientation, are paired up. These structuring element pairs are then applied onto the input image individually as shown in Figure 2 (b). The output image is generated by applying a grayscale normalization on the summation of all the result images generated by the morphological filtering. Here, we imagine that the target breast densities are composed of numerous straight 'pen-strokes' in different directions and sizes, just like these linear structuring elements. By extracting these 'lines' in different orientations, we are also extracting the target objects. At a certain scale, the MMS extracts objects whose outlines lie approximately in the space between the circles with a diameter of $D1$ and $D2$ as shown in Figure 3 (a). This is because that the 'strokes' inside the target objects tend to survive the morphological operations done by structuring elements $L(D1)$ and $L(D2)$.



(a) Theoretical model of objects that MMS filters extract.



(b) Real masses extracted by MMS filters on one scale (scale3).

Figure 3. The morphological model of the objects that MMS extracts. $D1$ and $D2$ here stand for the diameters of the two circles, as well as the magnitudes used in the linear structuring elements.

The morphological sifters are applied at multiple scales targeting lesions within different size ranges to cope with the size variation of masses. If the size range of masses is $[Area_{\min}, Area_{\max}]$, the pixel size of the image is P, the number of scales used is M, and the resizing factor used in the pre-processing stage is R, we can roughly estimate the magnitude range of linear structuring elements as,

$$[DI_{\min}, DI_{\max}] = \left[2 \times \frac{\sqrt{\frac{Area_{\min}}{\pi}}}{P \times R}, 2 \times \frac{\sqrt{\frac{Area_{\max}}{\pi}}}{P \times R} \right]. \quad (2)$$

Then, an 'exponential' morphological structuring element bank $\{L(D1(i)), L(D2(i)) \mid (i = 1, 2, ..., M)\}$ is generated. M is the number of scales used, and the scale interval SI is

$$SI = \left(\frac{DI_{\max}}{DI_{\min}} \right)^{\frac{1}{M}}. \quad (3)$$

The magnitudes of the linear structuring element pair at the ith scale ($Scale_i$) are

$$\{D1(i) = DI_{\min} \times SI^{i-1}, D2(i) = DI_{\min} \times SI^{i} \mid (i = 1, 2, ..., M)\}. \quad (4)$$

Therefore, at $Scale_i$, the MS filtered image $f'(i)$ is

$$f'(i) = \sum_{n=0}^{N-1} \{ f - [f \circ L(D2(i), \theta(n))] \} \circ L(D1(i), \theta(n)) \,. \quad (5)$$

Thus, the morphological sifter at $Scale_i$ targets elements within the size range of

$$[Area_{D1(i)}, Area_{D2(i)}] = [\pi/4 \times D1^2(i),\ \pi/4 \times D2^2(i)] \,. \quad (6)$$

Generally, as the number of linear structuring elements N increases, the smoother the processed image becomes. However, this also comes with an increased computational complexity. Figure 4 shows the examples of applying the morphological sifters for different N. Here, N is set as 18 ($\Delta\theta = 10^\circ$) and f' is normalized into 16-bit in our case. The number of scales is set as 4.

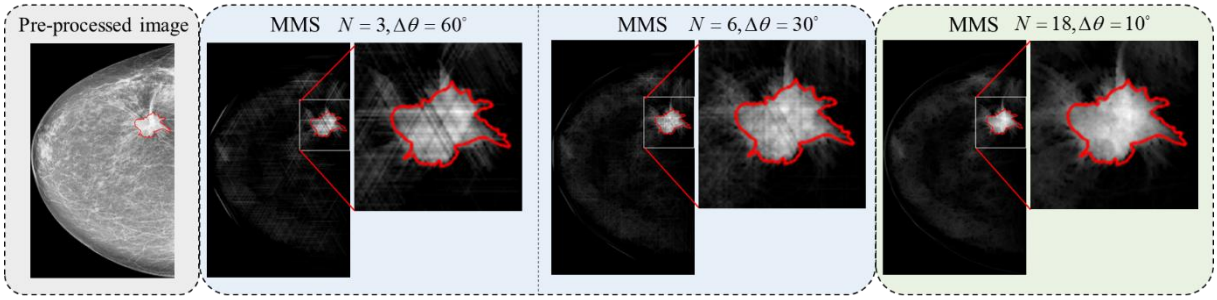


Figure 4. An example of the multi-scale morphological sifters (MMS) at one of the scales using $N = 3$, $N = 6$, and $N = 18$. The red outlines denote the ground truth of the lesions.

2.3.2 Multi-level Otsu thresholding

After multi-scale morphological sifting, the multi-level Otsu thresholding (MLT) method is then applied on the processed images at each scale. Otsu's method is a classic thresholding method (Otsu, 1979), which maximizes the between-class variance based on the histogram of an image (Gonzalez and Woods, 2008). Multi-level Otsu is an extension of the original Otsu method and is a reliable way to generate the optimal thresholds to segment an image (Backes and Bruno, 2008; Liao et al., 2001). Figure 5 shows the mechanism of region candidate segmentation using MMS and MLT. The MLT generates a series of thresholds $[T_k \mid k = 1, ..., K]$. For each threshold T_k, if a pixel value in the filtered image $f'(x, y) > T_k$, then the pixel value in

the output binary image $B(x,y)=1$, if $f'(x,y) \le T_k$, then $B(x,y)=0$. K thresholds should generate K binary images representing the segmentations at each level of threshold. However, in order to save space, at each scale, we plot the contours of the segmentations generated from different thresholds together on the same image as shown in 'The segmentation outputs from MLT' in Figure 5. Generally, the more scales and the more levels in multilevel Otsu, the more likely for the algorithm to capture the accurate outlines of the lesions, but with a higher computational complexity. Here, we use four scales and 16-level multilevel Otsu to segment the images. Only the patches within the targeted size range (equation 6) at each scale are selected as the region candidates as shown in Figure 5.

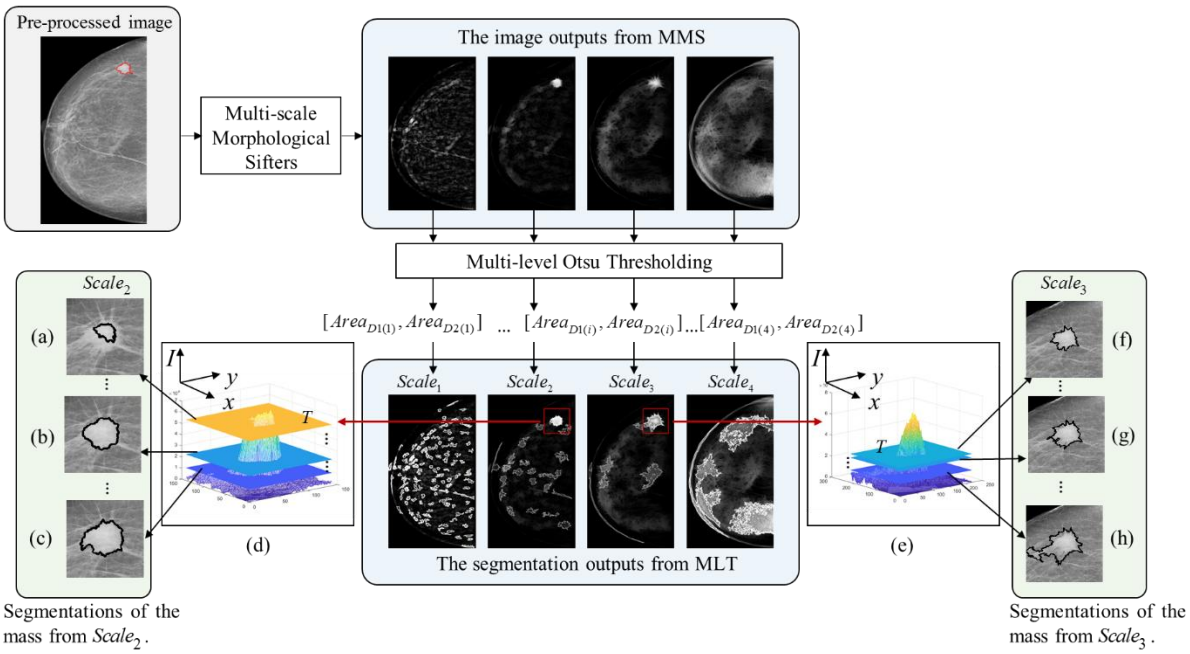


Figure 5. Region candidate segmentation based on multi-scale morphological sifting and multi-level Otsu thresholding on 4 scales. (d) and (e) are the x-y-Intensity plots of the lesion area in $Scale_2$ and $Scale_3$, where the flat surfaces show how different thresholds slice through the intensity value and generate the segmentations. (a) ~ (c) show several segmentation examples at the lesion area on $Scale_2$, while (f) ~ (h) show some examples at the lesion area on $Scale_3$. Both $Scale_2$ and $Scale_3$ generate relatively satisfactory segmentations of the lesion.

2.4 Region candidate classification using self-grow cascaded random forests

After the region candidates are generated, intensity features such as average intensity of pixels within the candidate region, kurtosis, skewness, contrast (te Brake et al., 2000), shape features such as eccentricity, extent, solidity, circularity (Cascio et al., 2006), radius, and texture features such as entropy, grey-level co-occurrence matrix (GLCM), inertial momentum (Cascio et al., 2006), are extracted. Haar-like features (Viola and Jones, 2001) are also used, and they are defined as in Figure 6 below. Three types of Haar features, horizontal, vertical and diagonal are operated at two sizes d and $0.5d$, where d is the longest distance between the pixels on the boundary and the centre of the region.

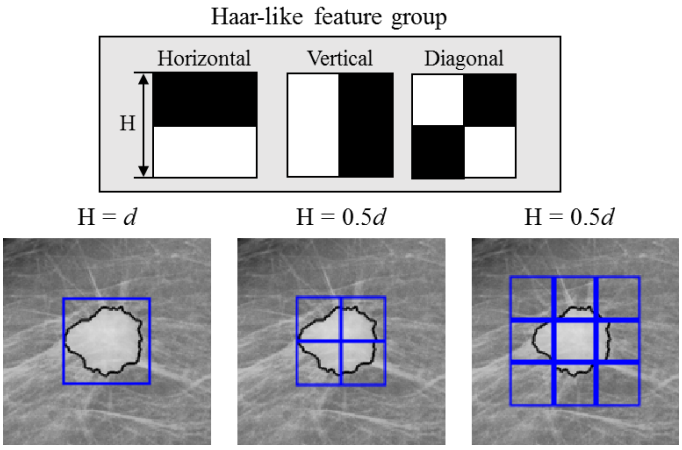


Figure 6. Haar-like feature group. d is the longest distance between the pixels on the boundary of the region and the centre of the region.

The previous stage generates many more non-lesion patches than mass patches. The ratio between positive to negative samples is approximately 1:160 in the training set, which indicates that the training set is highly imbalanced. Here, we design an ensemble learning method, self-grown cascaded random forests (self-grown CasRFs), to deal with the data imbalance problem. The random forest (RF) is chosen as the base classifier due to its advantages that it has relatively high accuracy, efficiency and is less prone to overfitting (Breiman, 2001).

The growth of the multi-layered self-grown cascaded random forests is shown in Figure 7 (a). There are two main parameters for RF, number of trees used in the forest (ntree), and number of features used in each tree (mtry). At $Layer_j$, all the positive and current negative samples (all negative samples for $Layer_1$) are put through a RF with default settings (DRF, ntree = 500, mtry = $\sqrt{\text{number of features}}$) as shown in Figure 7 (b). The negative samples are then sorted in ascending order of (posterior) probabilities generated from the DRF. To balance the two classes, if the number of positive samples is *NP*, the first *NP* negative samples are chosen to establish a balanced training set. Since the negative samples have been ranked in ascending order, the chosen negative samples are the least mass-like candidates and therefore relatively easy to distinguish from the positive candidates. Then a RF is trained on the balanced training set using a grid search for parameters ntree between $[100,1000]$, and mtry between $[\frac{1}{2}\sqrt{\text{number of features}}, 2\sqrt{\text{number of features}}]$, to find a model that reaches a sensitivity threshold *ST*. If no RF can reach *ST* in the current round of grid search, top 10% of the negative samples (with the highest probabilities, more likely to be lesions) are removed from the selected training set as shown in Figure 7 (b). After a RF model that meets *ST* is attained, the currently unused negative samples are classified by the RF model at $Layer_j$ and the true negative (TN) samples are discarded from the training. The remaining FPs and all positive samples then become $Input\ Samples_{j+1}$ and the learning moves on to $Layer_{j+1}$. As the cascade grows out more layers, the relatively easy non-lesion samples are gradually removed and the classification task becomes increasingly difficult. When the number of FPs stops to decrease, it indicates that the positive and negative samples can no longer be separated apart by the CasRFs, and the cascade stops to grow, as shown in Figure 7 (a). The formation of CasRFs is adaptive to the imbalance problem, requiring no need to manually choose the number of layers. The CasRFs are trained on the samples gathered from all the scales in the previous

stage. Here, a MATLAB RF package adapted by Abhishek Jaiantilal (Jaiantilal, 2013) based on (Liaw and Wiener, 2002) is used.

The testing process is similar to the training process. If testing samples are classified as negative (probability < cut-off value) by a RF classifier at the current layer, then they are removed and identified as normal regions. If they are classified as positive (probability ≥ cut-off value), then they are put into the next RF classifier in the cascade. The testing phase of CasRFs is illustrated in Figure 8. Since the design of the CasRFs is to filter out the easier (less lesion-like) negative samples first and then the difficult (more lesion-like) ones later, the negative samples in the testing data that are less likely to be masses are gradually discarded from the layers of RF models and only the samples left at the last layer are regarded as the suspicious regions.

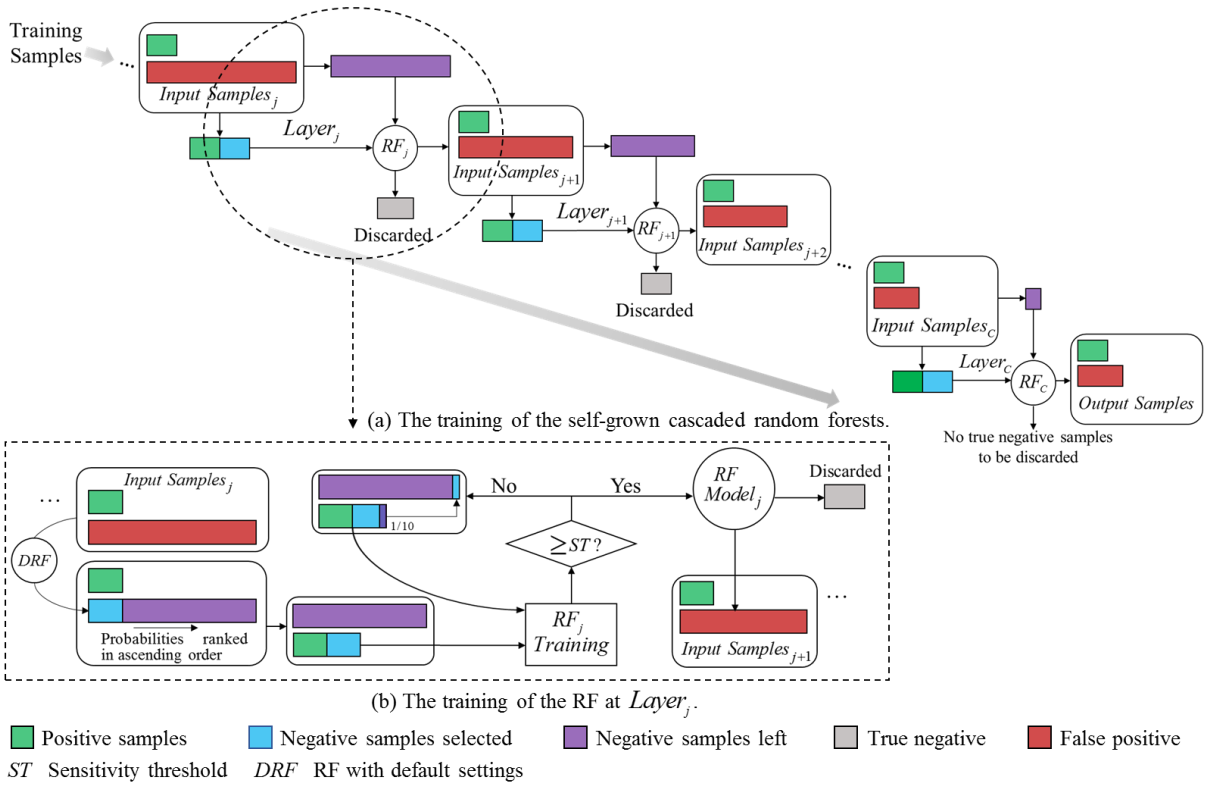


Figure 7. The training of the self-grown cascaded random forests. (a) shows an overview of the training of the whole cascade, and (b) shows the training of a layer. DRF stands for RF with default settings.

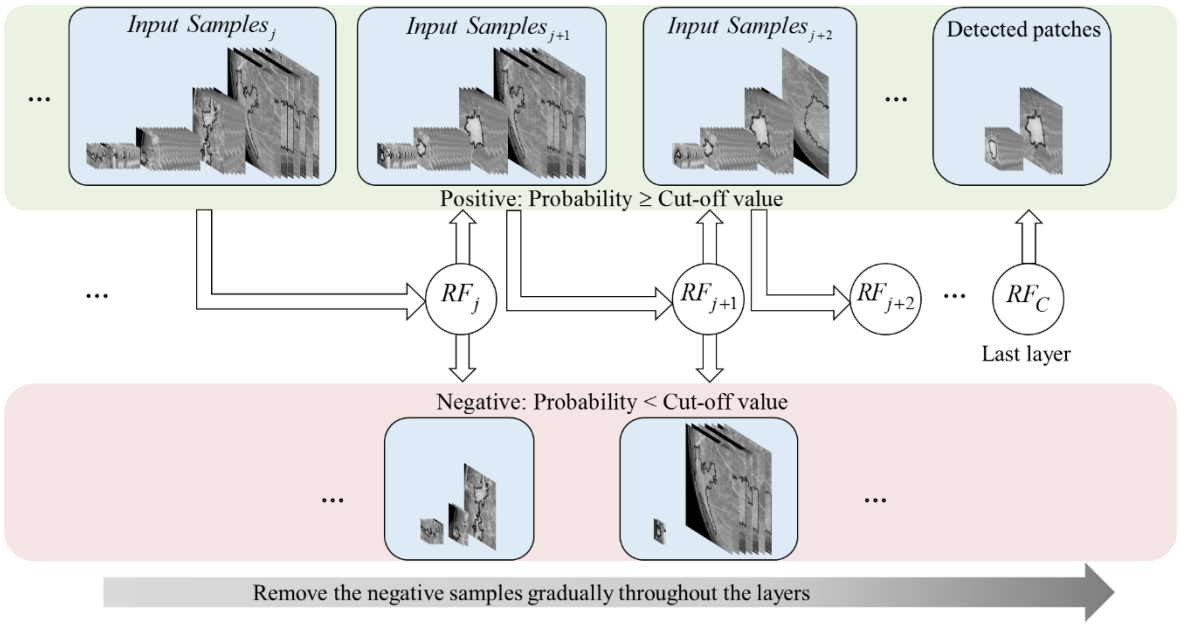


Figure 8. A schematic of the testing process using the CasRFs.

2.5 Post-processing and visualizing the results

After the classification, the outlines of the detected patches are plotted on the mammograms and a heat map of the probability is also generated as shown in Figure 9 (a), (b). However, there are often multiple identified suspicious patches overlapping with each other as shown in Figure 9 (c). To improve the visualization of the detection output and evaluate the segmentation accuracy, a label fusion method is designed to merge the overlap regions into one region by analysing the contrast between the inner area and the outer area of the region. Figure 9 (e) shows the technique of calculating the local contrast. Here, the contrast is defined as the subtraction between the average intensity of pixels within patch P_q and the average intensity of pixels within the background area P_q". The patches are firstly merged with in their scales and then between the scales. The outline of the region with the highest contrast is kept as the final contour of the detected region as shown in Figure 9 (f).

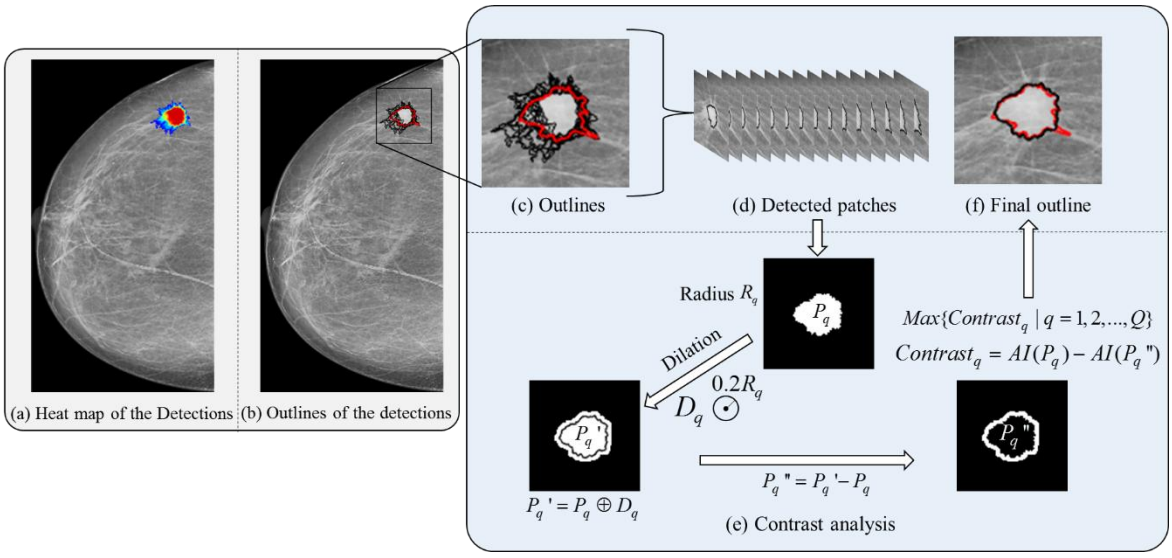


Figure 9. A schematic of the label fusion through local contrast analysis. (a) is the heat map of the detections from the CasRFs and all the detected patches are plotted on (b). (d) shows the detected patches and they go through a contrast analysis approach as shown in (e). P_q represents the inner region and P_q'' represents the outer region of a detected patch. *AI* stands for average intensity of pixels within a certain region. The outline of the patch that has the highest contrast is selected as the final outline of the detection. (f) shows the final fused outline.

3 Experiments

Our system is evaluated on two public available datasets, INbreast and DoD BCRP. For the evaluation on INbreast, we use the same evaluation method, repeated random sub-sampling validation (Dubitzky et al., 2007), as studies (Dhungel et al., 2015a, 2016; Dhungel et al., 2017). In these studies, the INbreast dataset is randomly split into training, validation and testing sets five times. In our work, we used the same testing set while we combine the training and validation sets together as a larger training set in each validation since random forest cross-validates itself internally when building the trees. For the evaluation on DoD BCRP, we simply use the pre-separated training and testing sets specified in this dataset.

During the classification of region candidates, the training samples are labelled as positive or negative for training according to their dice similarity index (DSI) (Dice, 1945) to the ground

356 truth. The DSI between a region A and ground truth B is defined as $2\times|A\cap B|/(|A|+|B|)$ (Dice,
357 1945). The patches with the highest DSI are labelled as positive and those with a DSI lower
358 than 0.1 are labelled as negative in the training sets. Overall, major inputs required by the
359 system include a size range of the lesions (which has been specified in the INbreast
360 documentation (Moreira et al., 2012)), a number of linear structuring elements used in MMS
361 (set as 18 in this study, recommend values larger than 10), a number of scales used in MMS
362 (set as 4 in this study), a number of levels in the multi-level thresholding (set as 16), and a
363 sensitivity threshold ST for CasRFs (set as 0.99 for INbreast and 0.97 for DoD BCRP). The
364 cut-off value inside random forests during training remains as the default value 0.5. All
365 experiments are carried on a Dell desktop with Intel Core i7-4790 CPU @ 3.60GHz, 16GB
366 RAM.

367 **4 Results**

368 The MMS filter working with MLT can capture 100% of the masses in INbreast at a
369 minimum DSI with the ground truth of 0.4, and 98% of the segmentations have a DSI with
370 the ground truth higher than 0.7. Figure 10 shows a performance comparison of the region
371 candidate segmentation between our previous method (using circular structuring elements)
372 and the proposed method. It can be seen the proposed method produces significantly better
373 segmentations compared to our previous work (Min et al., 2017). Since DoD BCRP provides
374 no accurate segmentation of lesions in the annotations, we could not evaluate the ROI
375 segmentation on this dataset quantitatively, instead, we present some examples of the region
376 candidate segmentation in Figure 11, where the masses are spiculated with an irregular shape
377 or architectural distortion. It can be seen that the ROI segmentation method based on MMS
378 and MLT still performs well on irregular masses.

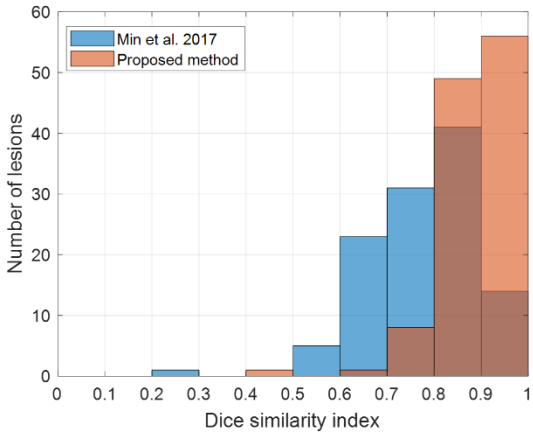


Figure 10. Performance comparison of region candidate segmentation between the proposed method and our previous study (Min et al., 2017).

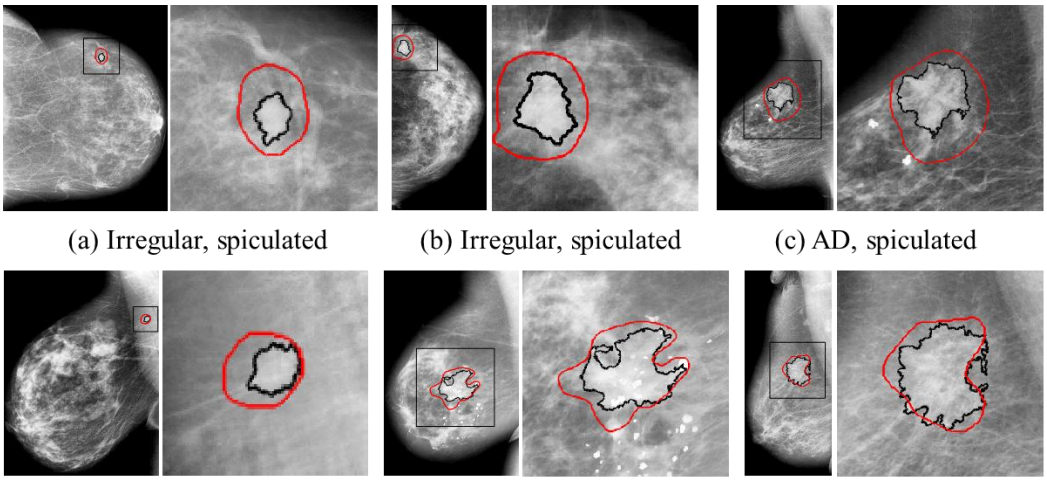


Figure 11. Examples of the region candidate segmentation on DoD BCRP. The red lines stand for the ground truth and the black lines stand for the segmentation of the lesion using MMS and MLT. Note that only the segmentations of the lesions are outlined in these images. AD stands for architectural distortion.

Before the label fusion, the system outputs heat maps based on the probabilities generated from CasRFs. Figure 12 shows examples of the heat maps showing the confidence of the detections. The boxplots of FP rate and DSI on INbreast dataset are shown in Figure 13. After the label fusion, if a detected region has a DSI $\geq DT$ ($DT \in \{0,1\}$), it is regarded as a true positive, if not, it is regarded as a true negative. For validation, the repeated random sub-sampling method is used.(Dhungel et al., 2017; Dubitzky et al., 2007) When the threshold

DT is set as 0.2, the average sensitivity across all the validations is 0.93 at 4.73 FPs per image (FPI) before the label fusion and at an average FPI of 1.64 FPs/image after the label fusion by contrast analysis. The average sensitivity S_A and average FPI FPI_A here are defined as,

$$S_A = \frac{\sum_{v=1}^{V} \text{number of lesions detected } (v)}{\sum_{v=1}^{V} \text{number of lesions}(v)}, \quad (7)$$

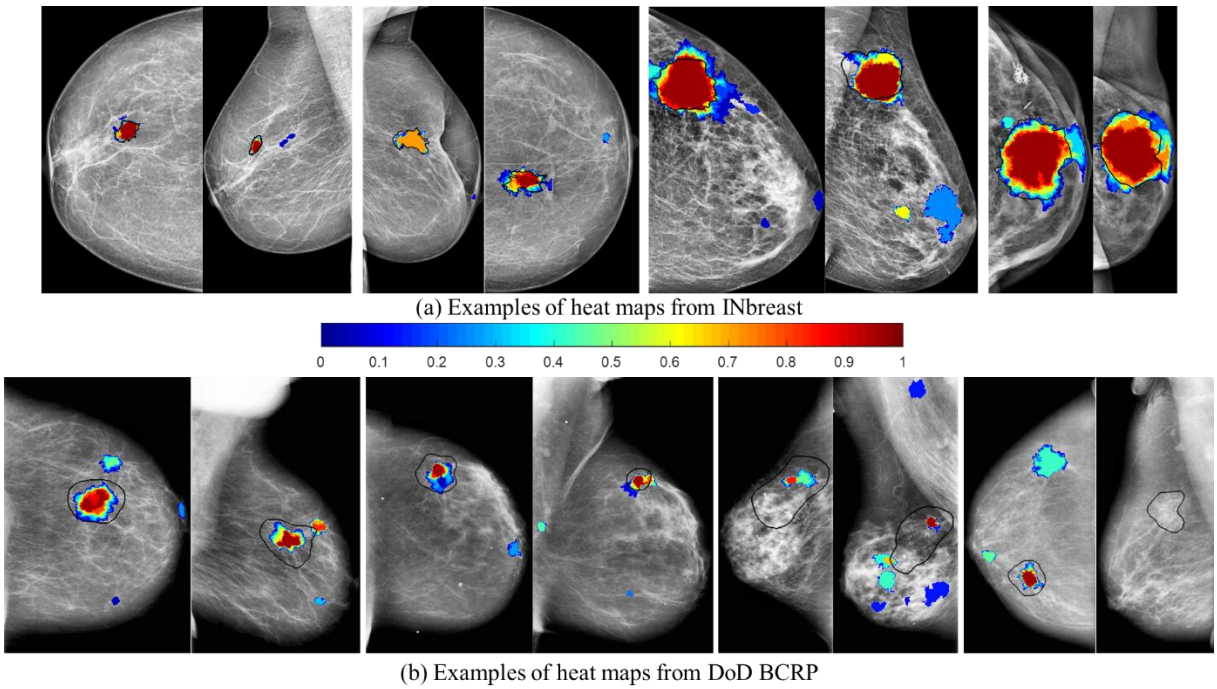
$$FPI_A = \frac{\sum_{v=1}^{V} \text{number of FPs } (v)}{\sum_{v=1}^{V} \text{number of images}(v)}, \quad (8)$$

where v stands for the v^{th} validation and V stands for the number of validations which is 5 in this work. A normalized partial area under the FROC curves (PAUC) is also calculated to provide an integral measurement of the performance. The PAUC is defined as $PAUC = \frac{1}{|\beta - \alpha|}\int_{\alpha}^{\beta} TPR\, dFPI$, where $[\alpha, \beta]$ stands for the range of FPI and set as $[0,5]$ in our case. The PAUC is 0.91, and 0.89 for the proposed method and our previous method respectively when $DT = 0.2$. The PAUC is 0.90 and 0.88 for the proposed method and our previous method when $DT = 0.5$.

The average DSI of the segmentation on INbreast is 0.85 ± 0.10 (minimum DSI > 0.4). The receiver operating characteristic (ROC) and free response operating characteristic (FROC) curves before the label fusion are shown in Figure 14 (a) and (b). Figure 15 shows the FROC curves on the testing set after the label fusion under the condition of $DT = 0.2$ and $DT = 0.5$, comparing with previous studies (Dhungel et al., 2017; Min et al., 2017). Figure 16 shows the relations between TPR, FPI and cut-off value in CasRFs before and after the label fusion. As shown in the FROC curve in Figure 15(c), the detection system can achieve an average sensitivity of 0.91 ± 0.06 at approximately 1.2 FPs per image when the average DSI is

413 0.86 ± 0.08 (minimum DSI > 0.6). The segmentation examples of lesions in various sizes and
414 shapes from INbreast are shown in Figure 17.

415 Additional evaluation of the proposed method is also conducted on DoD BCRP. Since DoD
416 BCRP only provides an approximate manual annotation, it can be difficult to evaluate the
417 segmentation performance. Here, we decide if the overlap ratio (Ben-Ari et al., 2017;
418 Dhungel et al., 2015a; Reichel and Cole, 2016) between a detected region and the ground
419 truth is higher than 0.7, the lesion is regarded as detected. Here, the overlap ratio between two
420 regions *A* and *B* is defined as $A \cap B / \min(A, B)$ (Reichel and Cole, 2016; Szymkiewicz, 1934).
421 Thus the sensitivity of detection on DoD BCRP is 0.81 when the sensitivity threshold *ST* in
422 the CasRFs is set as 0.97 and the cut-off value in the individual RFs is 0.5 by default, at a FPI
423 of 3.59 after label fusion. The comparison between this work and other state-of-art methods
424 in terms of TPR, FPI and segmentation accuracy (DSI) where applicable is shown in Table 1.

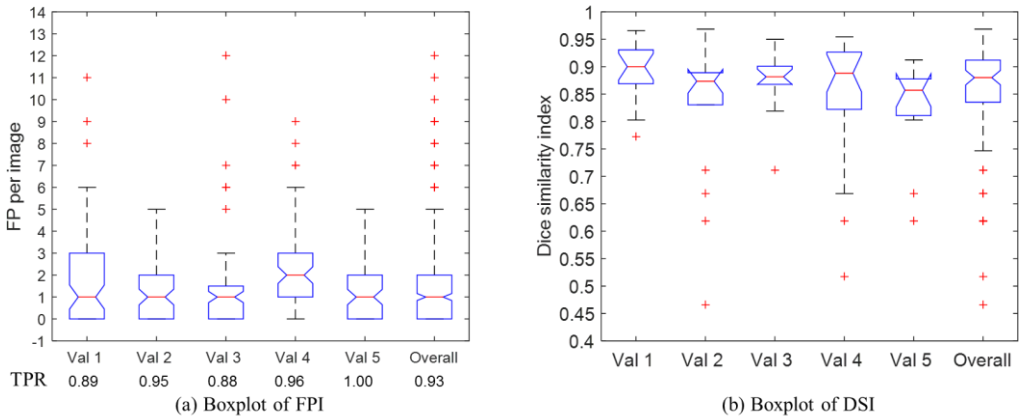


(a) Examples of heat maps from INbreast

(b) Examples of heat maps from DoD BCRP

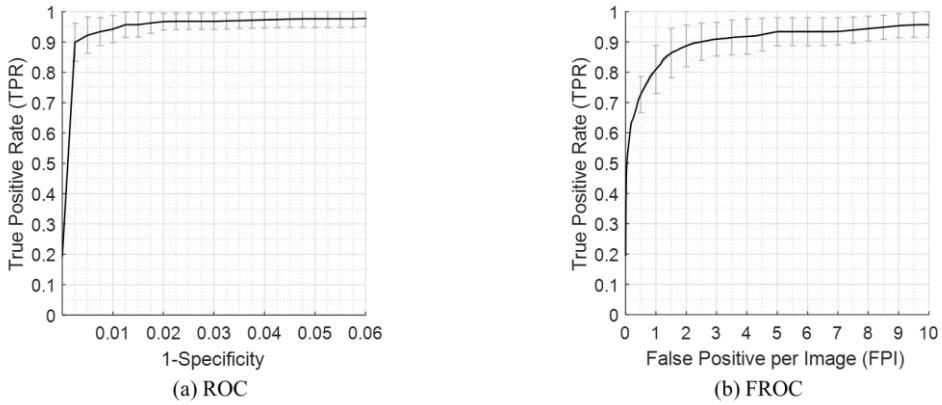
425

426 Figure 12. The examples of detection heat maps and the colour bar showing the mapping of probability values
427 into the colour map. The ground truth is labelled with black lines.



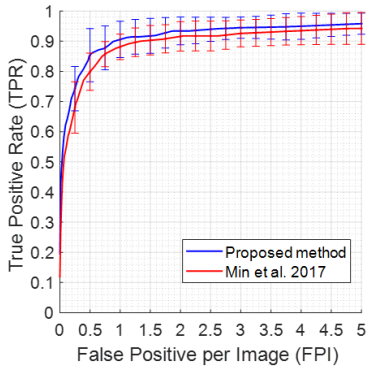
428

429 Figure 13. Boxplots of sensitivity & FP rate and DSI. 'Val' is short for 'validation' and TPR stands for true
430 positive rate (sensitivity). (a) shows the boxplots of FPI for all the validations and the TPR at each validation. A
431 boxplot of the overall FPI together with the average TPR across all the validations is also shown in (a). (b)
432 shows the boxplots of DSI (segmentation accuracy) at each validation and also an overall boxplot of all the
433 detected masses among all the validations. The median of the overall average DSI is 0.88.

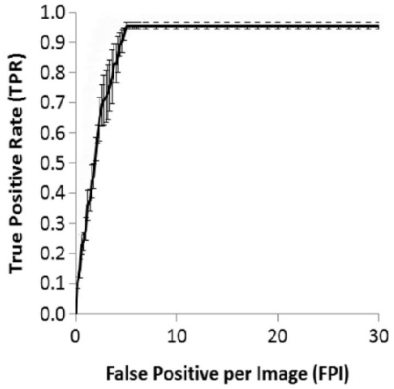


434

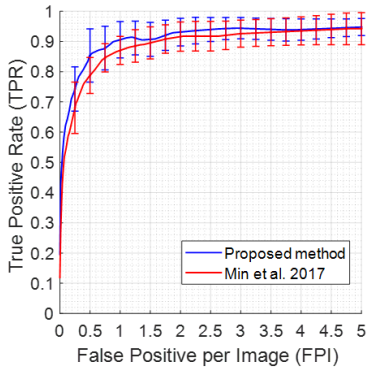
435 Figure 14. ROC and FROC of the classification on the testing samples from INbreast before label fusion.



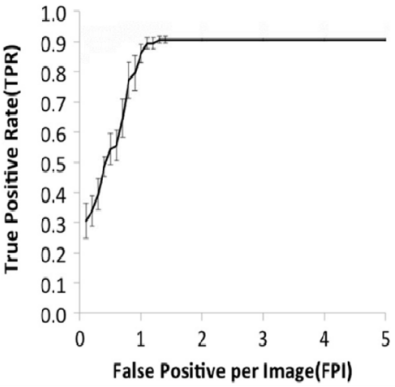
(a) FROC of the proposed method and Min et al. 2017 (*DT*=0.2)



(b) FROC of Dhungel et al. 2017 (*DT*=0.2)

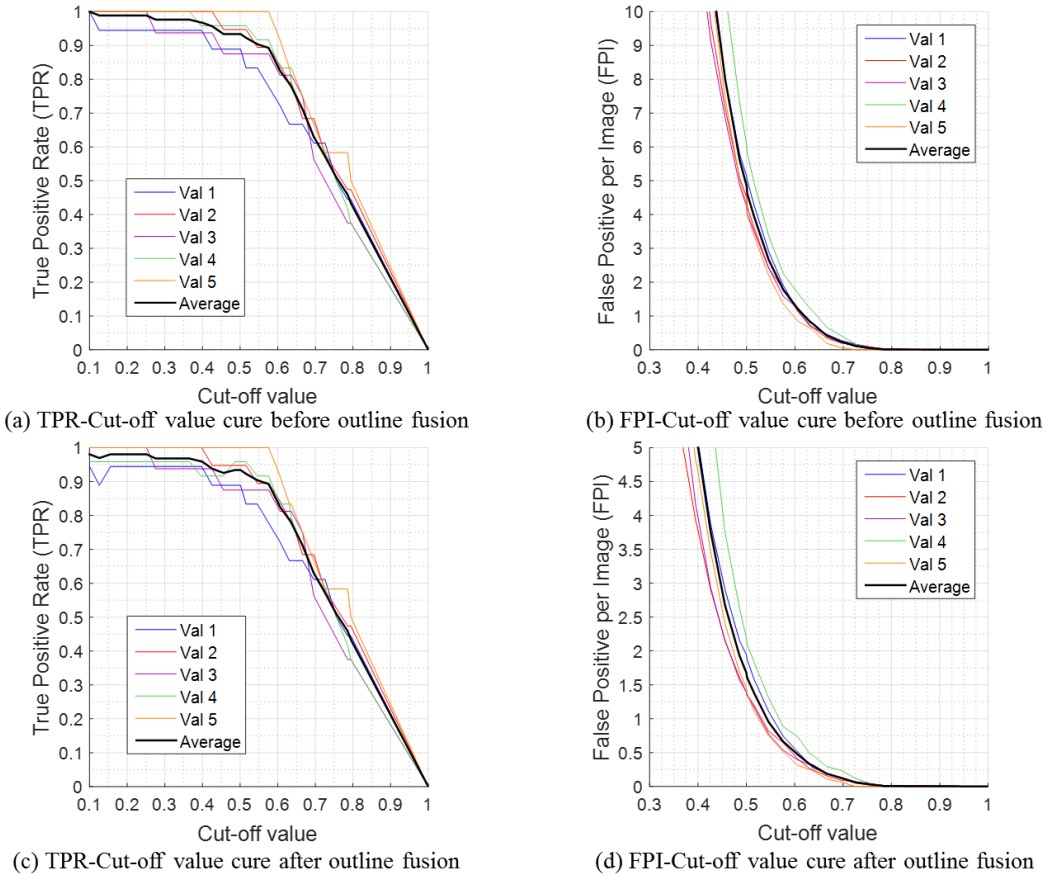


(c) FROC of the proposed method and Min et al. 2017 (*DT*=0.5)



(d) FROC of Dhungel et al. 2017 (*DT*=0.5)

Figure 15. The FROC curves on the testing dataset after label fusion. (a) and (b) show the FROC curves of the proposed method, our previous work (Min et al., 2017) and study (Dhungel et al., 2017) with $DT = 0.2$. (c) and (d) show the FROC curves of the proposed method, our previous work (Min et al., 2017) and study (Dhungel et al., 2017) with $DT = 0.5$.



(a) TPR-Cut-off value cure before outline fusion (b) FPI-Cut-off value cure before outline fusion

(c) TPR-Cut-off value cure after outline fusion (d) FPI-Cut-off value cure after outline fusion

Figure 16. The relationship between TPR, FPI and the cut-off value (default 0.5) before and after the label fusion on INbreast. 'Val' is short for 'validation'. The black line represents the average TPR and FPI when given different cut-off values ($DT = 0.2$).

Table 1. Performance comparison between the proposed method and previous publications. The third column indicates whether the method generates both detection and segmentation of the lesions or only detection. √ means the method detects as well as segments the lesions. TPR stands for true positive rate (sensitivity), FPPI stands for FPs per image, and DSI stands for dice similarity index.

Methods	Dataset	Segmentation	Avg. TPR @ FPPI	Avg. DSI
(Kozegar et al., 2013)	INbreast	√	0.87@3.67	-
(Dhungel et al., 2015a)	INbreast		0.96@1.2,0.87@0.8	-
(Dhungel et al., 2016) & (Dhungel et al., 2017)	INbreast	√	0.90@1.3	0.85
(Min et al., 2017)	INbreast	√	0.94@1.99	0.80
The proposed method	INbreast	√	0.93@1.64,0.91@1.2	0.85,0.86
(Beller et al., 2005)	DoD BCRP	√	0.70@8	-

(Dhungel et al., 2015a)	DoD BCRP		0.75@4.8,0.7@4	-
(Min et al., 2017)	DoD BCRP	√	0.77@3.93	-
The proposed method	DoD BCRP	√	0.81@3.59	-

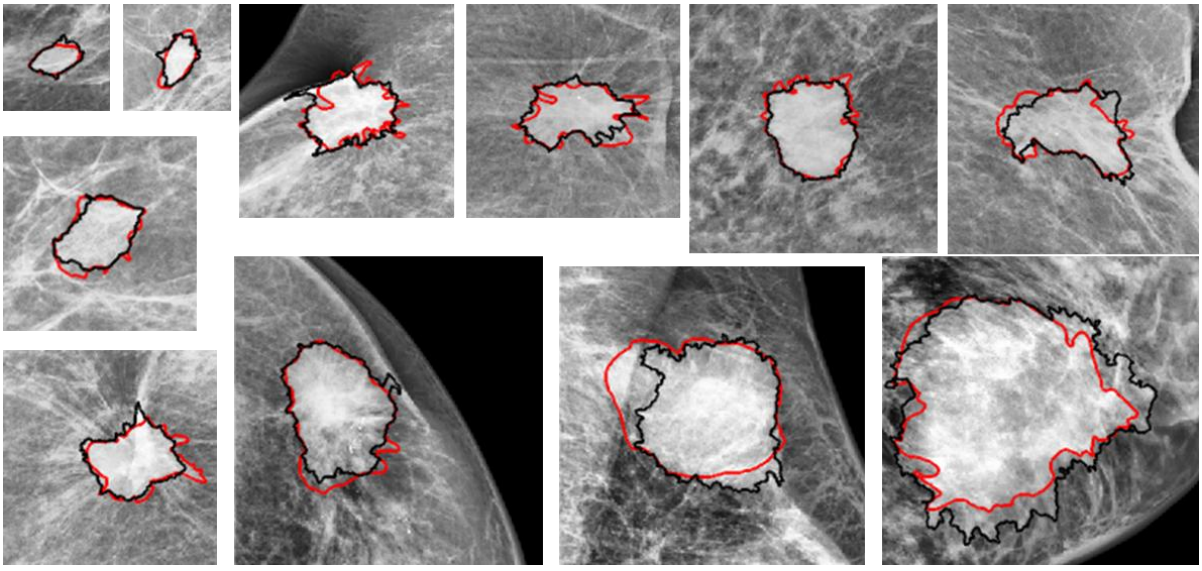


Figure 17. Segmentation examples of masses in various shapes and sizes from INbreast dataset. The black lines represent the segmentation from our method, and the red lines are the ground truth.

5 Discussion

5.1 Unsupervised segmentation using morphological sifting and multi-level thresholding

In this work, a ROI segmentation method based on MMS and MLT is designed to separate the lesions from the background tissue. As mentioned in section 1, the main challenges for breast mass segmentation are the large variations of lesions in size, shape and contrast (Oliver et al., 2010). Generally, lesions appear in irregular shape and spiculated margin are highly suspicious in malignancy. In order to cope with these challenges, the multi-scale morphological sifters are developed to extract the regions of interest (ROI) within the mass size range. Applying the morphological sifters at multiple scales can limit the size variation at each individual scale. The new morphological sifting method adopts linear structuring elements that are mapped in a multiple directions and length to imitate the pattern of masses

as shown in Figure 2. After MMS and MLT, the average DSI between the best segmented region candidates and the ground truth is 0.89±0.07 among all the cases in INbreast, which indicates the segmentation method performs well regardless of the variations in the mammographic appearance of the lesions. In previous studies, disk-shaped structuring elements have been used to extract mass-like patterns (Chu et al., 2015; Min et al., 2017). They are effective in extracting densities of a semi-circular shape, however they may not be as effective on masses with highly irregular shapes and spiculated margins that are more likely to be malignant (Karssemeijer and te Brake, 1996). Figure 18 shows the comparison between applying linear (this work) and disk-shape structuring elements (Min et al., 2017) to extract spiculated masses with relatively irregular shapes. Figure 10 has also provided a statistic comparison of the segmentation accuracy for lesions between these two methods. The new multi-scale morphological sifting method clearly show advantages over our previous method. As to the region candidate generation methods proposed by recent studies (Casti et al., 2016; Görgel et al., 2013; Kozegar et al., 2013; Zhang et al., 2016), these approaches mainly aim at locating suspicious regions (e.g. to find the focal area of the lesions) and did not report the segmentation accuracy of ROI generation, while the proposed MMS and MLT based method generates relatively accurate segmentations. Moreover, this approach is not affected by the existence of the pectoral muscle. There are special cases where in the MLO view, lesions can fully or partially overlap with the pectoral muscle as shown in Figure 11 (d), and removing pectoral muscle means at least part of the lesion will also be removed. Many studies remove the pectoral muscle in pre-processing (Casti et al., 2016; Soulami et al., 2017; Tan et al., 2015; Zhang et al., 2016), while the proposed method does not need to, therefore it can extract lesions that overlap with the pectoral muscle. This ROI generation method is the foundation of the detection and segmentation of the whole system. Compared with studies based on convolutional neural network (CNN) (Dhungel et al. 2015, 2016 and

2017), the MMS sifter is much simpler and computationally efficient. It removes the irrelevant parts of the breast, extracts the suspicious regions accurately at the early stage of the algorithm and provides exact regions of interest for the later learning step. This helps to avoid the need of iterative searching as in CNN based algorithms.

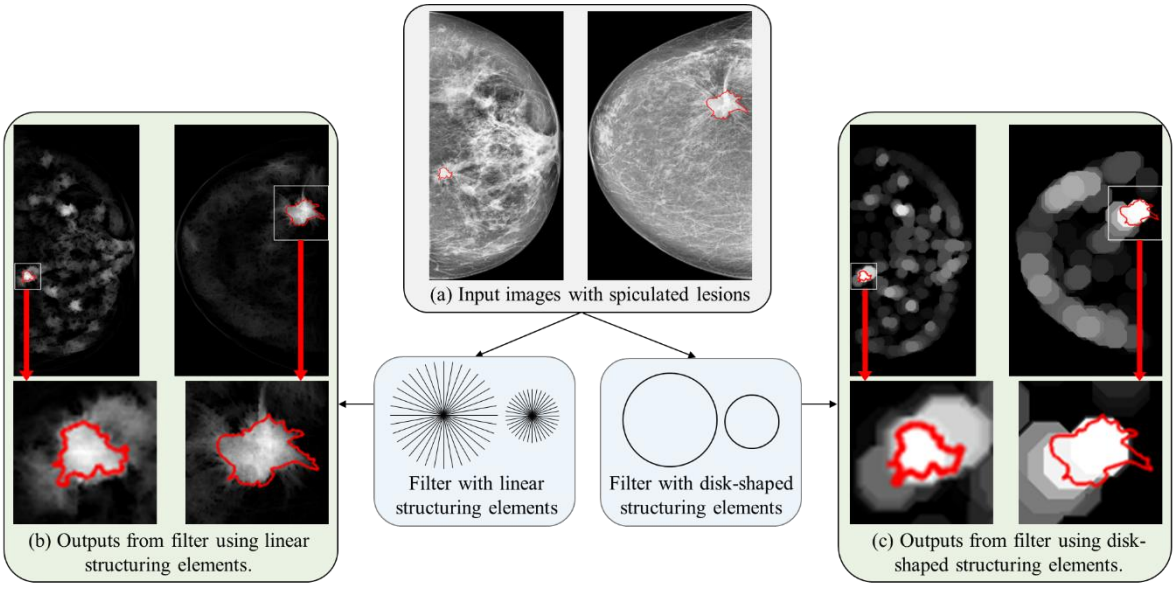


Figure 18. Comparisons between using linear and disk-shape structuring elements. The two examples contain lesions with spiculated margin. The red lines represent the ground truth.

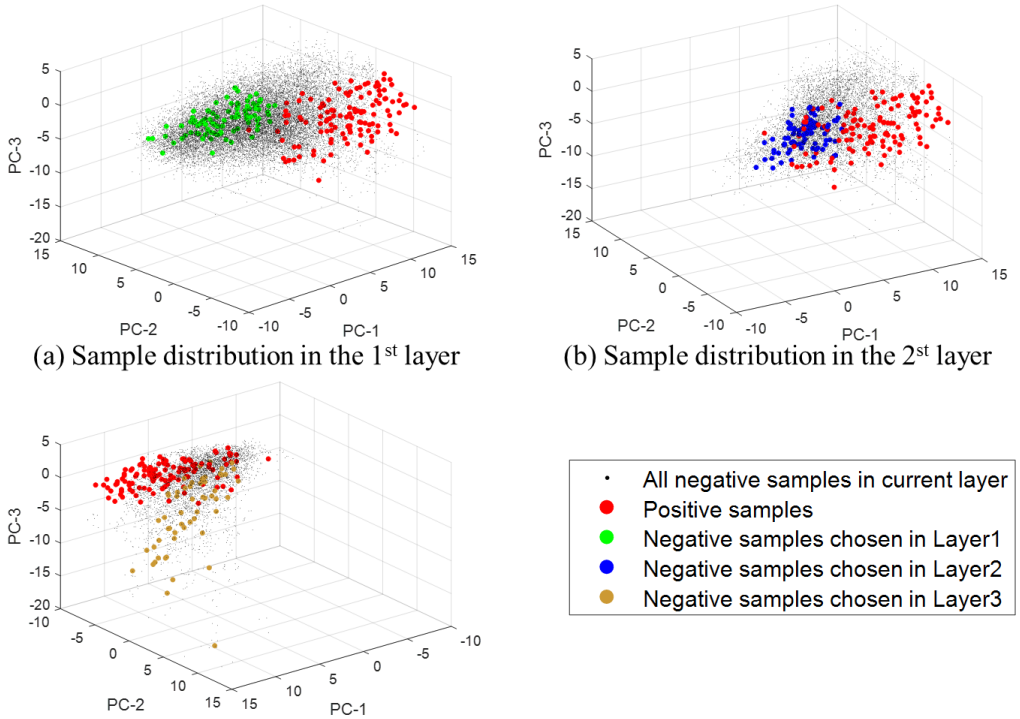
Although the morphological sifting method performs well on most types of lesions, it is still not quite as effective in coping with some architectural distortions that have vague or scattered spicules, and lack significant central densities. An architectural distortion (AD) is defined as an interruption of normal breast pattern with lines, often appearing as a star-shaped distortion, with no definite mass visible (Moreira et al., 2012). Figure 21 shows two cases of AD from INbreast and DoD BCRP. Figure 21 (a) shows an AD with visible central densities that is well segmented and detected by the proposed algorithm. Figure 21 (b) shows an AD that has no obvious (very small) focal area, and scattered spicules in various angles. The segmentation of the AD in Figure 21 (b) only captures a small density near the centre of the AD, instead of identifying the centre of the lesion precisely or contouring the whole lesion.

This is mainly due to the fact that there are not as many ‘strokes’ passing the morphological sifters in an area where there are no obvious central densities and the spicules are scattered. However, the frequency of this type of architectural distortion is relatively low. There are around 5 ADs appearing in this pattern only from DoD BCRP, which our method has difficulties to generate relatively satisfactory segmentations. The proposed algorithm is supposed to be a general model to extract majority types of masses, while these characteristic ADs are relatively rare and may need methods specially designed to identify (Rangayyan et al., 2010). There are methods designed to detect line structures that can be used to detect architectural distortions. Study (Zwiggelaar et al., 2004) compared four linear structure detecting approaches in mammographic images (line operator, orientated bins, Gaussian derivatives and ridge detector). Study (Matsubara et al., 2015) adopted direction and background filters to extract the line structures to detect ADs. Rami Ben-Ari *et al* extracted ROIs by sampling the parenchymal tissues (segmented by an unsharp mask filter) where ADs are likely to be found (Ben-Ari et al., 2017). However, it is still unclear whether these methods would be effective on capturing ADs that have relatively faint, scattered spicules and no significant central densities.

5.2 Tackling class imbalance using self-grown CasRFs

After the generation of region candidates, there is a huge imbalance between the positive and negative samples. To cope with the class imbalance problem, we adopted an ensemble of random forests structured as a cascade. The novelty of the CasRFs lies in the under-sampling guided by probability-ranking and its ability to grow adaptively according to the skewness of the training data. As we have stated in Section1, many previous studies adopt naïve random under-sampling when training each individual base classifier. However, this type of random under-sampling could lead to the possibility that the positive and negative samples in the

subset may not be separable. If the subset is too challenging for the classifier, it could affect the performance of the whole cascade at an early stage. By using the probability-ranking based under-sampling, only the more discriminative samples (according to the probability ranking) are used to train the RF at a current layer. By doing this, the training process is not only stabilized, but also giving the RF a relatively less challenging training subset. Figure 19 shows an example of the negative samples chosen in the first three layers during the training process. It can be seen that the negative samples chosen in the three layers are the samples that are clearly separable from the positive samples. The selected negative samples start from being rather far away from the positive samples in layer 1 to being closer to the positive samples in layer 2 and 3.



(a) Sample distribution in the 1st layer (b) Sample distribution in the 2st layer

(c) Sample distribution in the 3st layer

Figure 19. Example of the negative samples selected to form the training subset with all the positive samples in layer 1, layer 2 and layer 3 during the training in one of the validations. The dimension of the sample data has been reduced by principal component analysis (PCA). The first three principal components are used in this figure.

The CasRFs also give the user control over the detection rate during the training since attaining high sensitivity is crucial in mass detection. To reach the ideal sensitivity set by the user during the training of each individual RF, the CasRFs gradually reduce the number of negative samples (starting from the more difficult ones to remove) in the initially balanced training set. By doing this, we in essence reverse the imbalance by letting the positive samples become the majority class and the RF will naturally focus more on the detection of positive samples, without any need to adjust weights to increase the cost of misclassifying the minority samples (Bria et al., 2016; Viola and Jones, 2002) or manipulate the threshold in the RF (Wei et al., 2015).

The CasRFs only use part of the negative samples during the whole training process, since after building the RF model at a current layer, the unused negative samples are put through the model and some of them that do not contain any useful information will be eliminated without the need to be used for training in the next layer. This can be observed in Figure 19, where there are fewer negative samples in Figure 19 (b) than in Figure 19 (a) since some of the non-selected negative samples have been discarded by the RF model built in layer 1. This will also reduce the computational cost especially when the training set is highly skewed.

Unlike studies (Baumann et al., 2013; Tang et al., 2012) which used fixed number of layers, the CasRFs can adapt to highly skewed training data without the need to set the number of layers in the cascade (i.e. the number of RFs), by setting an explicit stopping criteria. The CasRFs perform relatively well when facing a class imbalance ratio of 1:160 in this work, and also achieved satisfactory results in the previous study where the imbalance ratio can be as high as 1:2000 (Min et al., 2017), while in other imbalance learning related studies, the imbalance ratio is around 1:6 (Wei et al., 2015), 3:10 (Viola and Jones, 2002) and 1:20 (Kozegar et al., 2013) respectively. The feature extraction and training can be carried out on a

regular desktop without the need of using graphic processing units like deep learning based studies (Dhungel et al., 2017; Ribli et al., 2017).

The design of the CasRFs was firstly introduced in our preliminary work (Min et al., 2017). However, during the training in this work, all the negative samples are used to initialize the first layer instead of random selecting a subset of negative samples as in (Min et al., 2017). And the samples from all scales are put together as the training set instead of training the CasRFs on each individual scale as in (Min et al., 2017), thus we could get a more sufficient number of positive samples (masses). If trained on each individual scale, the system achieves an average sensitivity of 0.90 at a FP rate of 1.45FPs/image at the default cut-off value of 0.5. This shows that in our case, having more positive samples in the training set does benefit the sensitivity, however also slightly increase the FP rate. We also compare the classification performance of the self-grown CasRFs with RUSBoost (Seiffert et al., 2010). The RUSBoost uses decision tree as base learner, with the maximal number of decision splits per tree set as the number of training instances. For the ensemble structure, the number of learning circles is set as 1000, and the learning rate for shrinkage is set as 0.1. Figure 20 shows the FROC curves of the proposed self-grown CasRFs and RUSBoost. The PAUC is 0.91 and 0.89 for CasRFs and RUSBoost respectively. The CasRFs also take less time to classify the samples than RUSBoost. For instance, given a certain testing image represented by its feature matrix, it takes approximately 4s for CasRFs and 14s for RUSBoost. The better overall performance and less execution time make the CasRFs more favourable than RUSBoost.

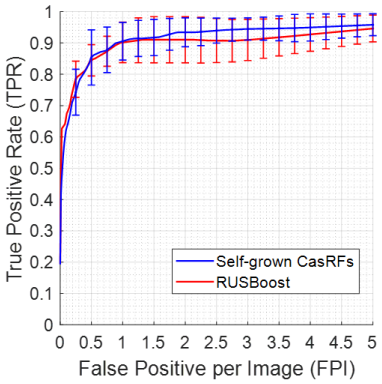


Figure 20. The FROC curves of the self-grown CasRFs and the RUSBoost.

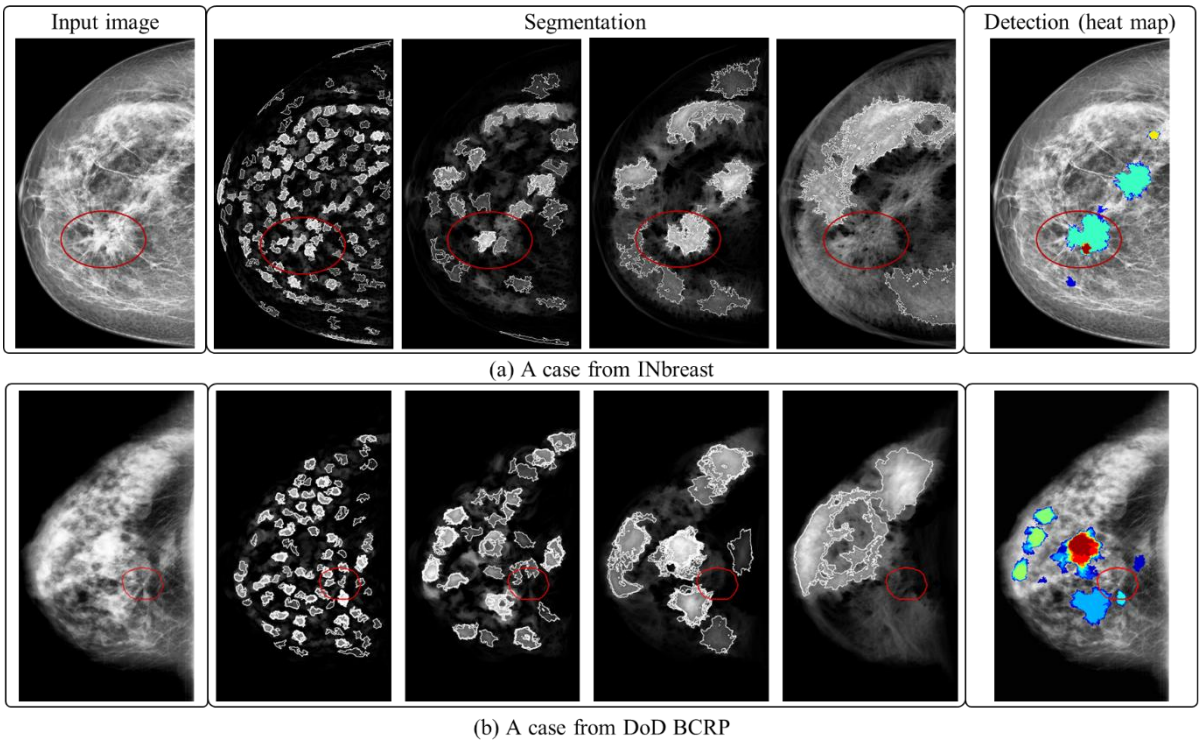


Figure 21. Two examples of the segmentation affected by architectural distortions. (a) is from INbreast and it has visible central densities within the distortion. (b) is from DoD BCRP and it contains a spiculated distortion with scattered spicules and no central density. The red lines indicate the ground truth.

5.3 Visualizing the detection and segmentation results

After classifying the samples using CasRFs, each patch sample was assigned a probability score by summing up the scores generated from all the RFs in the cascade. These scores are used in generating heat maps as shown in Figure 12 where only the positive patches are

marked. The heat map is a very useful tool to visualize the confidence of the detection and can be used as the final detection output. Figure 16 (a), (b) analyse the relations between TPR, FPI and cut-off value of CasRFs (before label fusion). To further clean up the outlines of the detections, we adopted a simple label fusion method to merge the overlapped regions by choosing the patch with the highest contrast against the background. Comparing Figure 16 (a) and (c), it can be noticed that the label fusion method does have the possibility of sacrificing the sensitivity sometimes as TPI drops within cut-off value 0.4~0.45 in validation 4. This happens due to the incorrect merging of overlap regions. An alternative fusion method is to choose the patch with the highest predicted probability. However, the average DSI calculated under this method is 0.84±0.10, which is slightly lower than the one under contrast analysis. We believe there are more sophisticated ways to merging the overlap raw detections such as expectation-maximization algorithm (Warfield et al., 2004) and atlas-based method (Langerak et al., 2010).

5.4 Performance evaluation

Since full-field digital mammographic (FFDM) has replaced the screen-film due to the higher image quality (de Munck et al., 2016) and DoD BCRP lacks an accurate ground truth, we mainly evaluated the system's performance on INbreast and only provided the performance on DoD BCRP as a supplemental result. Table 1 shows the performance comparison between this method and previous publications evaluated on INbreast and DoD BCRP. Our new method yields competitive results in both mass detection and segmentation. The validation approach on INbreast by (Eltonsy et al., 2007) is not known, and therefore, it could not be determined if our method's performance is certainly more favourable. Previous study (Dhungel et al., 2015a) shows the best mass detecting performance using deep learning and fixed two layers of RFs. However, it only locates the suspicious regions using bounding

boxes, without a precise segmentation of the regions. (Dhungel et al., 2016) and (Dhungel et al., 2017) present both detection and segmentation results. In these studies, the suspicious regions are located using method from the previous study (Dhungel et al., 2015a), then lesions are segmented from the background using another deep learning based algorithm (Dhungel et al., 2015b). Compared with (Dhungel et al., 2016) and (Dhungel et al., 2017), our method achieves a higher average sensitivity and DSI at a slightly lower FP rate as shown in Table 1. It is worth noting that our method has a much simpler structure with only one stage of machine learning compared with the multiple learning stages adopted in previous studies, each having very high computational complexity (Dhungel et al., 2015a, 2016; Dhungel et al., 2017). We also provide a comparative guide to the mass detection performances of two commercial systems, Cyclopus CAD (CyclopusCAD Ltd, Palermo, Italy), SecondLook (iCAD Inc. OH, USA), reported in study (Cascio et al., 2014). Evaluated on a FFDM dataset, CyclopusCAD achieved a sensitivity of 76.9% at 0.73 FPs per image, and iCAD yielded a sensitivity of 61.5% at 0.28 FPs per image for breast mass detection. Figure 22 shows the FROC curves of the commercial systems and the proposed method. Note that the FPI is solely calculated on normal images since the FPI of the two commercial systems in Figure 22 was only calculated on healthy cases (Cascio et al., 2014). It can be seen that the FROC curve of the proposed method in Figure 22 is above the ones of the two commercial systems. However, we still cannot be certain that our method performs better than these commercial systems since they are not tested on the same data.

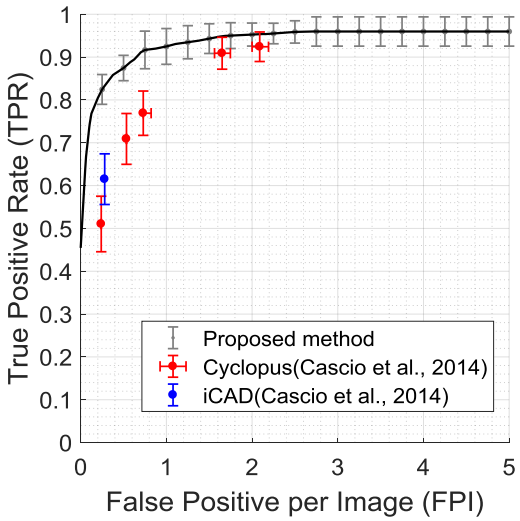


Figure 22. FROC curves of the proposed method on normal images and two commercial systems reported in study (Cascio et al., 2014).

As to the performance on DoD BCRP, the MMS using linear structuring elements does show advantages in identifying spiculated lesions (except for some architectural distortions as discussed in section 5.1) compared to the disk-shape structuring elements as presented in (Min et al., 2017). In DoD BCRP, the majority of the lesions are spiculated lesions in irregular shapes, and the rest are architectural distortions and lesions in oval or round shapes. The missed lesions are either spiculated in irregular shapes or ADs. Only 50% (9/18, overlap ratio > 0.7) of the ADs are detected in the testing set from DoD BCRP. The low sensitivity on ADs is not unusual since study (Baker et al., 2003) reported only fewer than 40% of the ADs (image sensitivity) from the dataset used for evaluation were detected by two most widely available commercial CAD systems. The reason the proposed method does not perform well on ADs is that ADs are more characteristic and rarer compared with other types of lesions in the evaluation dataset. Moreover, the proposed method is still more oriented towards detecting lesions with visible central densities. In the future, we would like to explore the detection of architectural distortions, especially the ADs that have no central densities with scattered and thin radiating lines as discussed in 5.1.

6 Conclusion

In this paper, we presented a mammographic breast mass detection and segmentation system using multi-scale morphological sifting and a self-grown, dynamically multi-layered, cascaded random forests algorithm. The MMS using linear structuring elements is able to extract lesions in various shapes and sizes while removing the background tissue. The self-grown CasRFs provide an effective and adaptive solution to imbalanced learning by adopting a probability-ranking based under-sampling approach and distributing the class imbalance throughout the layers of the cascade. In general, the proposed system achieves promising results in both detection and segmentation on digital and screen-film mammograms. In the further work, we would like to explore the possibility of combining the MMS based region candidate generation method with CNN to avoid hand-selecting features and exhaustive searching for CNN itself. We would consider adopting transfer learning (West et al., 2007) since the size of the training set might be limited for building CNN from scratch. Further study is also needed to improve the system's ability to detect architectural distortions.

7 Acknowledgement

We would like to thank Dr. Neeraj Dhungel for providing the INbreast validation sets used in this work.

Hang Min is supported by the China Scholarship Council.

8 References

DoD BCRP Spiculated Mass Detection Evaluation Data. http://marathon.csee.usf.edu/Mammography/DDSM/BCRP/bcrp_mass_01.html

Andreea, G.I., Pegza, R., Lascu, L., Bondari, S., Stoica, Z., Bondari, A., 2011. The role of imaging techniques in diagnosis of breast cancer. J. Curr. Health Sci 37, 241-248.

Backes, A.R., Bruno, O.M., 2008. A new approach to estimate fractal dimension of texture images, International Conference on Image and Signal Processing. Springer, pp. 136-143.

Baker, J.A., Rosen, E.L., Lo, J.Y., Gimenez, E.I., Walsh, R., Soo, M.S., 2003. Computer-aided detection (CAD) in screening mammography: sensitivity of commercial CAD systems for detecting architectural distortion. Am. J. Roentgenol. 181, 1083-1088.

Baumann, F., Ehlers, A., Vogt, K., Rosenhahn, B., 2013. Cascaded Random Forest for Fast Object Detection, Scandinavian Conference on Image Analysis. Springer, pp. 131-142.

Beller, M., Stotzka, R., Müller, T.O., Gemmeke, H., 2005. An example-based system to support the segmentation of stellate lesions, Bildverarbeitung für die Medizin 2005. Springer, pp. 475-479.

Ben-Ari, R., Akselrod-Ballin, A., Karlinsky, L., Hashoul, S., 2017. Domain specific convolutional neural nets for detection of architectural distortion in mammograms, 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), . IEEE, pp. 552-556.

Bowyer, K., Kopans, D., Kegelmeyer, W., Moore, R., Sallam, M., Chang, K., Woods, K., 1996. The digital database for screening mammography, Third international workshop on digital mammography, p. 27.

Breiman, L., 2001. Random forests. Machine learning 45, 5-32.

Bria, A., Karssemeijer, N., Tortorella, F., 2014. Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications. Med. Image Anal. 18, 241-252.

Bria, A., Marrocco, C., Molinara, M., Tortorella, F., 2016. An effective learning strategy for cascaded object detection. Information Sciences 340-341, 17-26.

Cascio, D., Fauci, F., Iacomi, M., Raso, G., Magro, R., Castrogiovanni, D., Filosto, G., Ienzi, R., Vasile, M.S., 2014. Computer-aided diagnosis in digital mammography: comparison of two commercial systems. Imaging Med. 6, 13.

Cascio, D., Fauci, F., Magro, R., Raso, G., Bellotti, R., De Carlo, F., Tangaro, S., De Nunzio, G., Quarta, M., Forni, G., 2006. Mammogram segmentation by contour searching and mass lesions classification with neural network. IEEE Trans. Nucl. Sci. 53, 2827-2833.

Casti, P., Mencattini, A., Salmeri, M., Ancona, A., Mangeri, F., Pepe, M.L., Rangayyan, R.M., 2016. Contour-independent detection and classification of mammographic lesions. Biomedical Signal Processing and Control 25, 165-177.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321-357.

Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W., 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting, In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (Eds.), Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 107-119.

Chu, J., Min, H., Liu, L., Lu, W., 2015. A novel computer aided breast mass detection scheme based on morphological enhancement and SLIC superpixel segmentation. Med. Phys. 42, 3859-3869.

de Munck, L., de Bock, G.H., Otter, R., Reiding, D., Broeders, M.J.M., Willemse, P.H.B., Siesling, S., 2016. Digital vs screen-film mammography in population-based breast cancer screening: performance indicators and tumour characteristics of screen-detected and interval cancers. Br. J. Cancer 115, 517-524.

Dhungel, N., Carneiro, G., Bradley, A.P., 2015a. Automated Mass Detection in Mammograms using Cascaded Deep Learning and Random Forests, 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, pp. 1-8.

Dhungel, N., Carneiro, G., Bradley, A.P., 2015b. Deep structured learning for mass segmentation from mammograms, 2015 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2950-2954.

Dhungel, N., Carneiro, G., Bradley, A.P., 2016. The automated learning of deep features for breast mass classification from mammograms, International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 106-114.

Dhungel, N., Carneiro, G., Bradley, A.P., 2017. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med. Image Anal. 37, 114-128.

737 Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297-
738 302.
739 Dubitzky, W., Granzow, M., Berrar, D.P., 2007. Fundamentals of data mining in genomics and
740 proteomics. Springer Science & Business Media.
741 Eltonsy, N.H., Tourassi, G.D., Elmaghraby, A.S., 2007. A Concentric Morphology Model for the
742 Detection of Masses in Mammography. IEEE Trans. Med. Imaging 26, 880-889.
743 Ganesan, K., Acharya, U.R., Chua, C.K., Min, L.C., Abraham, K.T., Ng, K.-H., 2013. Computer-
744 aided breast cancer detection using mammograms: a review. IEEE Reviews in Biomedical
745 Engineering 6, 77-98.
746 Gonzalez, R.C., Woods, R.E., 2008. Digital image processing / Rafael C. Gonzalez, Richard E.
747 Woods, 3rd ed.. ed. Harlow : Pearson/Prentice Hall, Harlow.
748 Görgel, P., Sertbas, A., Ucan, O.N., 2013. Mammographical mass detection and classification using
749 Local Seed Region Growing-Spherical Wavelet Transform (LSRG-SWT) hybrid scheme. Comput.
750 Biol. Med. 43, 765-774.
751 Gromet, M., 2008. Comparison of computer-aided detection to double reading of screening
752 mammograms: review of 231,221 mammograms. Am. J. Roentgenol. 190, 854-859.
753 He, H., Ma, Y., 2013. Imbalanced learning: foundations, algorithms, and applications. John Wiley &
754 Sons.
755 Jaiantilal, A., 2013. *Randomforest-matlab* [Online]. Available:
756 https://github.com/jrderuiter/randomforest-matlab [Accessed].
757 Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D., 2011. Global cancer statistics. CA
758 Cancer J. Clin. 61, 69-90.
759 Kang, P., Cho, S., 2006. EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems,
760 International Conference on Neural Information Processing. Springer, pp. 837-846.
761 Karssemeijer, N., te Brake, G.M., 1996. Detection of stellate distortions in mammograms. IEEE Trans.
762 Med. Imaging 15, 611-619.

763 Kharel, N., Alsadoon, A., Prasad, P., Elchouemi, A., 2017. Early diagnosis of breast cancer using
764 contrast limited adaptive histogram equalization (CLAHE) and Morphology methods, Information
765 and Communication Systems (ICICS), 2017 8th International Conference on. IEEE, pp. 120-124.
766 Kimori, Y., 2011. Mathematical morphology-based approach to the enhancement of morphological
767 features in medical images. J. Clin. Bioinforma. 1, 33.
768 Kozegar, E., Soryani, M., Minaei, B., Domingues, I., 2013. Assessment of a novel mass detection
769 algorithm in mammograms. J. Cancer Res. Ther. 9, 592.
770 Langerak, T.R., van der Heide, U.A., Kotte, A.N., Viergever, M.A., Van Vulpen, M., Pluim, J.P.,
771 2010. Label fusion in atlas-based segmentation using a selective and iterative method for
772 performance level estimation (SIMPLE). IEEE Trans. Med. Imaging 29, 2000-2008.
773 Li, H., Wang, Y., Liu, K.R., Lo, S.-C., Freedman, M.T., 2001. Computerized radiographic mass
774 detection. I. Lesion site selection by morphological enhancement and contextual segmentation.
775 IEEE Trans. Med. Imaging 20, 289-301.
776 Liao, P.-S., Chen, T.-S., Chung, P.-C., 2001. A fast algorithm for multilevel thresholding. J. Inf. Sci.
777 Eng. 17, 713-727.
778 Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18-22.
779 Liu, L., Li, J., Wang, Y., 2015. Breast mass detection with kernelized supervised hashing, 2015 8th
780 International Conference on Biomedical Engineering and Informatics (BMEI). IEEE, pp. 79-84.
781 Martins, L.d.O., Cardoso de Paiva, A., Corrêa Silva, A., Braz Junior, G., Gattass, M., 2009. Detection
782 of Masses in Digital Mammograms using K-Means and Support Vector Machine. ELCVIA.
783 Electronic letters on computer vision and image analysis 8, 39-50.
784 Matsubara, T., Ito, A., Tsunomori, A., Hara, T., Muramatsu, C., Endo, T., Fujita, H., 2015. An
785 automated method for detecting architectural distortions on mammograms using direction analysis
786 of linear structures, 2015 37th Annual International Conference of the IEEE Engineering in
787 Medicine and Biology Society (EMBC). IEEE, pp. 2661-2664.

Min, H., Chandra, S.S., Dhungel, N., Crozier, S., Bradley, A.P., 2017. Multi-scale mass segmentation for mammograms via cascaded random forests, 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, pp. 113-117.

Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S., 2012. INbreast: toward a full-field digital mammographic database. Acad. Radiol. 19, 236-248.

Murthy, S.N., Kumar, A., Sheshadri, H., 2013. Mass Detection and Classification using Machine Learning Techniques in Digital Mammograms. International Journal of Computer Applications 76.

Oliver, A., Freixenet, J., Martí, J., Pérez, E., 2010. A review of automatic mass detection and segmentation in mammographic images. Med. Image Anal. 14, 87-110.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62-66.

Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K., 1987. Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing 39, 355-368.

Rangayyan, R.M., Banik, S., Desautels, J.L., 2010. Computer-aided detection of architectural distortion in prior mammograms of interval cancer. J. Digit. Imaging 23, 611-631.

Reichel, U.D., Cole, J., 2016. Entrainment analysis of categorical intonation representations. Proc. P&P, Munich, Germany.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2017. Detecting and classifying lesions in mammograms with Deep Learning. arXiv preprint arXiv:1707.08401.

Schnabel, J.A., Giger, M.L., Karssemeijer, N., 2013. Breast Image Analysis for Risk Assessment, Detection, Diagnosis, and Treatment of Cancer. Annu. Rev. Biomed. Eng. 15, 327-357.

Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A., 2010. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 40, 185-197.

Soulami, K.B., Saidi, M.N., Tamtaoui, A., 2017. A CAD System for the Detection of Abnormalities in the Mammograms Using the Metaheuristic Algorithm Particle Swarm Optimization (PSO), In:

El-Azouzi, R., Menasche, D.S., Sabir, E., De Pellegrini, F., Benjillali, M. (Eds.), Advances in Ubiquitous Networking 2: Proceedings of the UNet'16. Springer Singapore, Singapore, pp. 505-517.

Szymkiewicz, D., 1934. Une conlribution statistique à la géographie floristique. Acta Societatis Botanicorum Poloniae 11, 249-265.

Tan, M., Qian, W., Pu, J., Liu, H., Zheng, B., 2015. A new approach to develop computer-aided detection schemes of digital mammograms. Phys. Med. Biol. 60, 4413.

Tang, D., Liu, Y., Kim, T.-K., 2012. Fast Pedestrian Detection by Cascaded Random Forest with Dominant Orientation Templates, BMVC, pp. 1-11.

te Brake, G.M., Karssemeijer, N., Hendriks, J.H., 2000. An automatic method to discriminate malignant masses from normal tissue in digital mammograms1. Phys. Med. Biol. 45, 2843.

Varela, C., Tahoces, P.G., Méndez, A.J., Souto, M., Vidal, J.J., 2007. Computerized detection of breast masses in digitized mammograms. Comput. Biol. Med. 37, 214-226.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001. . IEEE, pp. I-511-I-518 vol. 511.

Viola, P., Jones, M., 2002. Fast and robust classification using asymmetric adaboost and a detector cascade. Adv. Neural Inf. Process. Syst. 2, 1311-1318.

Wang, Y. 2006. *Hierarchical Masses Detection Algorithms Based on SVM in Mammograms.* Master's Thesis, Xidian University, China.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23, 903-921.

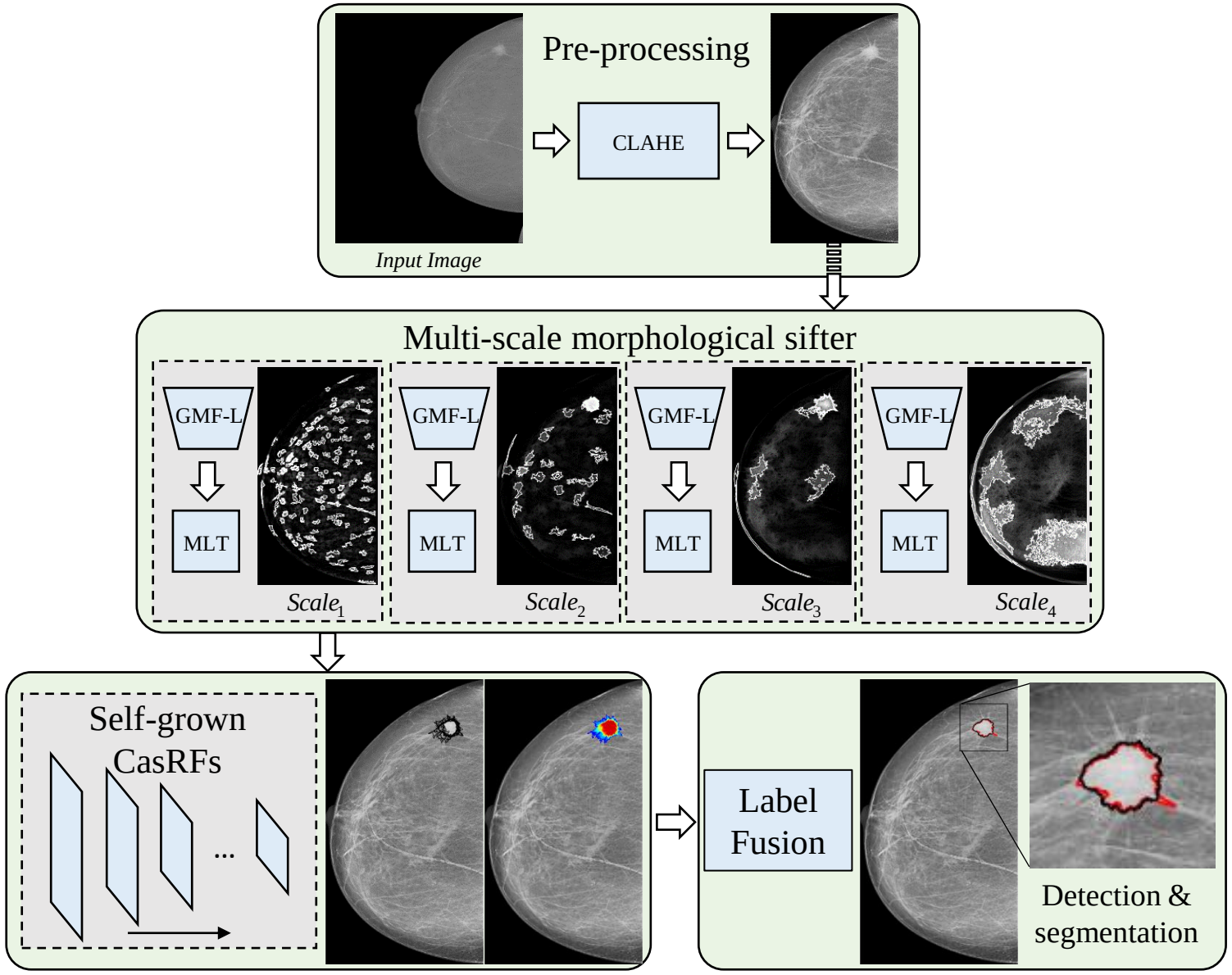
Wei, Z.-S., Yang, J.-Y., Shen, H.-B., Yu, D.-J., 2015. A cascade random forests algorithm for predicting protein-protein interaction sites. IEEE Trans. NanoBiosci. 14, 746-760.

West, J., Ventura, D., Warnick, S., 2007. Spring research presentation: A theoretical foundation for inductive transfer. Brigham Young University, College of Physical and Mathematical Sciences.

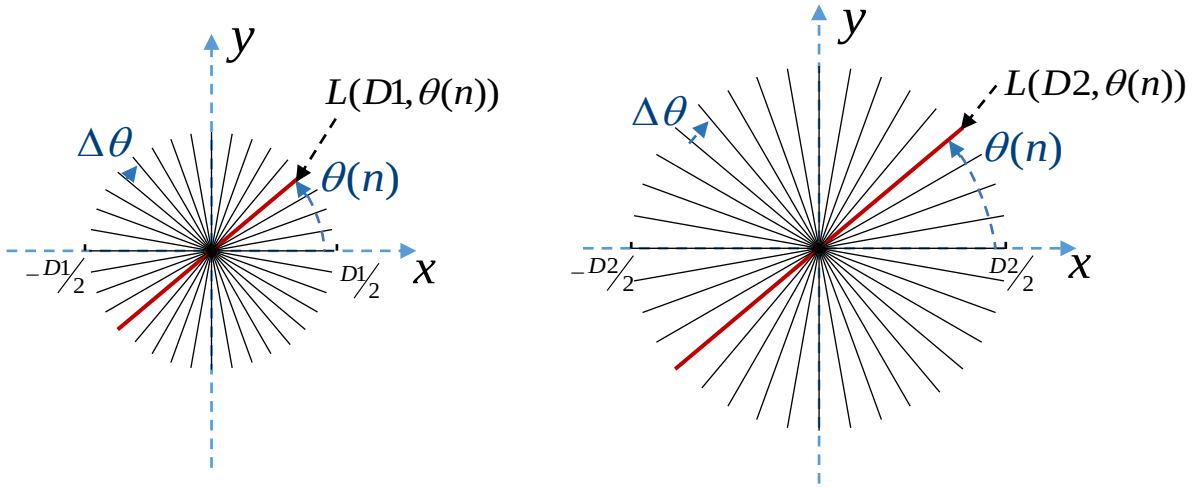
Zhang, Y.-D., Wang, S.-H., Liu, G., Yang, J., 2016. Computer-aided diagnosis of abnormal breasts in mammogram images by weighted-type fractional Fourier transform. Advances in Mechanical Engineering 8, 1687814016634243.

Zwiggelaar, R., Astley, S.M., Boggis, C.R., Taylor, C.J., 2004. Linear structures in mammographic images: detection and classification. IEEE Trans. Med. Imaging 23, 1077-1086.

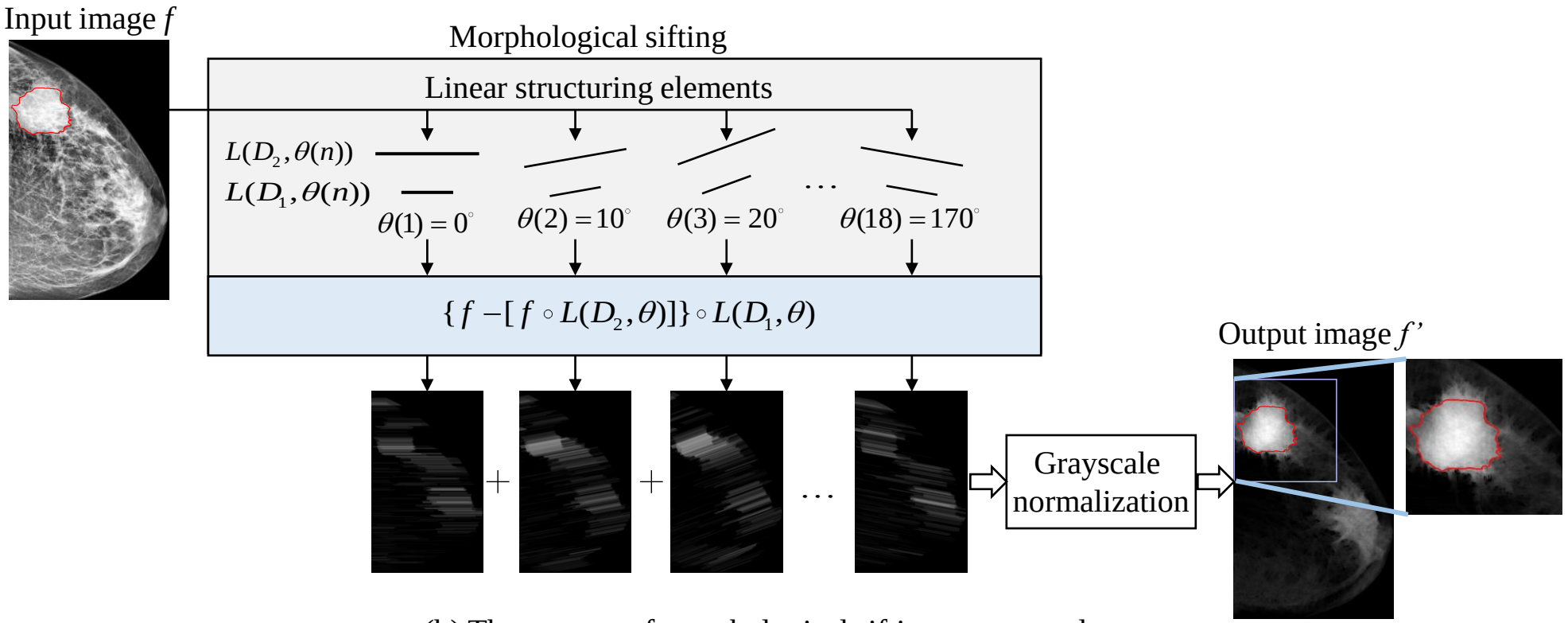
Figure


Pre-processing
CLAHE
Input Image
Multi-scale morphological sifter
GMF-L
MLT
Scale1
Scale2
Scale3
Scale4
Self-grown
CasRFs
...
Label
Fusion
Detection &
segmentation

Figure

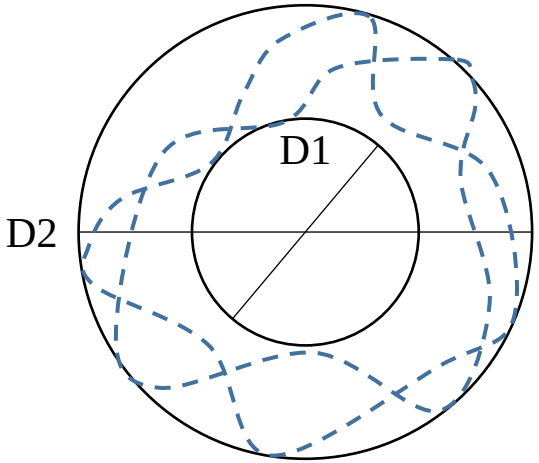


(a) Two sets of linear structuring elements. The lines marked in red represent a linear element pair $[L(D1,\theta(n)),L(D2,\theta(n))]$.

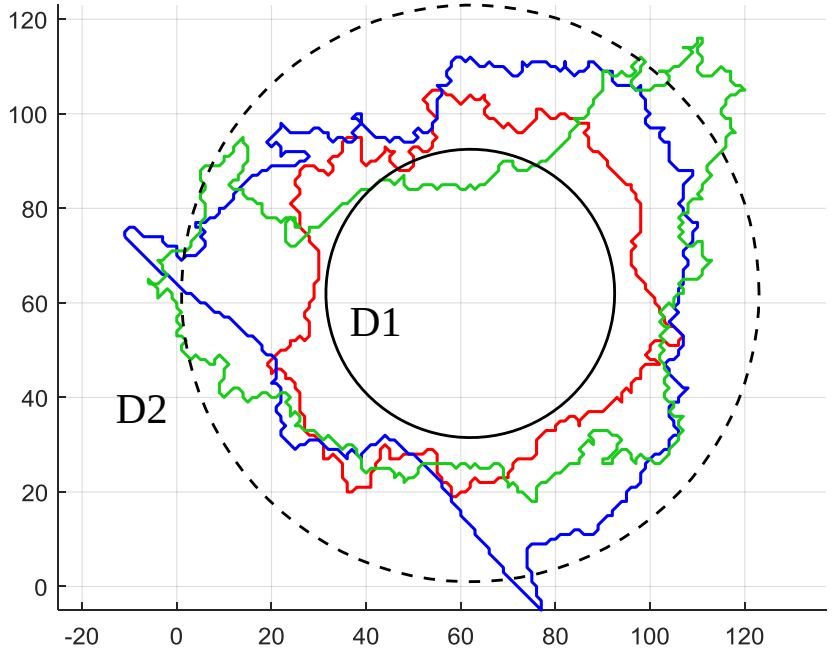


(b) The process of morphological sifting on one scale.

Figure

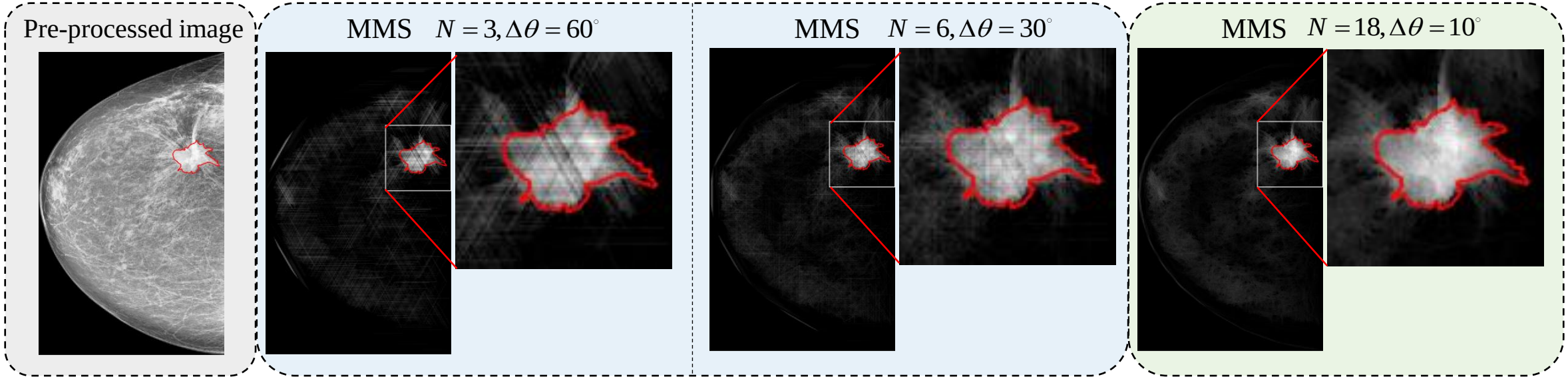


(a) Theoretical model of objects that MMS filters extract.

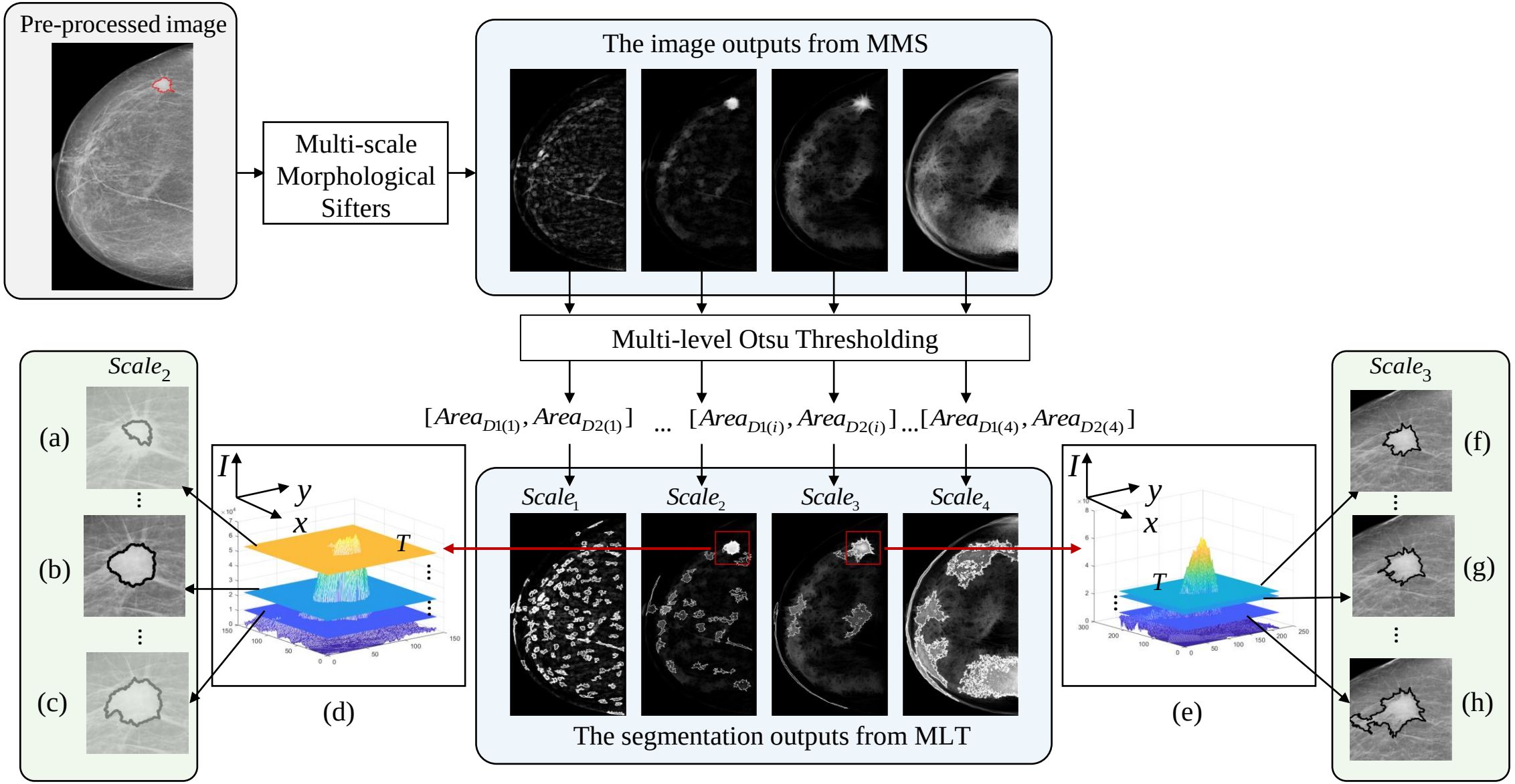


(b) Real masses extracted by MMS filters on one scale (scale3).

Figure


Pre-processed image
MMS $N = 3, \Delta\theta = 60^{\circ}$
MMS $N = 6, \Delta\theta = 30^{\circ}$
MMS $N = 18, \Delta\theta = 10^{\circ}$

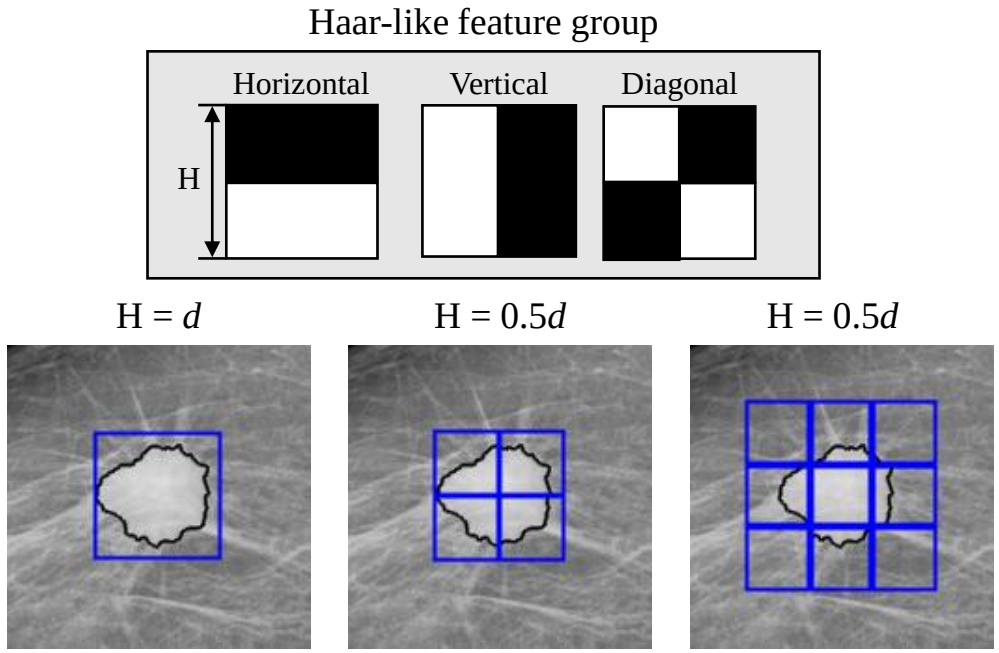
Figure



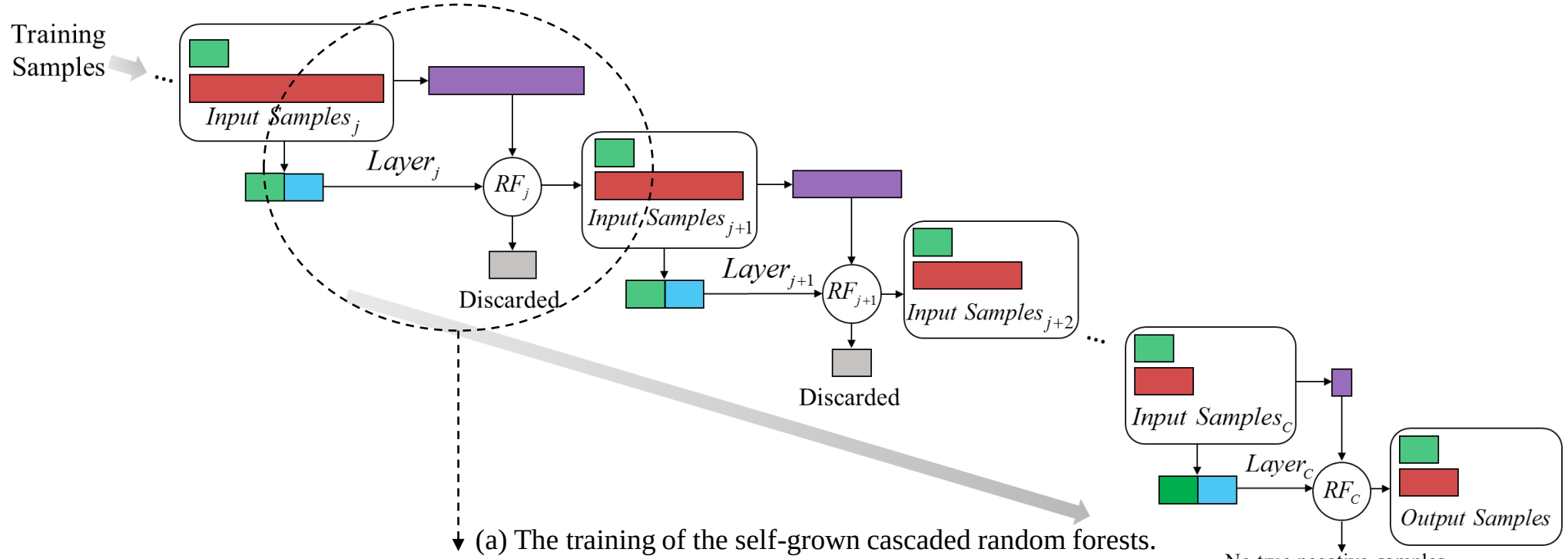
Segmentations of the mass from $Scale_2$.

Segmentations of the mass from $Scale_3$.

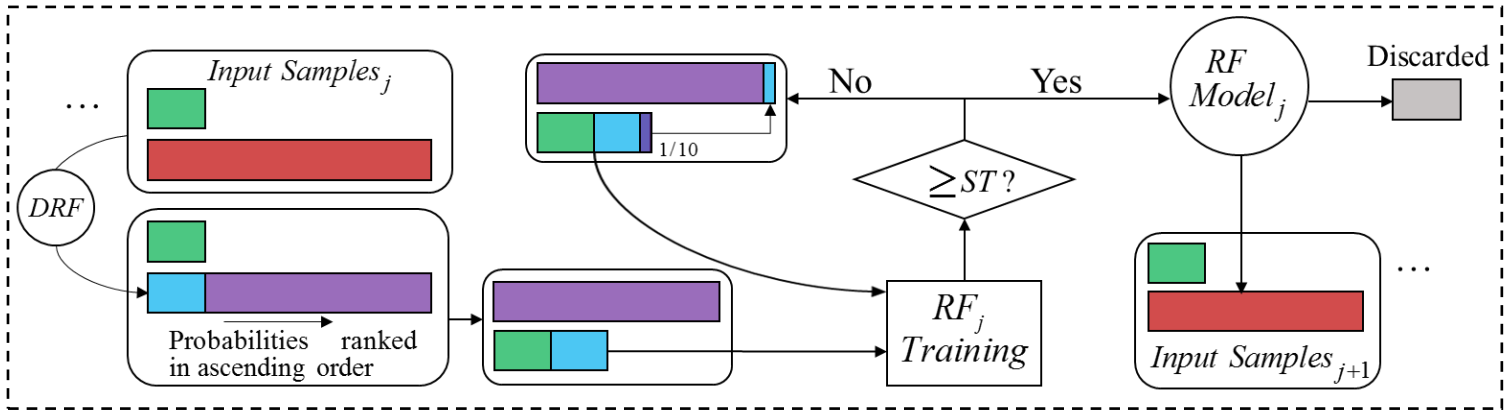
Figure


Haar-like feature group
Horizontal
Vertical
Diagonal
H
H = d
H = 0.5d
H = 0.5d

Figure



(a) The training of the self-grown cascaded random forests.

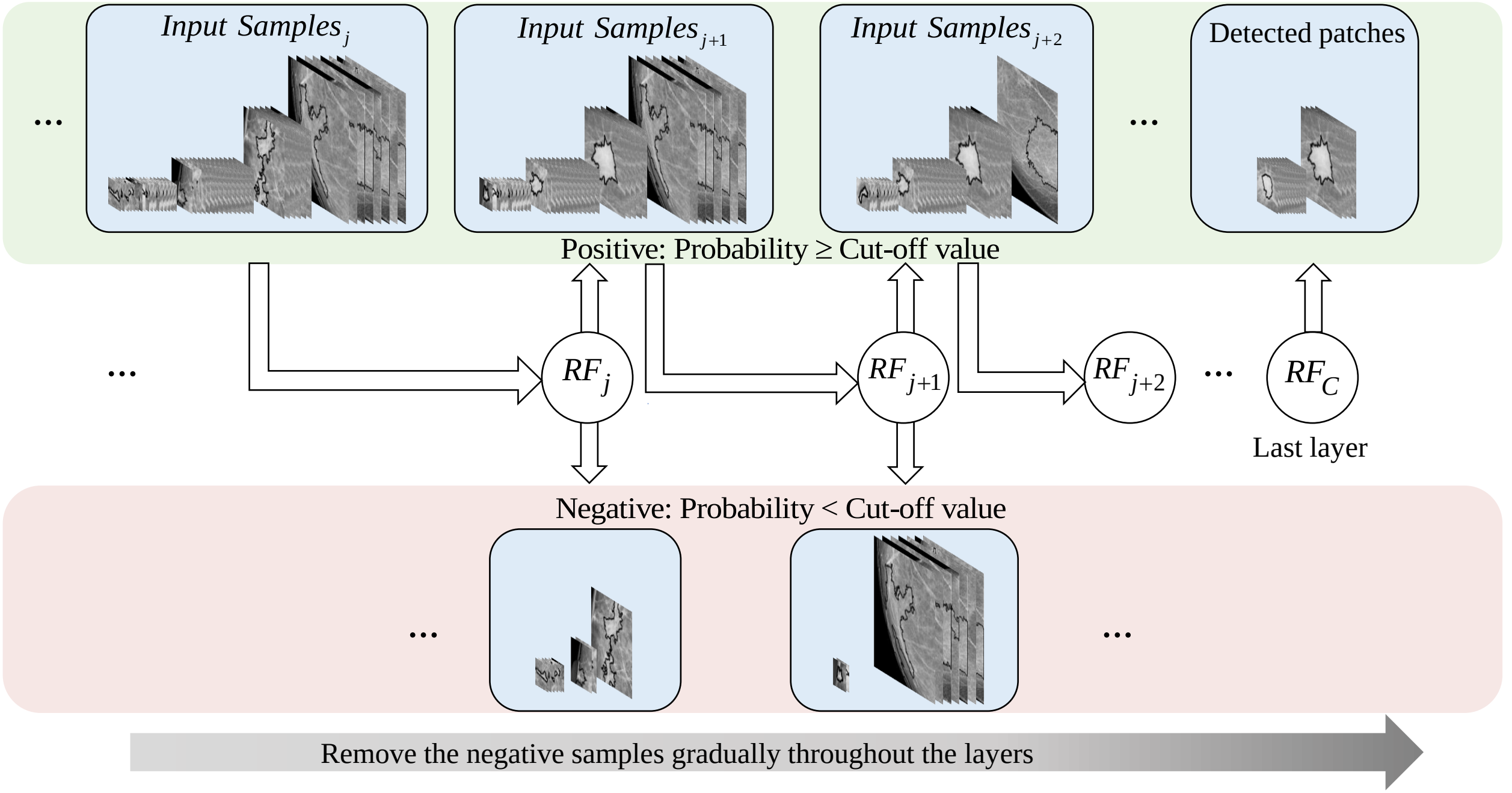


(b) The training of the RF at $Layer_j$.

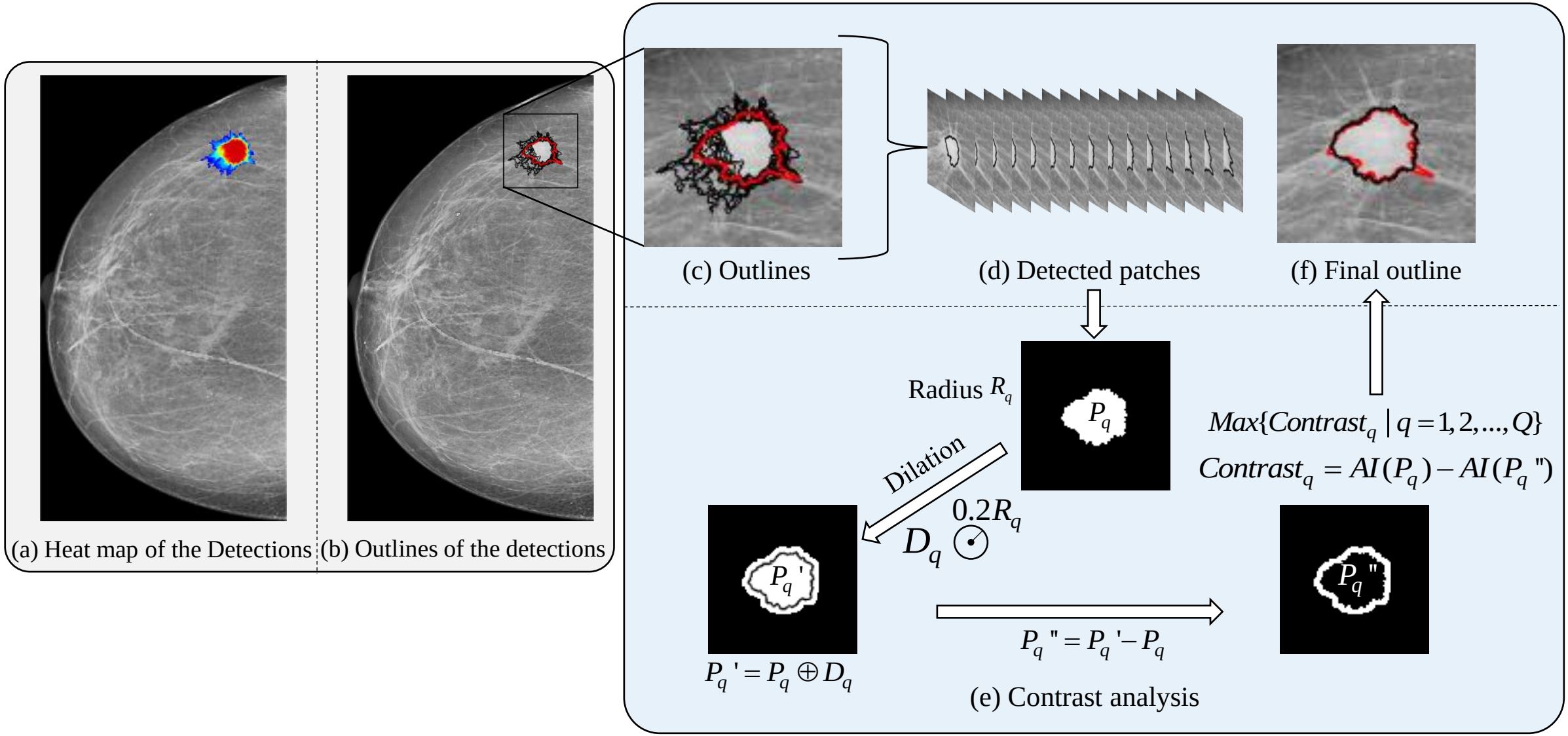
Positive samples Negative samples selected Negative samples left True negative False positive

ST Sensitivity threshold *DRF* RF with default settings

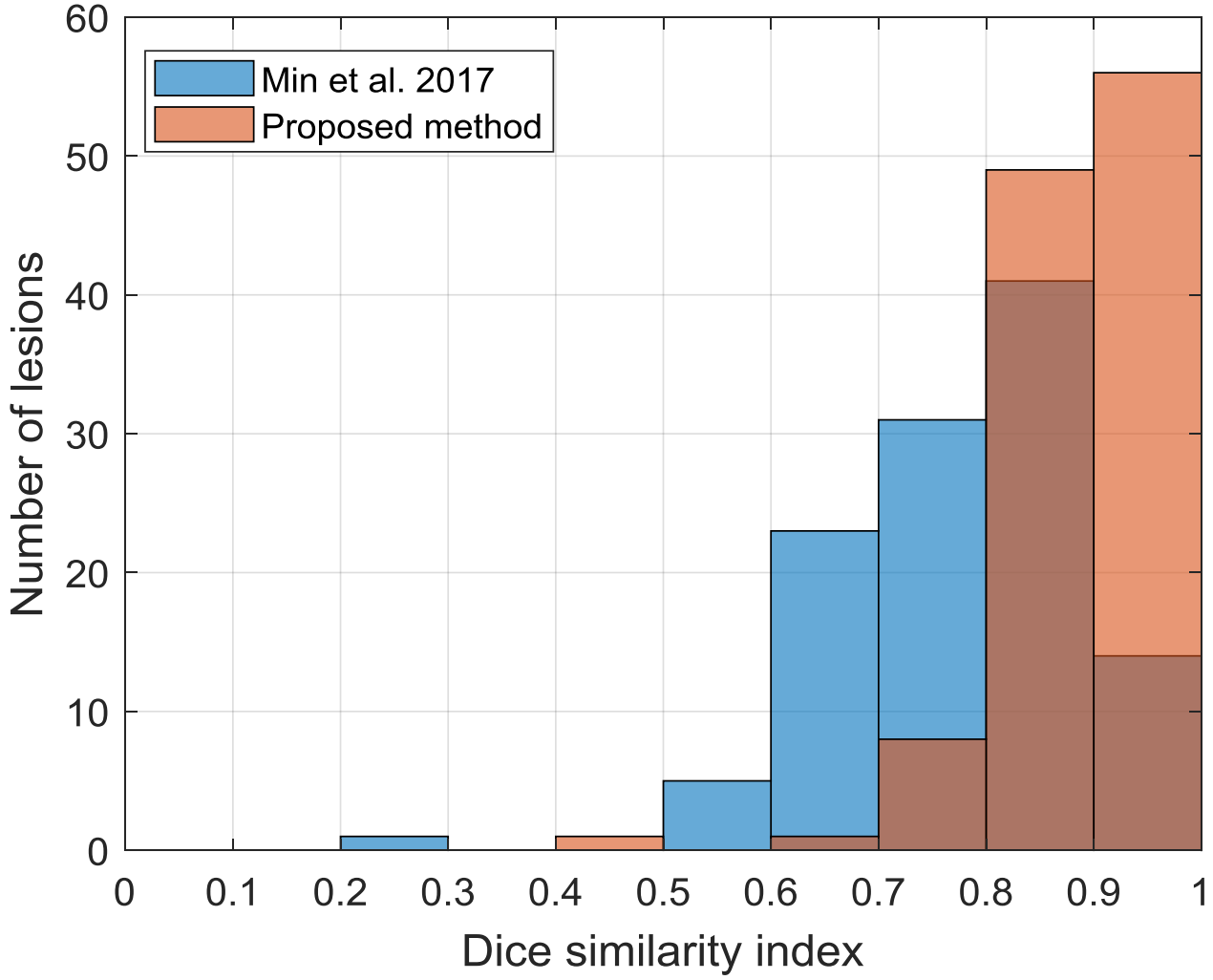
Figure


Input Samples j
Input Samples j+1
Input Samples j+2
Detected patches
Positive: Probability ≥ Cut-off value
RF j
RF j+1
RF j+2
RF C
Last layer
Negative: Probability < Cut-off value
Remove the negative samples gradually throughout the layers

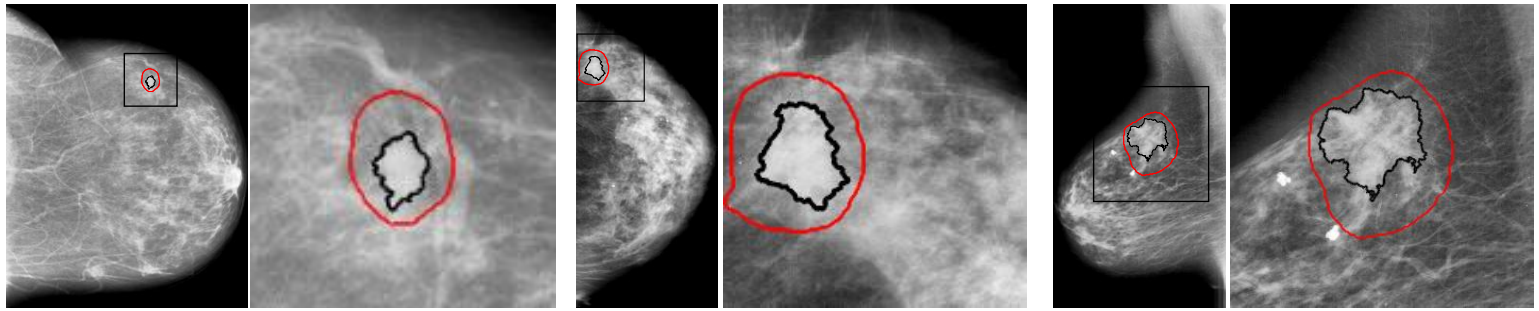
Figure


(a) Heat map of the Detections
(b) Outlines of the detections
(c) Outlines
(d) Detected patches
(f) Final outline
Radius R_q
P_q
Dilation
$0.2R_q$
D_q
P_q'
$P_q'' = P_q' - P_q$
$P_q'' $
$P_q' = P_q \oplus D_q$
$Max\{Contrast_q \mid q = 1, 2, ..., Q\}$
$Contrast_q = AI(P_q) - AI(P_q'')$
(e) Contrast analysis

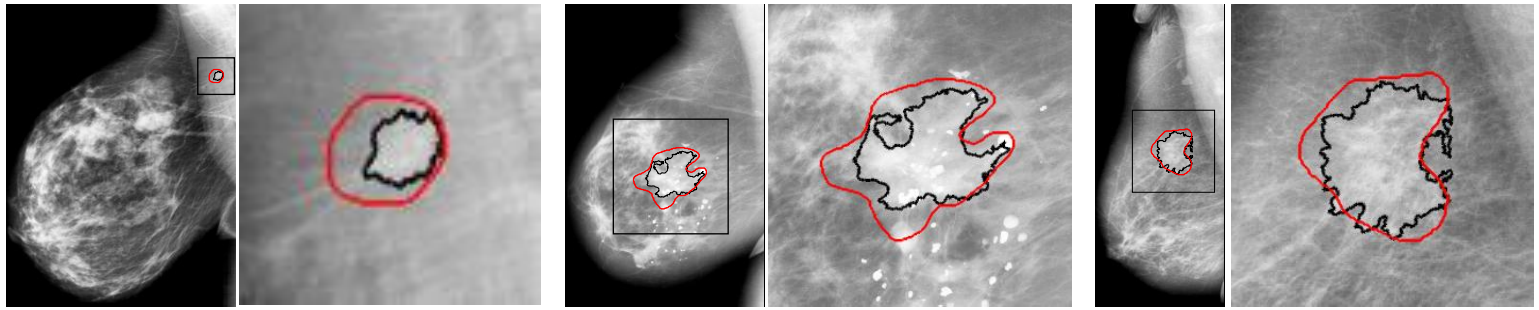
Figure


Min et al. 2017
Proposed method
Number of lesions
Dice similarity index
0
10
20
30
40
50
60
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure

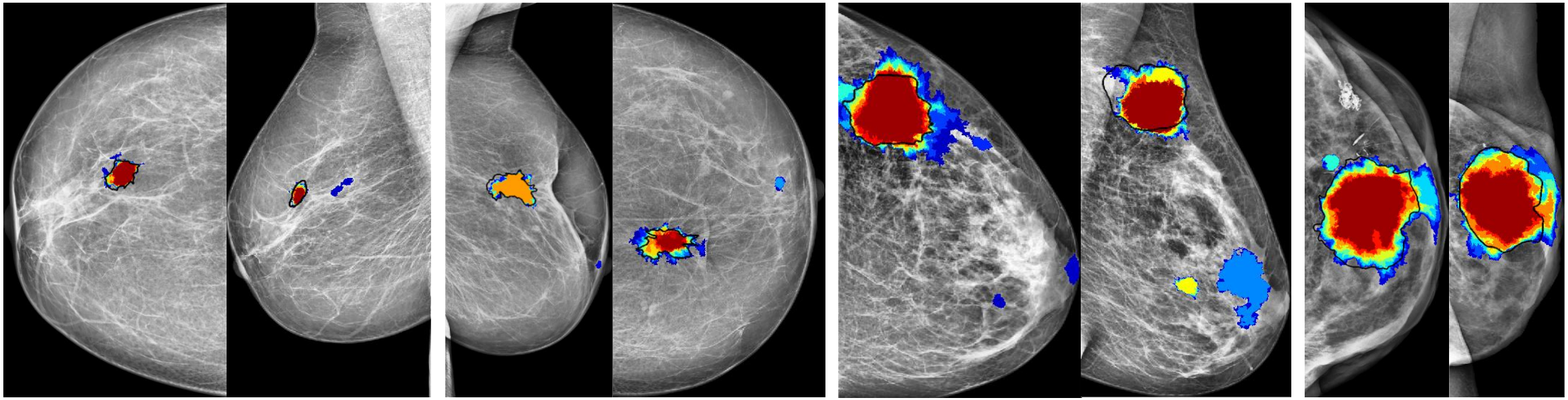


(a) Irregular, spiculated (b) Irregular, spiculated (c) AD, spiculated

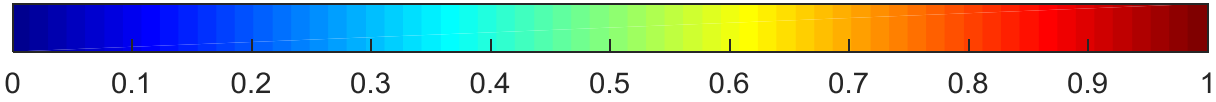


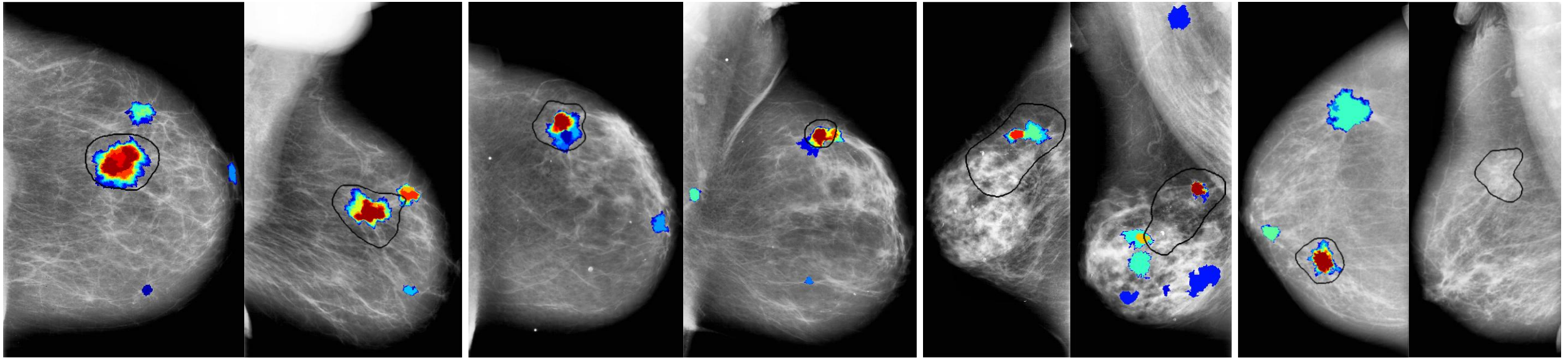
(d) Irregular, spiculated (e) Irregular, spiculated (f) AD, spiculated

Figure



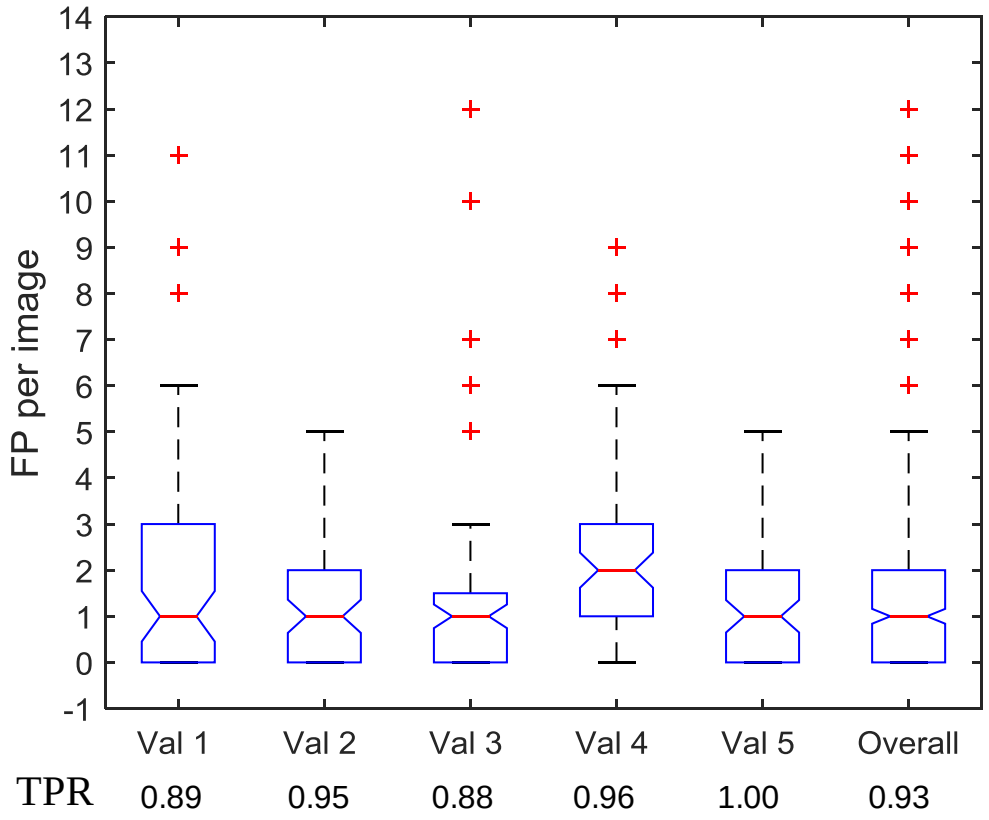
(a) Examples of heat maps from INbreast



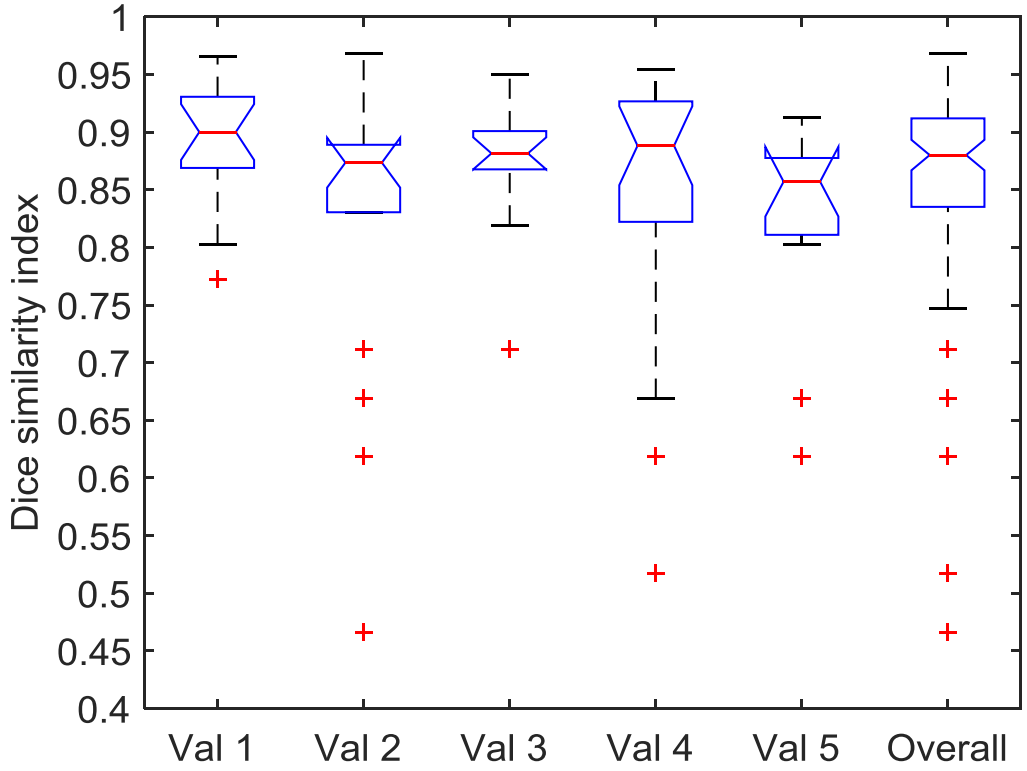


(b) Examples of heat maps from DoD BCRP

Figure

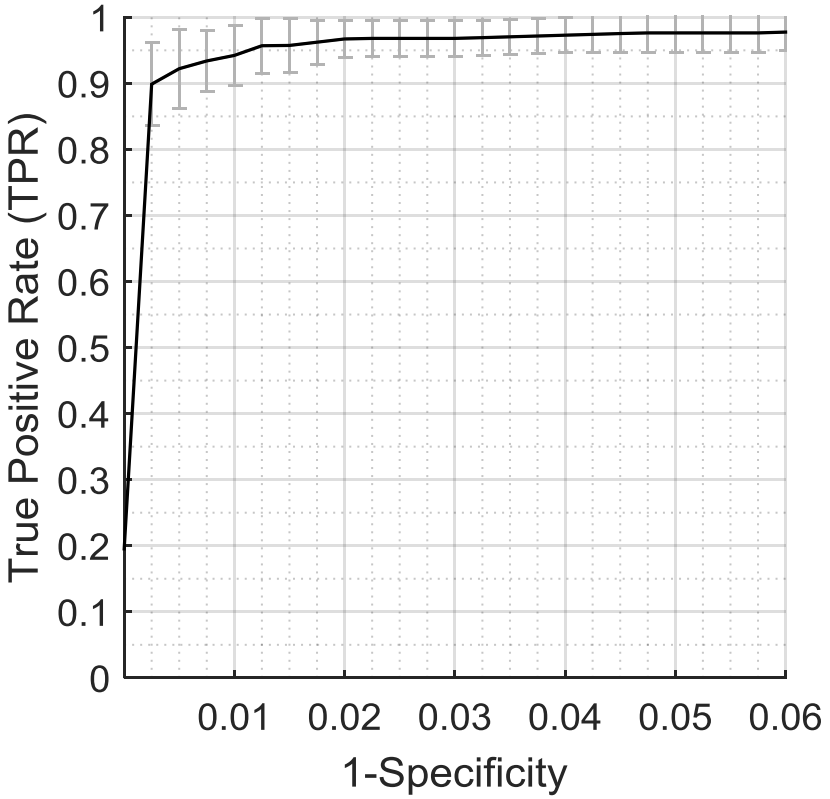


(a) Boxplot of FPI

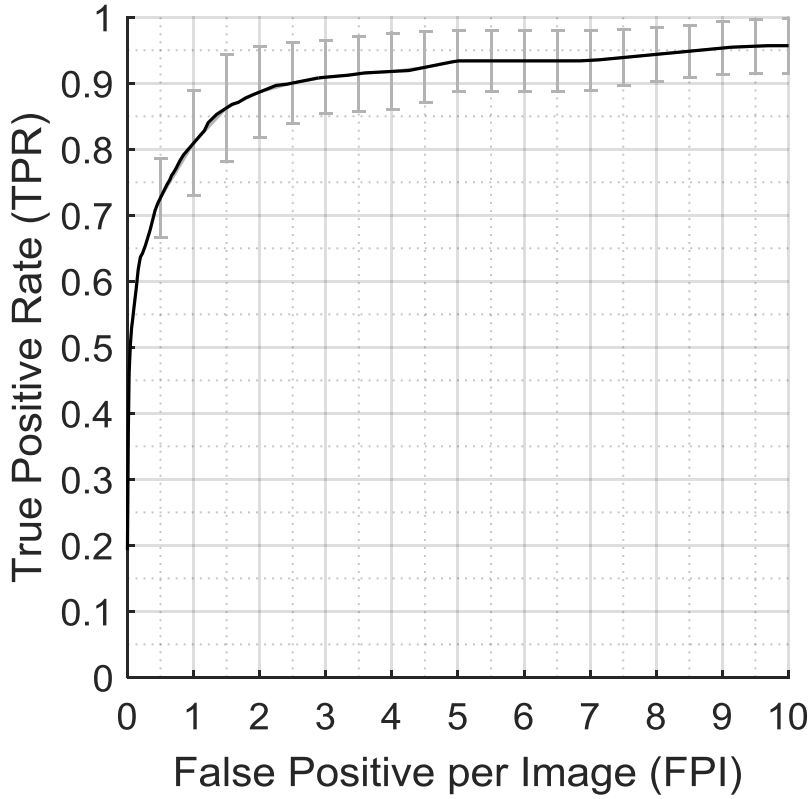


(b) Boxplot of DSI

Figure

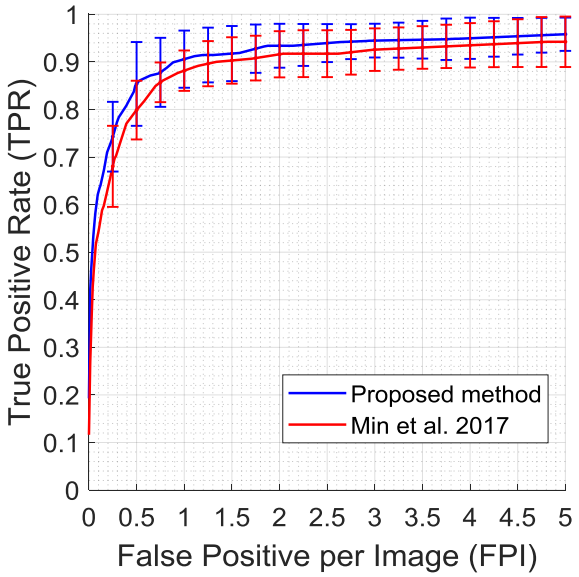


(a) ROC

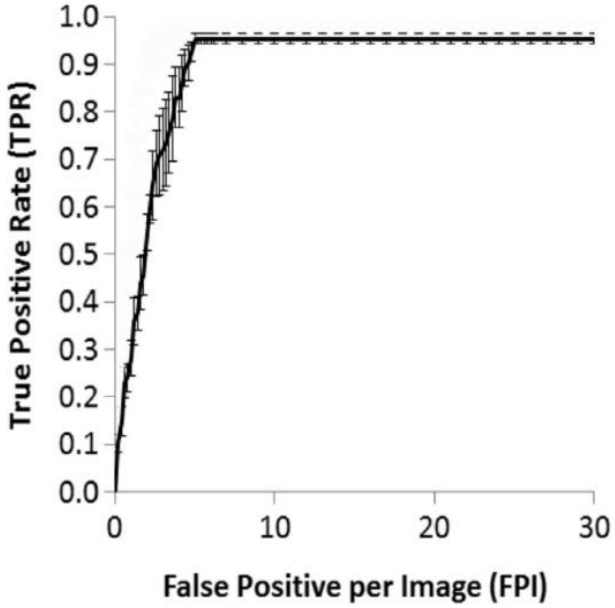


(b) FROC

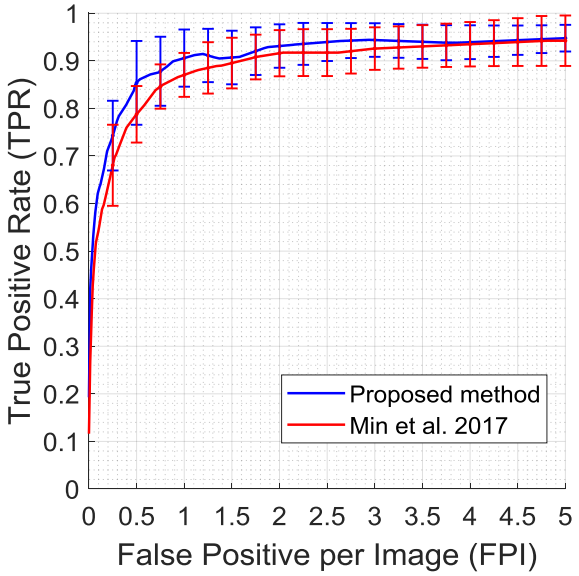
Figure



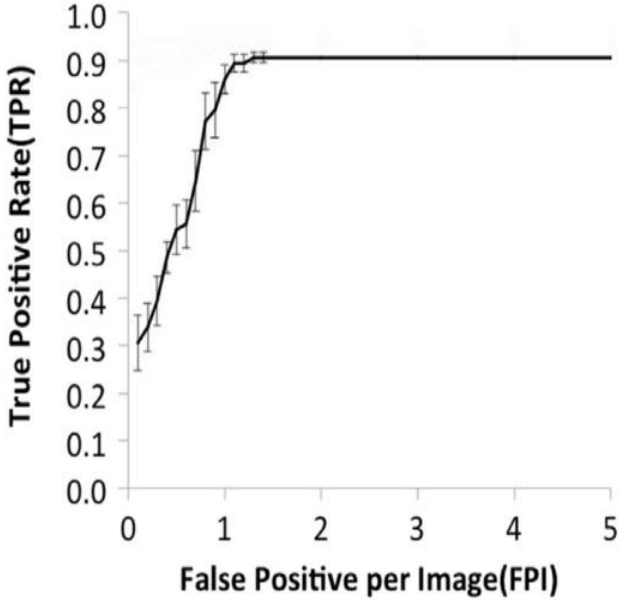
(a) FROC of the proposed method and Min et al. 2017 (*DT*=0.2)



(b) FROC of Dhungel et al. 2017 (*DT*=0.2)

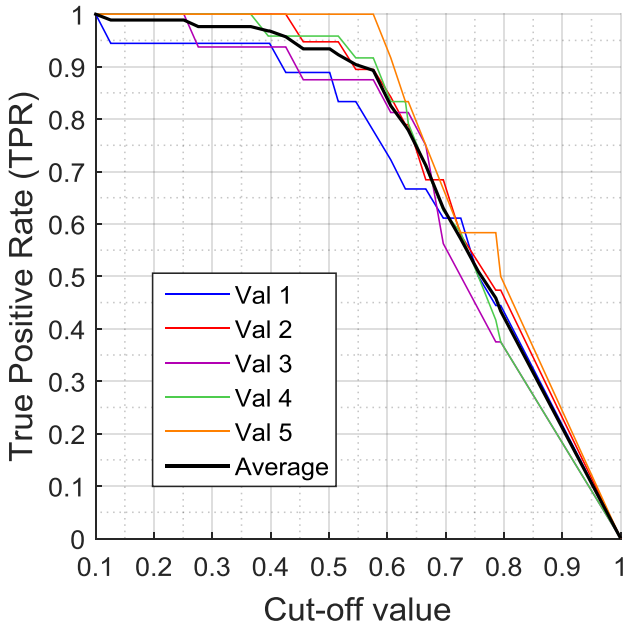


(c) FROC of the proposed method and Min et al. 2017 (*DT*=0.5)

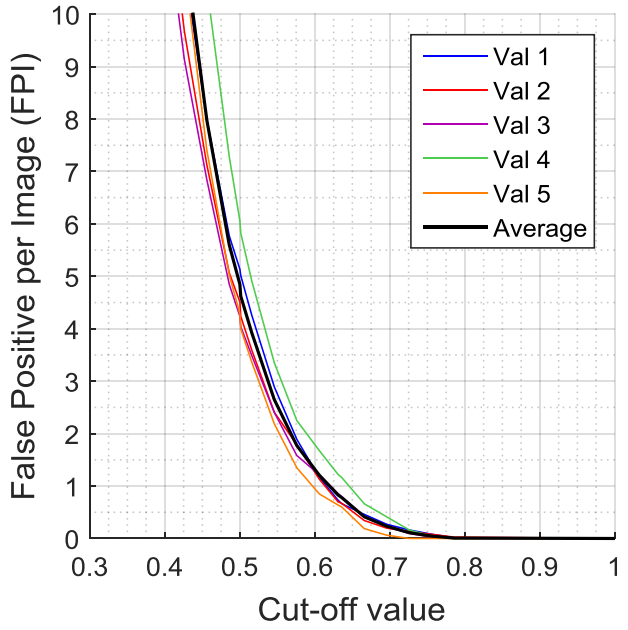


(d) FROC of Dhungel et al. 2017 (*DT*=0.5)

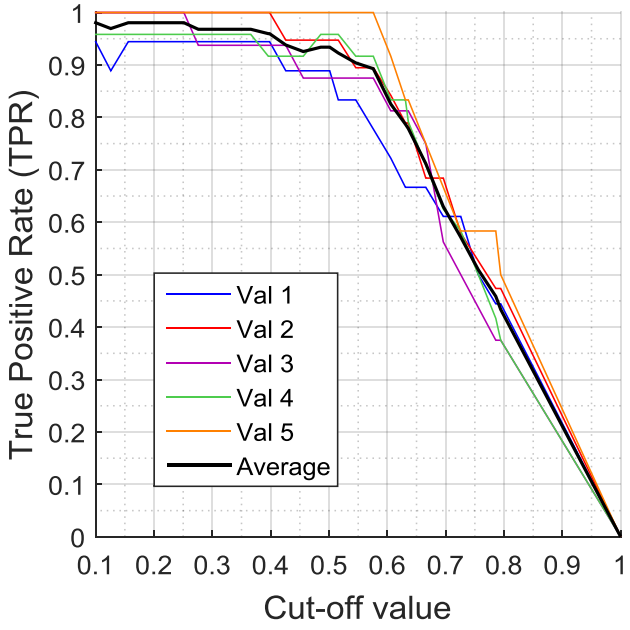
Figure



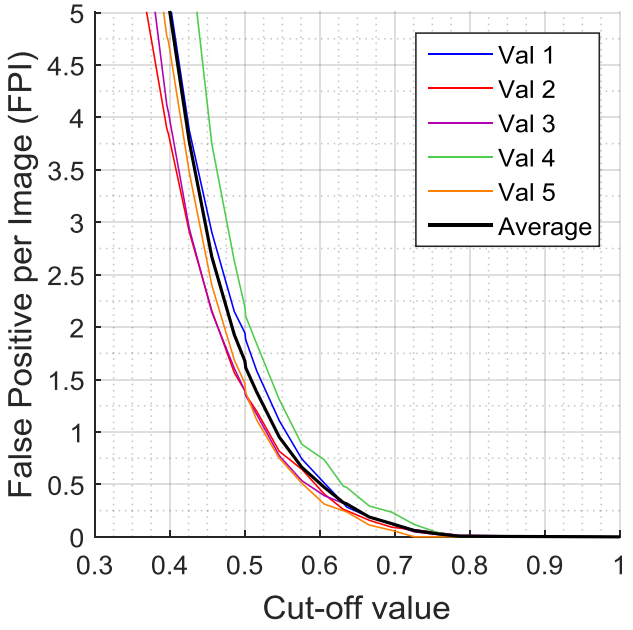
(a) TPR-Cut-off value cure before outline fusion



(b) FPI-Cut-off value cure before outline fusion

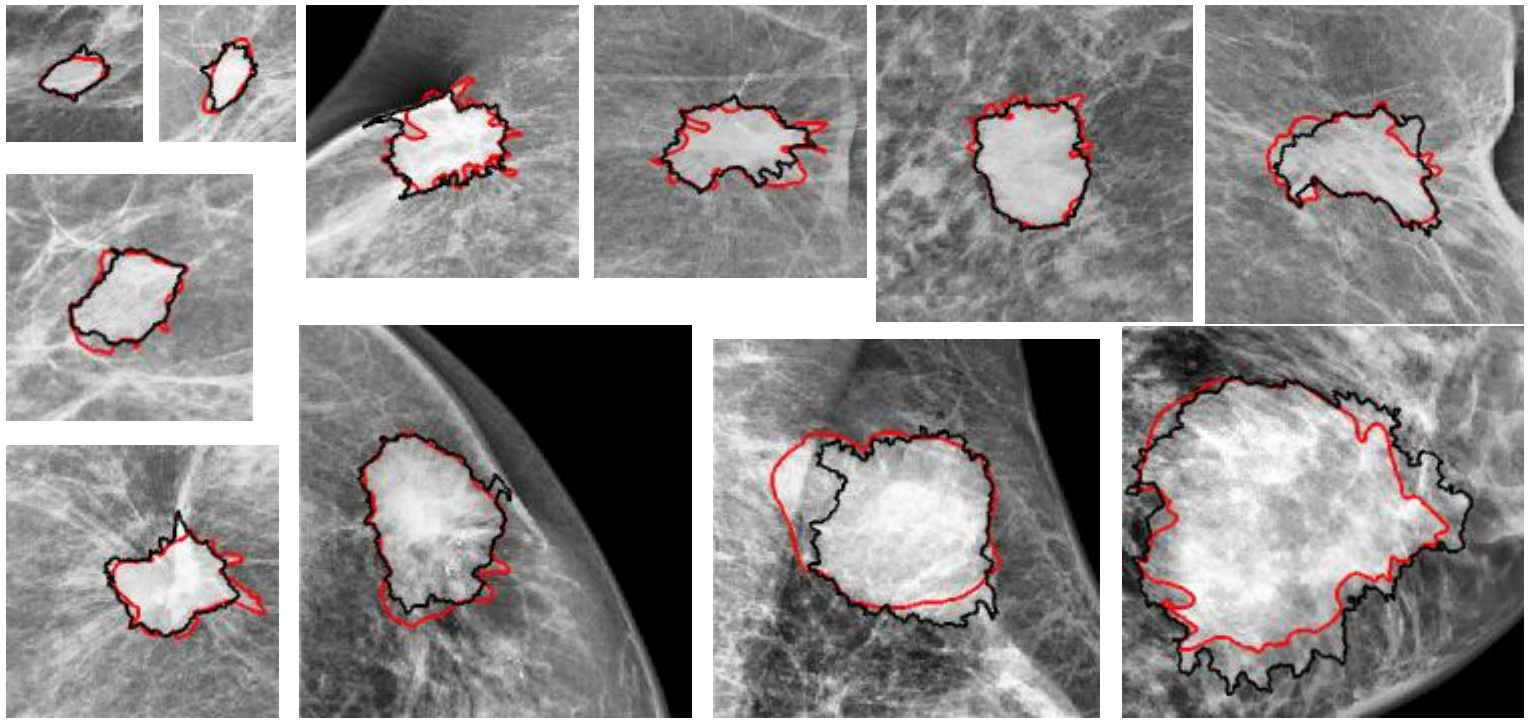


(c) TPR-Cut-off value cure after outline fusion

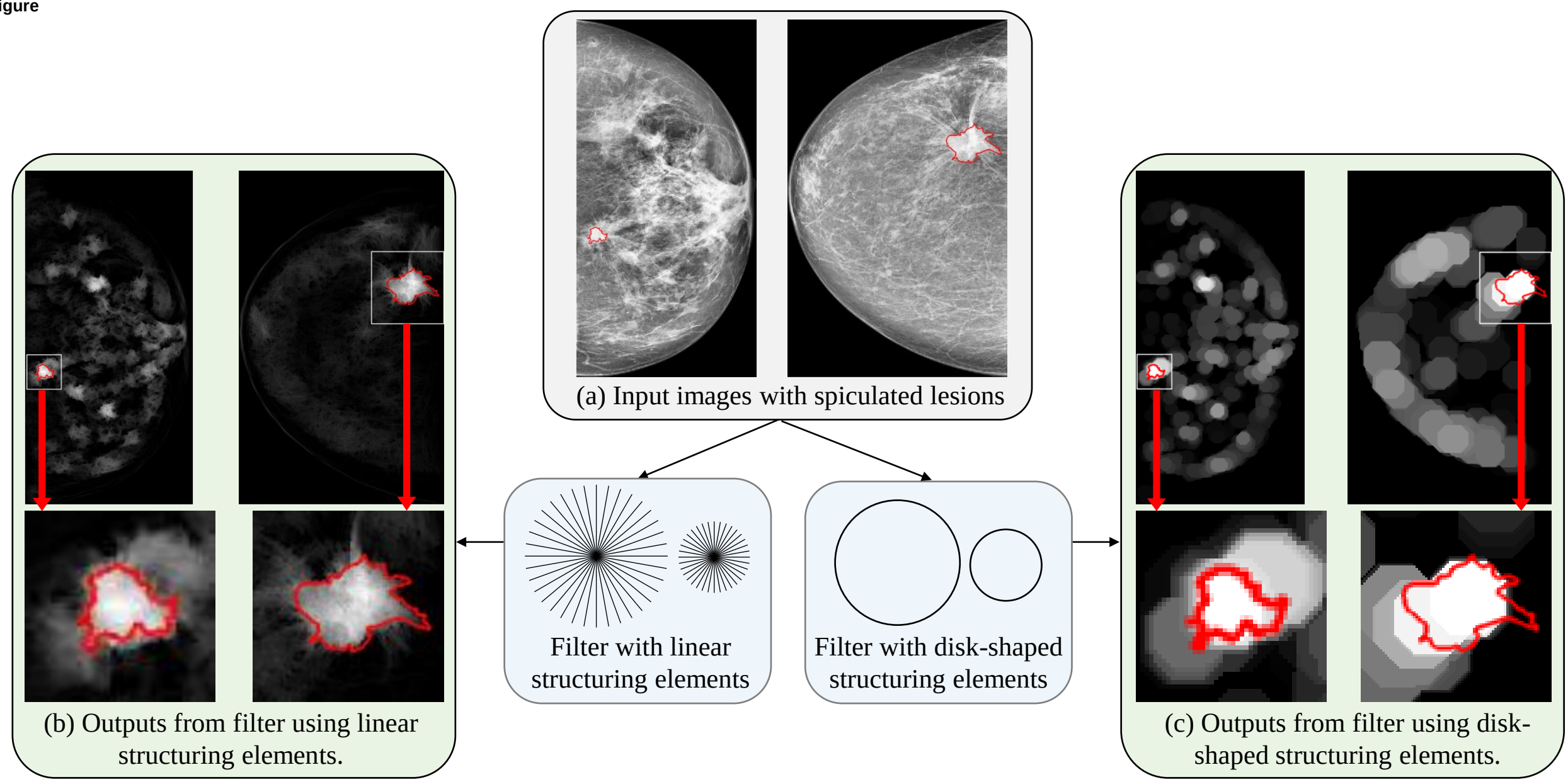


(d) FPI-Cut-off value cure after outline fusion

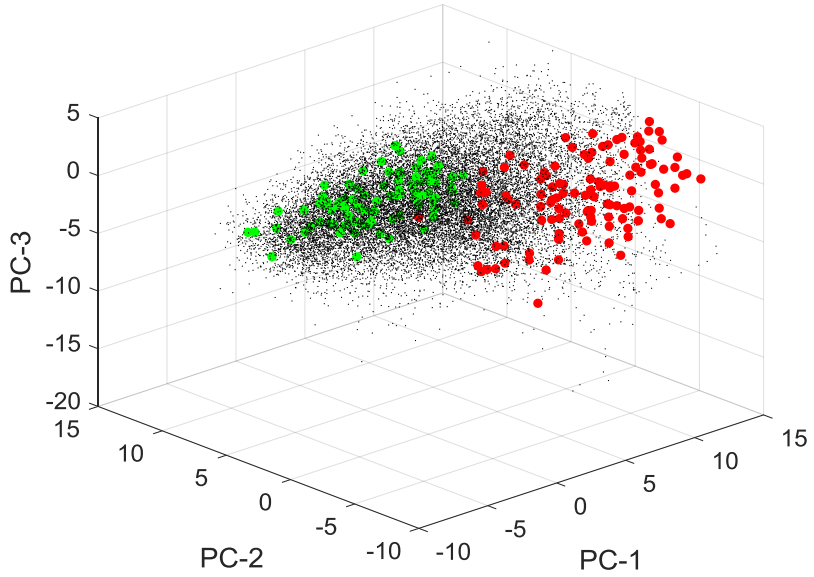
Figure



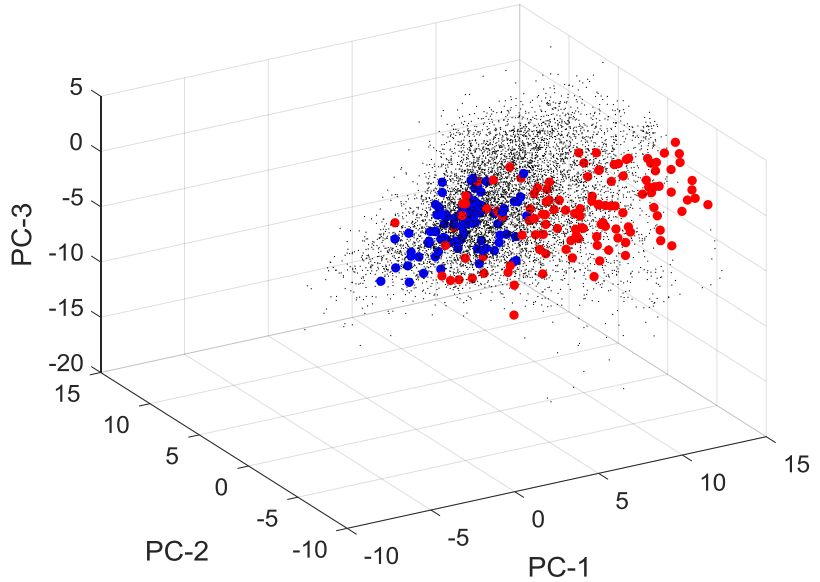
Figure


(a) Input images with spiculated lesions
Filter with linear structuring elements
Filter with disk-shaped structuring elements
(b) Outputs from filter using linear structuring elements.
(c) Outputs from filter using disk-shaped structuring elements.

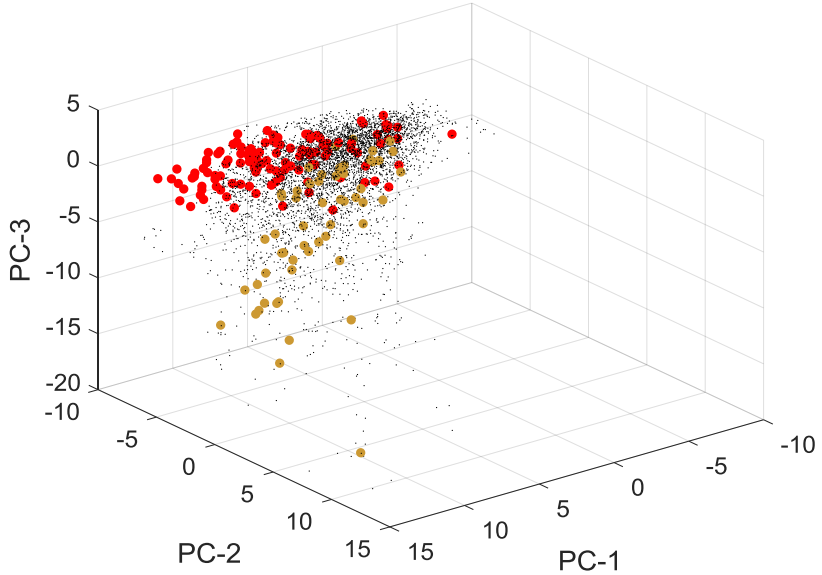
Figure



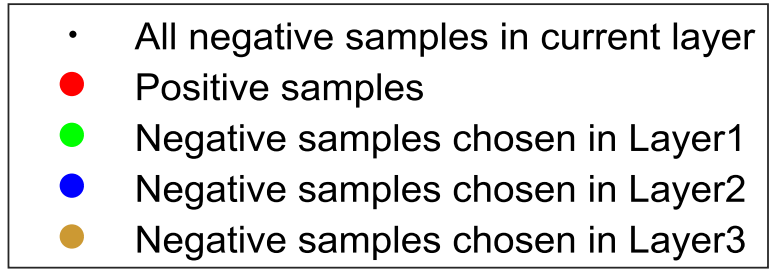
(a) Sample distribution in the 1st layer



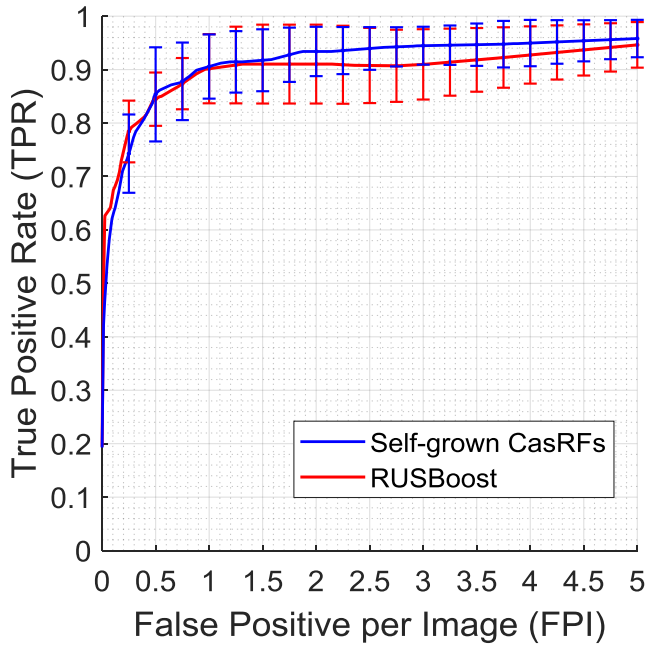
(b) Sample distribution in the 2st layer



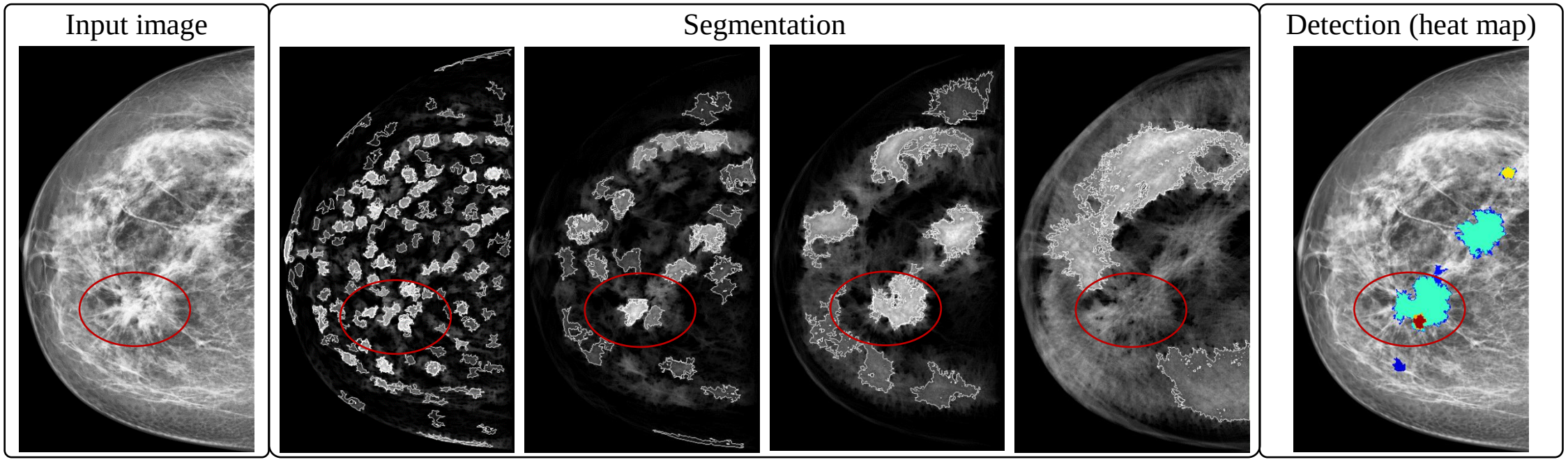
(c) Sample distribution in the 3st layer



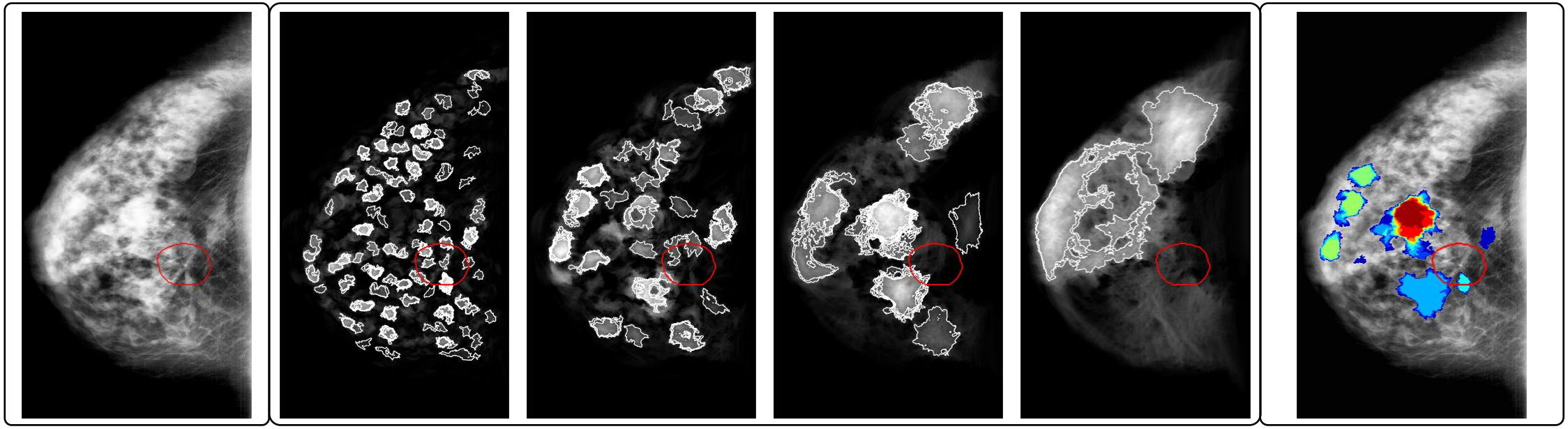
Figure


True Positive Rate (TPR)
False Positive per Image (FPI)
Self-grown CasRFs
RUSBoost
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Figure

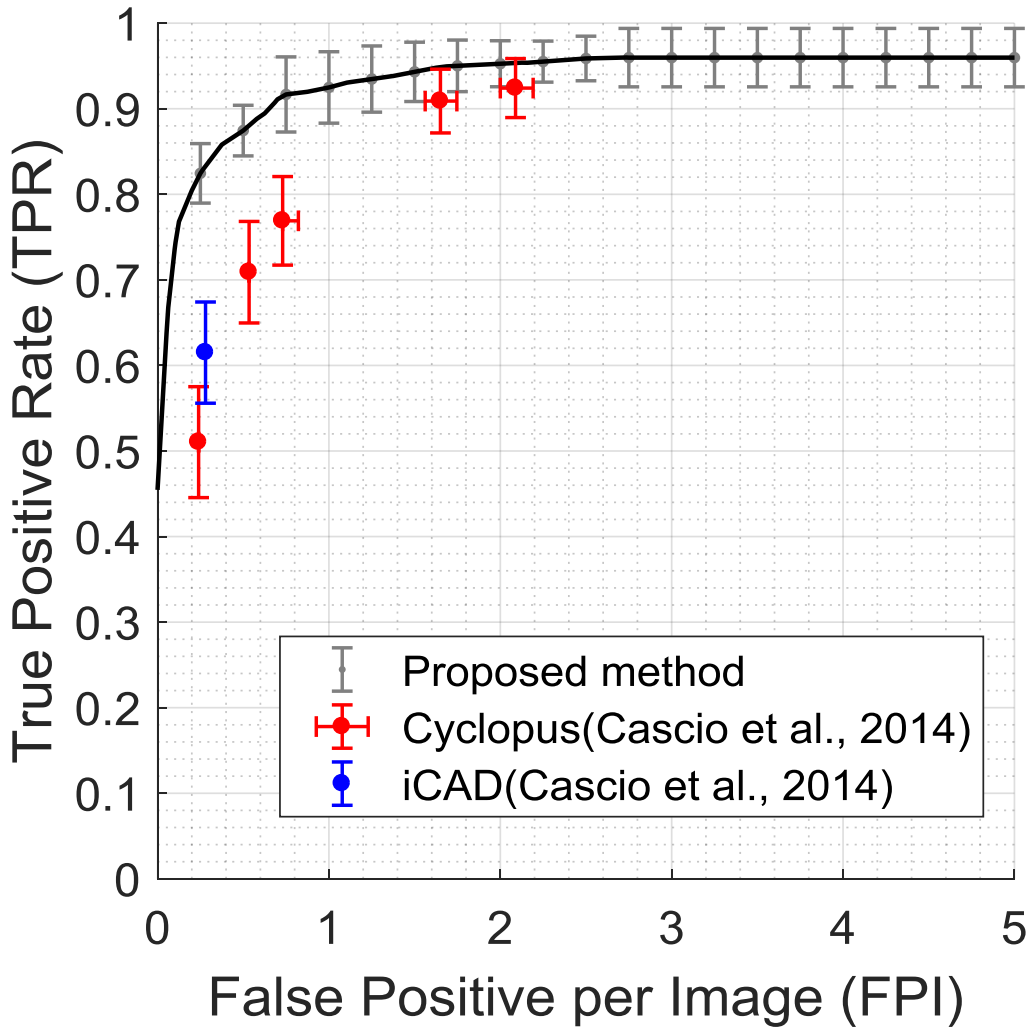


(a) A case from INbreast



(b) A case from DoD BCRP

Figure


True Positive Rate (TPR)
False Positive per Image (FPI)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
Proposed method
Cyclopus(Cascio et al., 2014)
iCAD(Cascio et al., 2014)