Multi-scale sifting for mammographic mass detection and segmentation

Article *in* Biomedical Physics & Engineering Express · January 2019

DOI: 10.1088/2057-1976/aafc07

CITATIONS
12

READS
355

4 authors, including:

Min Hang
Ingham Institute
7 PUBLICATIONS **229** CITATIONS
SEE PROFILE

Shekhar Suresh Chandra
The University of Queensland
86 PUBLICATIONS **842** CITATIONS
SEE PROFILE

Andrew P Bradley
Queensland University of Technology
223 PUBLICATIONS **11,832** CITATIONS
SEE PROFILE

Multi-scale sifting for mammographic mass detection and segmentation

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

Breast mass detection and segmentation are challenging tasks due to the fact that breast masses vary in size and appearance. In this work, we present a simultaneous detection and segmentation scheme for mammographic lesions that is constructed in a sifting architecture. It utilizes a novel region candidate selection approach and cascaded learning techniques to achieve state-of-the-art results while handling a high class imbalance. The region candidates are generated by a novel multi-scale morphological sifting (MMS) approach, where oriented linear structuring elements are used to sieve out the mass-like objects in mammograms including stellate patterns. This method can accurately segment masses of various shapes and sizes from the background tissue. To tackle the class imbalance problem, two different ensemble learning methods are utilized: a novel self-grown cascaded random forests (CasRFs) and the random under-sampling boost (RUSBoost). The CasRFs is designed to handle class imbalance adaptively using a probability-ranking based under-sampling approach, while RUSBoost uses a random under-sampling technique. This work is evaluated on two publicly available datasets: INbreast and DDSM BCRP. On INbreast, the proposed method achieves an average sensitivity of 0.90 with 0.9 false positives per image (FPI) using CasRFs and with 1.2 FPI using RUSBoost. On DDSM BCRP, the method yields a sensitivity of 0.81 with 3.1 FPI using CasRFs and with 2.9 FPI using RUSboost. The performance of the proposed method compares favorably to the state-of-the-art methods on both datasets, especially on highly spiculated lesions.

Keywords: Mammography, Breast mass detection and segmentation, Morphological sifting, Ensemble learning, Cascaded random forest

1. Introduction

Breast cancer is one of the leading causes of cancer-related 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 breast cancer (Schnabel et al., 2013). The detection of breast masses is generally regarded as more challenging compared with other breast abnormalities, not only due to the large variation in size and shape, 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 of the following stages. Firstly, region candidates are extracted from the mammograms, representing either abnormal or normal regions. To generate the region candidates, mammograms are partitioned into multiple regions/segments by analyzing the characteristics of pixels such as intensity, contrast and topographic representations (Schnabel et al., 2013). Region-driven approaches such as region-growing (Görgel et al., 2013), thresholding (Kozegar et al., 2013, Varela et al., 2007) and clustering (Martins et al., 2009), aim at segmenting mass-like structures from the background tissue as region candidates. However, these methods may not be sensitive to characteristic abnormal patterns, such as spiculations (Schnabel et al., 2013). Location-driven approaches such as region pooling (Ribli et al., 2017, Liu et al., 2015, Dhungel et al., 2015a) usually use bounding boxes distributed across the breast image to represent regions of interest (ROI). These methods aim at detection only and need further techniques for segmentation. Some systems with the detection and segmentation stage

integrated, however, still requires the users to reject false positive detections before the segmentation stage (Al-antari et al., 2018, Dhungel et al., 2017).

After region candidates are generated, machine learning algorithms are used for classifying the ROIs into lesion or normal tissue regions. Various machine learning algorithms, such as neural networks (Varela et al., 2007), support vector machines (Martins et al., 2009, Görgel et al., 2013), convolutional neural network (CNN) and random forest (Dhungel et al., 2015a, Kooi et al., 2017) have been applied to solve this problem. However, data imbalance is a common issue in this step, since the number of normal regions is significantly greater than the number of abnormal regions. To handle class imbalance, commonly used techniques include majority class under-sampling and minority class over-sampling (Dhungel et al., 2015a, Kozegar et al., 2013, Murthy et al., 2013, Jalalian et al., 2017). However, the sampled data may not be able to adequately represent the distribution of the original data (Kang and Cho, 2006, Bria et al., 2014).

In this work, we propose an end-to-end system that generates detection and segmentation results simultaneously, contrary to (Dhungel et al., 2017, Al-antari et al., 2018) that require users to reject false positive detections before the segmentation stage. The system consists of a novel multi-scale morphological sifting method which can generate accurate segmentations of the lesions, and an ensemble learning stage which can handle high class imbalance. Figure 1 shows the diagram of the proposed method. The multi-scale morphological sifting (MMS) utilizes morphological filters with oriented linear structuring elements to extract lesion-like patterns including linear spicules that are normally present in spiculated masses. The sifting process is applied on different scales so as to tackle the size variation of breast masses. To classify region candidates and handle class imbalance, two ensemble learning techniques, a novel self-grown cascaded random forest (CasRFs) and the random under-sampling boost (RUSboost) (Seiffert et al., 2010), are applied. Both methods adopt under-sampling techniques with ensemble structures, which can learn from all the given training data without artificially generating new data by oversampling or data augmentation. The CasRFs uses probability-ranked under-sampling within the cascade architecture, while the RUSboost is a hybrid of random under-sampling and boosting. The CasRFs presented is an modified version of the CasRFs in our preliminary work (Min et al., 2017). To the best of our knowledge, the RUSboost has not been applied in handling the class imbalance in mammographic mass detection. Here, we propose the usage of RUSboost as an alternative ensemble learning approach to the CasRFs that fits into the sifting framework and evaluate its performance on a highly skewed training dataset. Evaluated on two publicly available mammographic datasets, the proposed method achieves competitive performance compared to state-of-the-art methods that have significantly higher computational complexity (Dhungel et al., 2015a, Dhungel et al., 2017).

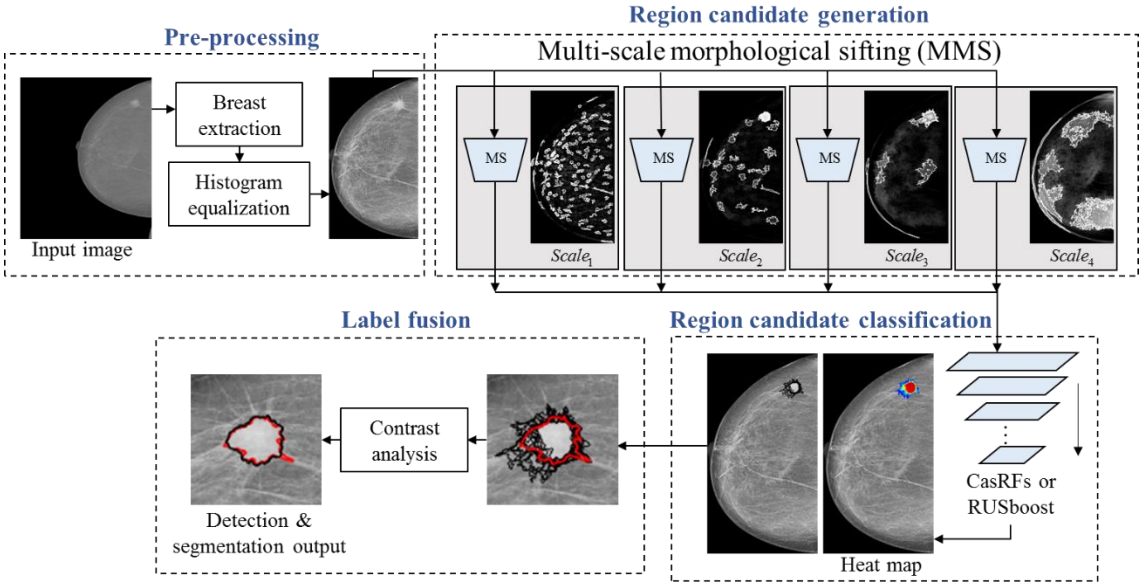


Figure 1. Diagram of the proposed mammographic mass sifting system. MS stands for morphological sifting. Red lines represent the annotation of the lesion and the black lines represent the detection outlines.

2. Methods

The sifting scheme utilizes a multi-scale morphological sifter for region candidate generation and an ensemble machine learning algorithm for region candidate classification. The MMS contains numerous linear morphological filters, an intensity and a size thresholding stage. The ensemble learning method (the CasRFs or

RUSboost) is designed to handle the class imbalance. The technical details of MMS and CasRFs are presented in this section.

2.1 Mammographic datasets

The system is evaluated on two public mammographic datasets, INbreast (Moreira et al., 2012) and DDSM BCRP (Bowyer et al., 1996). 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. DDSM BCRP is a screen-film mammographic dataset which contains many characteristic lesions such as spiculated, ill-defined masses and architectural distortions, which is suitable for testing the performance of the proposed method on highly irregular masses. INbreast is more clinical relevant compared to DDSM BCRP since most screening programs have adopted full-digital mammography. Therefore, INbreast is used as the primary evaluation dataset for our algorithm, and the performance on DDSM 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. The statistic details of these two datasets are presented in Table 1.

Table 1. Statistics of INbreast and DDSM BCRP.

Datasets	INbreast	DDSM BCRP	
		Training	Testing
Number of cases	115	39	40
Number of mammograms	410	156	160
Number of masses	116	80	81
Pixel size (μm)	70	43.5	
Contrast resolution (bit)	14	12	
Lesion size range (mm^2)	[15, 3689]	-	

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, Min et al., 2017, Danala et al., 2018). Here, we reduce the size of the mammograms by a factor of 4 using bi-cubic interpolation. For the INbreast dataset, the region where the pixels have a non-zero value is extracted as the breast region, since the background intensity of the INbreast mammograms is zero. For the DDSM BCRP dataset, a five-level multi-level Otsu thresholding (Otsu, 1979) is applied to the image. A binary image is generated by setting the pixels with an intensity between the lowest and the highest threshold to one and the rest to zero. Then the largest connected region in the binary image is extracted as the breast profile similar to (Dhungel, 2016, Dhungel et al., 2015a). The redundant background is then cropped away from the mammograms. The pixel values in the image are linearly rescaled to 16-bit and the contrast limited adaptive histogram equalization (CLAHE) is then applied (Pizer et al., 1987, Dhungel, 2016, Al-antari et al., 2018, Ball and Bruce, 2007, Neto et al., 2017). The number of rectangular contextual regions in CLAHE is set as 4×4 , and the contrast enhancement limit is set as 0.01 (the default). The parameters for CLAHE are determined experimentally on the training sets similar to (Ball and Bruce, 2007).

2.3 Region candidate generation using multi-scale morphological sifting

Morphological operations with structuring elements can be used to probe and analyze the images under the study for properties of interest (Gonzalez and Woods, 2008). By altering the size and shape of structuring elements, morphological filtering based approaches can extract mass-like patterns and suppress the background (Li et al., 2001a, Wang, 2006). In this work, we propose a new set of multi-scale grayscale morphological filters by using numerous pairs of oriented linear structuring elements (OLSE), as malignant mammographic densities are often surrounded by a radial pattern of linear spicules (Karssemeijer and te Brake, 1996). Since the morphological filters has the ability to sieve elements of interest from the background, we name the approach 'morphological sifting (MS)'. The algorithm is described as follows.

2.3.1 Single-scale morphological sifting

Firstly, two sets of OLSEs are defined as $\{L(ML_1, \theta(n)) \mid n = 0,1...N-1\}$ and $\{L(ML_2, \theta(n)) \mid n = 0,1...N-1\}$, where ML_1 , ML_2 stands for the magnitudes for the OLSEs, and $\theta(n)$ stands for the orientation of the OLSEs. $\theta(n)$

is equally spaced in $[0^\circ, 180^\circ]$. There are N elements in each set and the orientation difference between two adjacent lines is $\Delta\theta$ ($\Delta\theta = 180^\circ / N$). The number of OLSEs N affects the smoothness of the filtered image. Generally, as N increases, the smoother the processed image becomes at the cost of an increased computational complexity. $N = 18$ is a balanced value that generates relatively smooth processed image at a reasonable computational cost. One OLSE pair contains one OLSE from each set with the same orientation but different magnitudes as shown in Figure 2 (a).

The MS process on a single scale is described in the equation below.

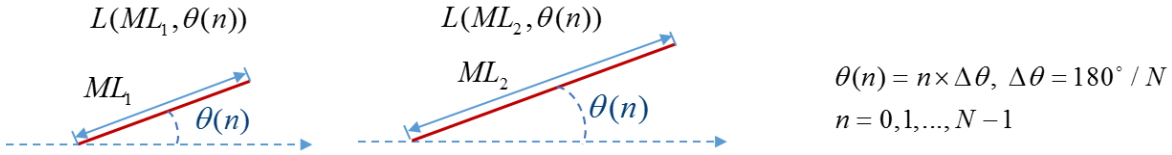
$$MS(n) = \{f - [f \circ L(ML_2, \theta(n))]\} \circ L(ML_1, \theta(n)) \quad (1)$$

$$f' = \sum_{n=0}^{N-1} MS(n) \quad (2)$$

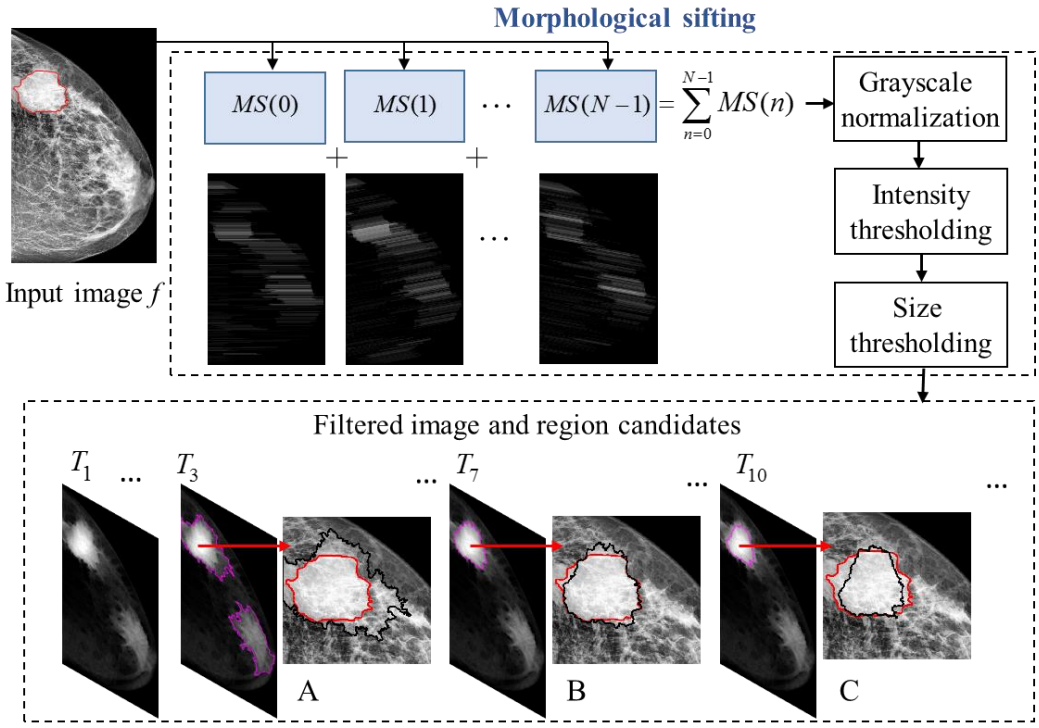
where f stands for the input image. '$\circ$' stands for morphological opening. f' is the output image after the MS. The output image f' is generated by applying a grayscale normalization (to 16-bit) on the summation of all the result images generated by the MSs over all the orientations as shown in Figure 2 (b).

The filtered image is then sieved through its intensity property by multi-level thresholding (MLT) (Otsu, 1979, Liao et al., 2001, Backes and Bruno, 2008). The MLT is utilized to generate a series of thresholds $[T_k \mid k = 1, ..., K]$. For each threshold T_k, if a pixel value at a pixel position (x, y) in the filtered image $f'(x, y) > T_k$, then the pixel value in the output binary image $B(x, y) = 1$, otherwise $B(x, y) = 0$. K thresholds should generate K binary images representing the segmentations at each level of threshold. The more levels of threshold (K) are used, the more likely for the algorithm to capture the accurate outlines of the lesions, but with a higher computational complexity. When $K = 16$, there are enough levels of threshold to segment the lesions in the training set accurately, and increase in K does not result in significant changes in segmentation.

At each level, the image is partitioned into numerous segments. These patches are then processed by a size thresholding. Only the patches within the area range of $[Area_{ML_1}, Area_{ML_2}]$ pass through as region candidates, where $Area_{ML_1} = \pi ML_1^2 / 4$ and $Area_{ML_2} = \pi ML_2^2 / 4$. The region candidate segmentation process on a single scale is shown in Figure 2 (b). There can be multiple region candidates generated at different levels of threshold overlapping with the annotation as shown in subfigures A, B and C in Figure 2 (b).



(a)



(b)

Figure 2. The process of morphological sifting on a single scale. (a) shows one oriented linear structuring element pair with the same orientation and different magnitudes. (b) shows the morphological sifting process. The red lines represent the annotation of the lesion. The black and magenta lines represent the generated region candidates. A, B and C are several examples of the region candidates generated at the lesion area. B is the best matching region candidate to the lesion on the current scale.

2.3.2 Multi-scale sifting

To enable the method to extract masses of different sizes, the MS is extended into a multi-scale process. Here, a logarithmic scale is used. 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 the OLSEs as,

$$[ML_{\min}, ML_{\max}] = \left[2(Area_{\min}/\pi)^{0.5}/PR,\ 2(Area_{\max}/\pi)^{0.5}/PR\right] \quad (3)$$

The logarithmic scale interval SI is defined as,

$$SI = \left(ML_{\max}/ML_{\min}\right)^{1/M} \quad (4)$$

Thus, on scale i, the magnitudes of the OLSE pair $ML_1(i)$ and $ML_2(i)$ are,

$$\{\mathrm{ML}_1(i) = ML_{\min} \times SI^{i-1}, \mathrm{ML}_2(i) = ML_{\min} \times SI^{i} \mid (i = 1, 2, ..., M)\} \quad (5)$$

And the filtered image on scale i is,

$$f'(i) = \sum_{n=0}^{N-1} \{f - [f \circ L(ML_2(i), \theta(n))]\} \circ L(ML_1(i), \theta(n)) \quad (6)$$

The lesion size range $[Area_{\min}, Area_{\max}]$ has been suggested in (Moreira et al., 2012). The number of scales M is set as 4, which can accommodate the size variance of the masses well with a reasonable computational complexity similar to (Dhungel et al., 2015a, Dhungel et al., 2017).

Note that there can be multiple region candidates generated at different levels of threshold from different scales overlapping with the annotation. Among all these region candidates, one of them will have the highest dice similarity index (DSI) (Dice, 1945) to the lesion, and this region candidate is regarded as the lesion's best matching region candidate (BMR). For instance, region candidate B in Figure 2 (b) is the BMR of the lesion. Here, the DSI between two regions ψ_1 and ψ_2 is defined as $DSI(\psi_1, \psi_2) = 2 \times (\psi_1 \cap \psi_2)/(|\psi_1| + |\psi_2|)$.

2.4 Region candidate classification using ensemble learning

After the region candidate segmentation, there are approximately 1.3 positive samples and 203 negative samples generated per mammogram used for training. The number of negative region candidates is approximately 160 times greater than the number of positive ones, which indicates a high class imbalance. To classify the region candidates and handle class imbalance, the self-grown CasRFs and RUSboost, are applied separately in this stage. The CasRFs can self-grow into a cascade of RFs where the class imbalance is distributed throughout the layers. It utilizes an under-sampling technique guided by the probability ranking of the majority samples to balance the training data for each base classifier. RUSboost is an ensemble learning technique that incorporates random under-sampling and boosting structure to handle class imbalance (Seiffert et al., 2010). It has not previously been adopted in mammographic mass detection. Here, we utilize RUSboost as an alternative learning method to CasRFs, and compare the performance of these two methods. The technical details about CasRFs are presented below.

2.4.1 Self-grown cascaded random forests

A step-by-step training process for the CasRFs is shown in Algorithm 1. 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). It has been successfully utilized in brain and abdominal abnormality detection (Mitra et al., 2014, Cherry et al., 2014).

During the training process on layer j, FP_j donates the false positive sample set and X is the sample set used to train the RF. X contains all the positive samples PS and a negative sample set X^n which is a subset of FP_j. X^n is selected based on the probability ranking of the negative samples, where the probability is the sample's probability of being positive. The probabilities are initialized by a default RF in Algorithm 1 step 3 and X^n consists of y negative samples with the lowest probabilities. A sensitivity threshold ST is given to the cascade to specify the true positive rate (TPR) on the training data. In step 5, a grid search is carried out to attain a RF model that meets the ST on the training set X. If $ntree$ stands for the number of trees and $mtry$ stands for the number of features tried at each split, the grid search ranges are $ntree \in [100,1000]$ and $mtry \in [0.5\sqrt{NF}, 2\sqrt{NF}]$, where NF stands for the number of features (Liaw and Wiener, 2002). If there is a RF that meets the ST as in step 6, the RF model is stored as the classifier RF_j on layer j. Otherwise, fewer negative samples are selected to build X as in step 9 and the training returns to step 4 to repeat the search. Throughout the layers, the cascade discards true negative samples and keeps the false positive samples for training. If the FP set stops to grow smaller, the cascade ceases to grow as in step 12.

As to the testing process, if a testing sample is classified as negative by a RF classifier RF_j at $layer_j$, then it is removed and identified as a normal region. If it is classified as positive, then it is passed onto RF_{j+1} at the next layer in the cascade, as shown in Figure 3. A testing sample will only be classified as positive (abnormal) if it is classified as positive by all the RFs in the cascade.

Algorithm 1 CasRFs training

PS, NS : sets of positive samples p and negative samples n

NP : number of positive samples

FP_j : the false positive (FP) sample set at layer j

ST : sensitivity threshold

X : a training sample set, X^p and X^n are the subsets of X only containing positive or negative samples

$rf(X)$: a RF trained on sample set X

$P(X)$: probabilities of sample set X generated by a RF

$V_y(P(X))$: subset of sample set X, containing y samples in X with the lowest probabilities

$Sen(rf(X))$: the sensitivity of a RF on sample set X

$AUC(rf)$: area under the receiver operating characteristic (ROC) curve of classifier rf

RF_j : the random forest (RF) classifier generated at layer j in the CasRFs

1 $j \leftarrow 1$, $FP_1 \leftarrow NS$, $varstop \leftarrow 0$

2 **while** $varstop = 0$

3 Initialize the probabilities for the FP samples: $P(FP_j) \leftarrow rf(PS \cup FP_j)$
Select negative samples and build a training set:
$X^p \leftarrow PS$, $y = NP$, $X^n \leftarrow V_y(P(FP_j))$, $X \leftarrow X^p \cup X^n$, $varstop1 \leftarrow 0$
4 **while** $varstop1 = 0$
5 Grid search to generate RFs, $G = \{rf_r(X) \mid r = 1,..., R\}$
6 **if** $\exists rf_r \in G,\ Sen(rf_r(X)) \geq ST$
7 $RF_j \leftarrow rf_r$, FP_{j+1} is attained by applying RF_j on FP_j and removing the true negatives from FP_j, $j = j+1$, $varstop1 \leftarrow 1$
8 **else**
9 Choose the rf with the highest AUC: $rf' = \arg\max_{rf_r \in G}(AUC(rf_r))$,
Update the probabilities, select fewer negative samples and rebuild the training set:
$P(X^n) \leftarrow rf'(X)$, $y = 0.9y$, $X^n \leftarrow V_y(P(X^n))$ $X \leftarrow X^p \cup X^n$
10 **end if**
11 **end while**
12 **if** $FP_{j-1} = FP_j = FP_{j+1}$
13 $varstop = 1$
14 **end if**
15 **end while**

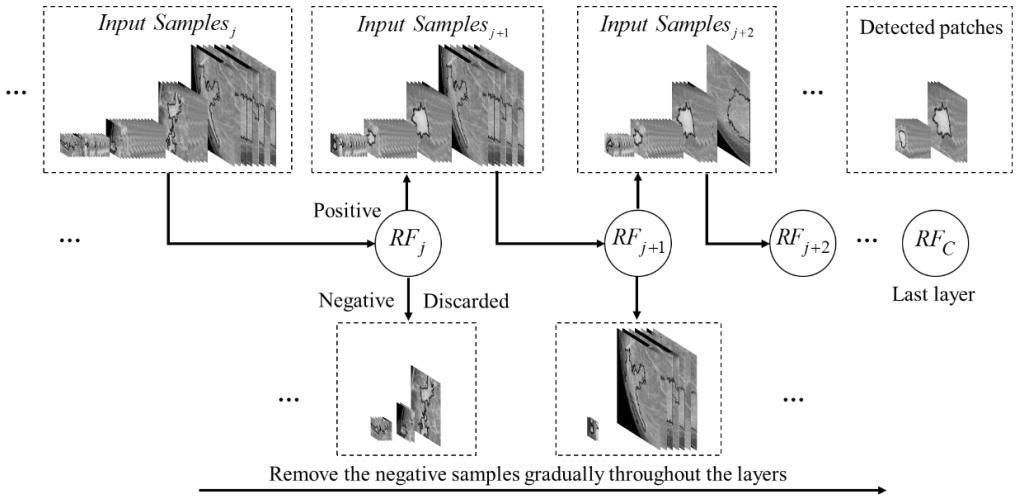


Figure 3. Sifting the testing samples through the CasRFs.

2.4.2 Feature extraction

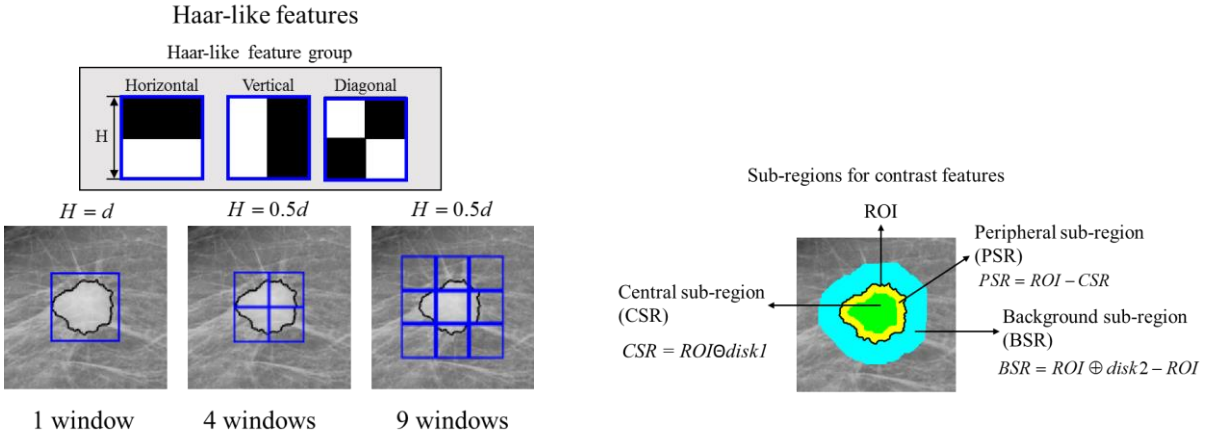
The features used in the ensemble learning are intensity, texture, contrast and shape features. A full list of all the features used can be found in Table 2. Intensity features 1~5 are calculated on both pre-processed images and the MMS filtered images within each region candidate. The gray level co-occurrence matrix (GLCM) features are calculated in four directions, horizontal, vertical, 45° and 135°. Contrast, correlation, energy and homogeneity are calculated on each direction. The design of Haar-like features is shown in Figure 4 (a). Horizontal, vertical and diagonal Haar-like features are calculated within a local window. H specifies the size of the window and d is the maximum radial distance between the center of the region and the pixels on the boundary. The Haar-like features are calculated within 14 windows in total. Contrast features 1~4 defined in (te Brake et al., 2000) are adopted in this work. To calculate these features, the ROI and its surrounding region are divided into the central sub-region, peripheral sub-region and background sub-region using morphological erosion and dilation as shown in Figure 4 (b). If the equivalent diameter of the ROI is D, the diameter of $disk1$ is $0.3D$ and the diameter of $disk2$ is $0.7D$. The contrast features (1~4) between the ROI and its background sub-region, central sub-region and peripheral sub-region are calculated on the pre-processed image. Only the contrast between the ROI and its background sub-region is calculated on the filtered image. A contrast difference (contrast feature 5) between the pre-processed image and the MMS filtered image is also calculated. If the difference of mean intensity between

the ROI and its background sub-region on the pre-processed image is C_1, the difference of mean intensity between the ROI and its background sub-region on the MMS filtered image is C_2, the mean intensity of ROI on the pre-processed image and on the filtered image are I_1 and I_2, the contrast difference is defined as $C_2 / I_2 - C_1 / I_1$.

Table 2. Hand-crafted features used in the proposed work.

	Intensity features	Shape features	Texture features	Contrast features
1	Maximum	Eccentricity	Entropy	Difference in mean intensity [b]
2	Minimum	Extent	GLCM related features	Normalized contrast measure [c]
3	Mean	Solidity	Haar-like features	Distance contrast measure [d]
4	Median	Circularity	Inertial momentum [a]	Matsutsita distance [e]
5	Standard deviation	Equivalent radius		Contrast difference [f]
6	Kurtosis			
7	Skewness			

[a] is defined in (Cascio et al., 2006).
[b, c, d, e] are contrast features defined in (te Brake et al., 2000).
[f] Contrast difference between the pre-processed image and the MMS filtered image.



(a) (b)

Figure 4. (a) Haar-like features. d is the maximum radial distance between the center of the ROI and the pixels on the boundary. (b) sub-regions for contrast features. The region surrounded by black line is the ROI. The green region is the central sub-region. The yellow region is the peripheral sub-region. And the cyan region is the background sub-region.

2.5 Post-processing using label fusion

Given a testing mammogram, all of the ROIs generated by MMS are passed through the ensemble classifier established during the training stage. There can be multiple detected patches overlapping with each other as shown in Figure 5. Heat maps are generated based on the probabilities of the detected regions. For a detected patch (DP), each layer of RF assigns a probability score ($prob$) to it, and the probability map for this patch is,

$$PM(x,y) = \begin{cases} \sum_{j=1}^{C} prob_j, (x,y) \in DP \\ 0, \ (x,y) \notin DP \end{cases} \quad (7)$$

where (x, y) represents a pixel position in the image and C is the number of RFs in the CasRFs. The final heat map is defined as the summation of the probability maps of all the detected patches $HM(x,y) = \sum_{b=1}^{B} PM_b(x,y)$, where B is the number of detected patches. The heat map is normalized to the range $[0,1]$. The heat maps can

visualize the probability of a detected region being a breast mass. Having a heat map allows the user to threshold the heat map to view the detections selectively, which can be helpful to the interpretation of the detections.

In order to generate a final segmentation for the lesion, the overlapping regions are fused by selecting the region with the highest contrast as the final segmentation. Here, the contrast is represented by the difference of mean intensity between the region and its background sub-region (te Brake et al., 2000). The background sub-region is the subtraction between the dilation of the ROI with a disk-shape structuring element and the ROI as shown in Figure 4 (b). The diameter of the disk is $0.2D$, where D is the equivalent diameter of the region.

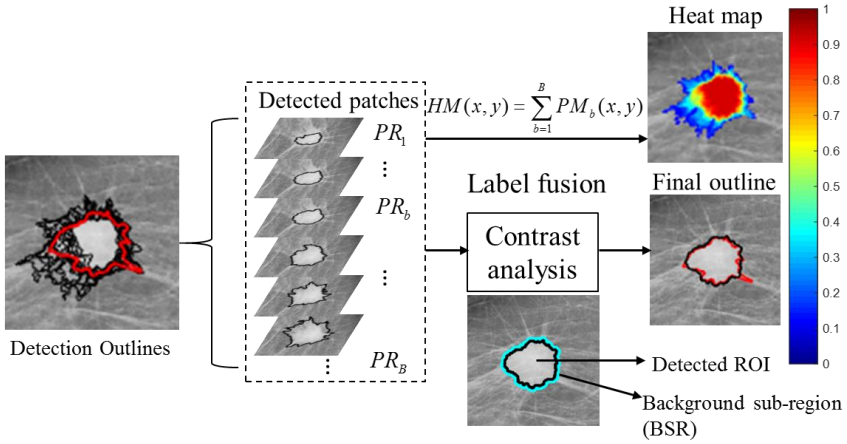


Figure 5. Heat map generation and label fusion. The black lines are the contours of the detected patches and the red lines are the contours of the mass annotation. The contrast is calculated between the detected ROI and its background sub-region (the cyan region).

2.6 Evaluation methods and experiment settings

Our system is evaluated on two public available datasets, INbreast and DDSM 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, Dhungel et al., 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 has inbuilt cross-validation. For the evaluation on DDSM BCRP, we simply use the pre-separated training and testing sets specified in this dataset. Note that the DDSM BCRP dataset is used as a secondary evaluation source and only the detection performance is evaluated on this dataset since it only provides approximate manual annotations of the lesions.

During the training stage, the training samples are labelled as positive or negative according to their DSI to the annotation. The BMRs of the lesions are labelled as positive and those candidates with a DSI lower than 0.1 are labelled as negative in the training sets. Since the DDSM BCRP does not have accurate contours of the lesions in the annotation, some of the training samples are manually labelled.

Table 3 shows a summary of primary parameter settings for the proposed method. The lesion size range $[Area_{\min}, Area_{\max}]$ suggested in (Moreira et al., 2012) is utilized for both datasets. For region candidate classification, a MATLAB RF package by Abhishek Jaiantilal (Jaiantilal, 2013) based on (Liaw and Wiener, 2002) is used to establish the CasRFs, while the Matlab ensemble learning toolbox is used to build the RUSboost. The CasRFs requires the input of a sensitivity threshold (ST) which specifies the desired sensitivity on the training set. ST can be decided by observing the trained model's performance on the training set, similar to (Bria et al., 2016). When ST is set as 0.99, there are approximately 1.1 false positives per image (FPI) on the training set on INbreast, which is close to the performance described in (Dhungel et al., 2017) that we intent to benchmark to. As to RUSboost, decision tree is used as the base learner. The maximum number of decision splits (MNS) is set as the number of training samples. The Number of learning cycles (NoC) is set as 1000, same as (García-Pedrero et al., 2017, Mounce et al., 2017). Increase in the number of cycles normally does not result in significantly different outcomes (Seiffert et al., 2010). The learning rate for shrinkage in boosting (LR) is set as 0.1 which is a popular choice and have been adopted in various publications (Hu et al., 2016, Paisitkriangkrai et al., 2014, Mounce et al., 2017). A Haar-like feature implementation by (Villamizar et al., 2006) is adopted with adaptations. All experiments are carried out on a Dell desktop with Intel Core i7-4790 CPU @ 3.60GHz, 16GB RAM.

Table 3. Parameter settings of the proposed method. N is the number of linear structuring elements. M is the number of scale used. K is the number of levels used in the multi-level thresholding. ST stands for the sensitivity threshold of the CasRFs. NoC stands for the number of learning cycles for RUSboost, and LR stands for the learning rate for shrinkage in boosting.

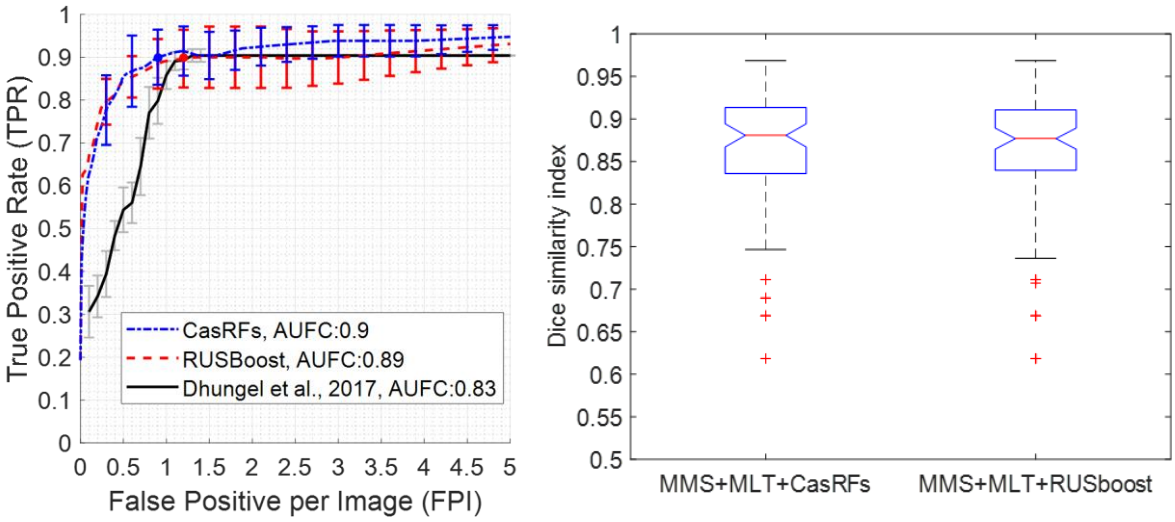
Stages	MMS				CasRFs	RUSboost	
Parameters	$[Area_{\min}, Area_{\max}]$ (mm^2)	N	M	K	ST	NoC	LR
INbreast & DDSM BCRP	15~3689	18	4	16	0.99	1000	0.1

3. Results

The MMS generates approximately 210 ± 78 region candidates per mammogram on INbreast. In this work, if a detected region has a $DSI \geq 0.6$ against the lesion annotation, it is regarded as a true positive detection, otherwise, as a false positive detection. Under this evaluation criterion, the minimum intersection over union (IoU) index between the bounding boxes of the detected region and the annotation is 0.52, which satisfies the $IoU \geq 0.5$ criterion used in (Dhungel et al., 2017). The free response operating characteristic (FROC) curves of the proposed work and (Dhungel et al., 2017) are shown in Figure 6 (a). The partial area under the FROC curve (AUFC) (Chakraborty, 2008) in the FPI range of $[0,5]$ is calculated for the proposed work and (Dhungel et al., 2017). The proposed work yields a AUFC value of 0.90 and 0.89 respectively using CasRFs and RUSboost, which is higher than the AUFC value of 0.83 achieved by (Dhungel et al., 2017). When the average TPR is 0.90, the proposed method generates 0.9 FPI using CasRFs (mark with the blue dot on Figure 6 (a)), and 1.2 FPI using RUSboost (marked with the red dot on Figure 6 (a)), which is lower than the 1.3 FPI in (Dhungel et al., 2017). The box-plots of the segmentation performance with CasRFs and RUSboost are shown in Figure 6 (b). Both of them have a median of 0.88. The average DSI for both ensemble classifiers is 0.86 ± 0.08. Since in clinical screening, the majority of the mammographic cases may not contain breast cancer (Kolb et al., 2002), it is also important to evaluate the detection specificity on the normal mammograms and breasts. Table 4 below shows the percentage of normal mammograms and breasts without detections in each testing set using CasRFs. Each breast contains two mammograms. It shows that the proposed system does not generate detections in some of the normal mammograms or breasts.

Additional evaluation of the detection performance is conducted on DDSM BCRP. The MMS generates approximately 210 ± 70 region candidates per mammogram in DDSM BCRP. Here, we decide if the overlap coefficient (OC) (Reichel and Cole, 2016, Dhungel et al., 2015a, Ben-Ari et al., 2017) between a detected region and the annotation is higher than 0.7, the lesion is regarded as detected. Here, the overlap coefficient between two regions ψ_1 and ψ_2 is defined as $OC(\psi_1, \psi_2) = \psi_1 \cap \psi_2 / \min(\psi_1, \psi_2)$ (Reichel and Cole, 2016, Szymkiewicz, 1934). The system achieves a TPR of 0.81 at 3.1 FPI using CasRFs, and a TPR of 0.81 at 2.9 FPI using RUSboost. To further evaluate the system's performance, we also calculated the detection sensitivity on masses with different levels of subtlety as shown in Table 5, since the DDSM BCRP dataset provides the subtlety rating for lesions, indicating the difficulty level of detecting the mass. The subtlety rate is in the range of 1 to 5. A lower subtlety rate suggests a less obvious mass and a higher difficulty level for detection (Bowyer et al., 1996). The system can achieve a sensitivity higher than 0.8 for masses with a subtlety level ≥ 2.

Figure 7 shows several detection and segmentation examples for lesions in various shapes and sizes from INbreast and DDSM BCRP using CasRFs. Table 6 shows the performance comparison between the proposed work and several state-of-the-art methods evaluated on the same datasets.



(a) (b)

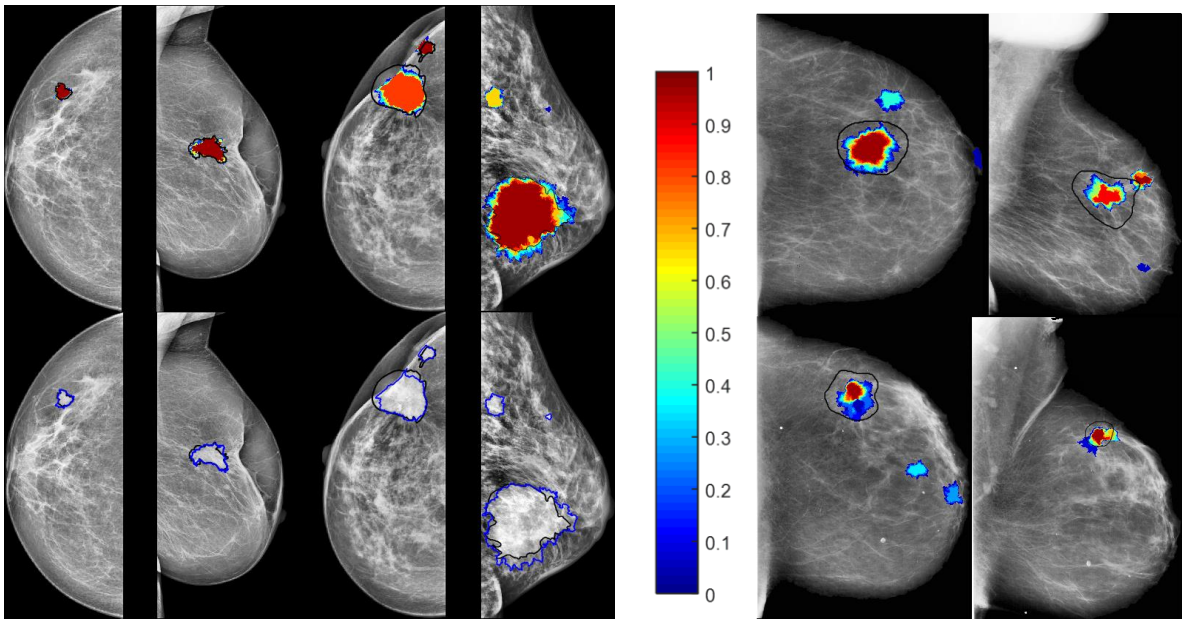
Figure 6. (a) FROC curves of the proposed work using MMS&CasRFs, MMS&RUSboost, and (Dhungel et al., 2017) on INbreast. The blue dot marks where the system using CasRFs achieves an average sensitivity of 0.90 ± 0.06 with 0.9 FPI, and the red dot marks where the system using RUSboost achieves an average sensitivity of 0.90 ± 0.08 with 1.2 FPI. AUFC stands for partial area under the FROC curve. (b) The box plots of the DSI of the system using CasRFs and RUSboost on INbreast.

Table 4. Percentage of normal mammograms and breasts with no detections in the INbreast dataset.

Testing sets	1	2	3	4	5
Pct. of normal mammograms without detections	0.75	0.58	0.88	0.63	0.30
Pct. of normal breasts without detections	0.50	0.44	0.75	0.25	0.20

Table 5. Detection sensitivity on masses in DDSM BCRP with different subtlety levels. TPR stands for true positive rate.

	Subtlety				
	1	2	3	4	5
TPR	0.30	0.86	0.81	0.89	0.92



(a) Detection heat maps and segmentations from INbreast (b) Detection heat maps from DDSM BCRP

Figure 7. Examples of detections and segmentations from INbreast and DDSM BCRP datasets using CasRFs. (a) Heat maps and final segmentations of several masses in various sizes and shapes from INbreast. (b) Heat maps of

the detected lesions in DDSM BCRP. The black lines represent the annotations. The blue lines represent the final segmentation of the lesions.

Table 6. Performance comparison between the proposed methods and previous publications. TPR stands for true positive rate, FPI stands for false positive per image and DSI stands for dice similarity index.

Datasets	INbreast			DDSM BCRP	
Methods	TPR	FPI	DSI	TPR	FPI
MMS+MLT+CasRFs	0.90 ± 0.06	0.9	0.86 ± 0.08	0.81	3.1
MMS+MLT+RUSboost	0.90 ± 0.08	1.2	0.86 ± 0.08	0.81	2.9
(Dhungel et al., 2017)	0.90 ± 0.02	1.3	0.85 ± 0.02	-	-
(Kozegar et al., 2013)	0.87	3.67	-	-	-
(Beller et al., 2005)	-	-	-	0.70	8
(Dhungel et al., 2015a)	0.96, 0.87	1.2, 0.8	-	0.75, 0.7	4.8, 4

4. Discussion

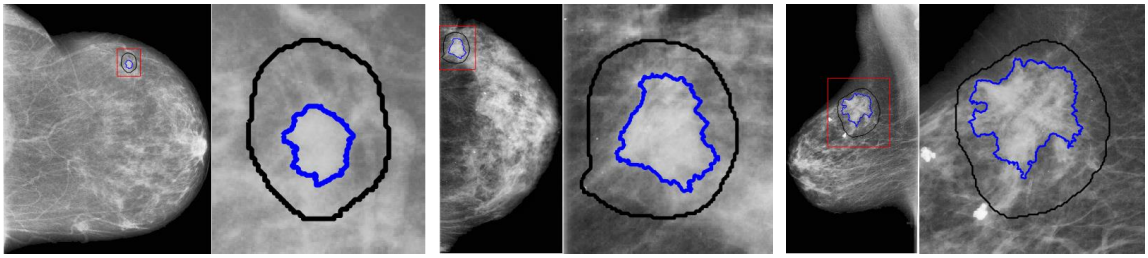
As shown in Figure 6 and Table 6, both MMS with CasRFs and MMS with RUSboost show competitive performance in lesion detection and segmentation on INbreast. The proposed method using CasRFs achieves the same average TPR at a lower FPI when compared to other methods evaluated on INbreast. The CasRFs slightly outperforms the RUSboost in detection on INbreast, while both classification methods has similar segmentation performance as shown in Figure 6 (b) since they share the same region candidate segmentation method. Although (Dhungel et al., 2017) appears to have a smaller variance in TPR at FPI = 1.3 as shown in Table 6, the variance becomes higher when $FPI \in [0,1]$ as shown in Figure 6 (a). The proposed work can achieve a much higher TPR at a similar variance to (Dhungel et al., 2017) when FPI is lower than 1. For instance, the proposed method achieves an average TPR of 0.9 ± 0.06 and 0.88 ± 0.06 using CasRFs and RUSboost respectively at 0.9 FPI, while (Dhungel et al., 2017) yields an average TPR of 0.80 ± 0.05 at the same level of FPI as shown in Figure 6 (a). It is important for a detection system to maintain a satisfactory TPR at a relatively low FPI, which could limit unnecessary call backs (Li et al., 2001b). The proposed method achieves a slightly higher average DSI compared with (Dhungel et al., 2017). (Dhungel et al., 2017) appears to have a lower variance in DSI, however, its segmentation stage relies on the detection output of (Dhungel et al., 2015a) with user intervention, requiring a manual rejection of false positive detections. This could lead to a lower variance and high accuracy because the segmentation is effectively based of a semi-automated detection system. As to the evaluation on DDSM BCRP, both the CasRFs based and RUSboost based methods outperformed previous publications. Note that the method presented in (Dhungel et al., 2015a) focuses on mass detection. It uses bounding boxes to represent detections and a lesion is regarded as detected if the detected box has an overlap ratio with the bounding box of the annotation higher than 0.2 (Dhungel et al., 2015a). In the more recent publication (Dhungel et al., 2017), the authors have extended the detection technique into a detection & segmentation system on INbreast by combining the detection system in (Dhungel et al., 2015a) and the segmentation algorithm in (Dhungel et al., 2015b). Therefore, we chose to compare with the authors' latest results on INbreast in (Dhungel et al., 2017). The detection performance in (Dhungel et al., 2015a) on DDSM BCRP is still used for comparison. It is worth noting that around 43% of the cases in INbreast contain masses and all of the cases in DDSM BCRP contain lesions, while in a population-based breast screening scenario, the percentage of cases with breast cancer can be only around 2% (Kolb et al., 2002). The fact that the public datasets may not reflect the real breast cancer population can pose a limitation to the evaluation of the proposed work.

The proposed mass sifting method can sieve the mass-like elements out of the mammograms through their morphological properties. The region candidate generation method analyses the spatial and intensity information

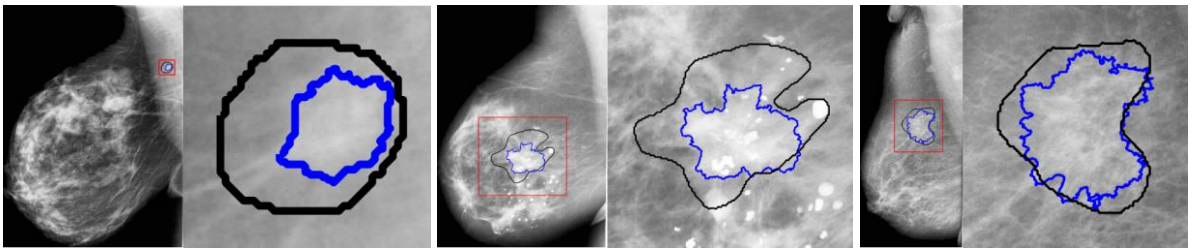
of the mammographic patterns by utilizing morphological filtering and thresholding. The MMS utilizes oriented linear structuring elements to imitate stellates in lesions and can generate explicit segmentations of the lesions as region candidates, which is a good foundation for the later classification. The average DSI between the annotations and their BMRs is 0.89 ± 0.07 on INbreast. Table 7 shows the segmentation performance comparison between the proposed region candidate generation method, single-scale morphological sifting, multi-level thresholding and the method in (Min et al., 2017). It can be observed that the MMS has the highest average DSI to the lesions at the lowest variance. The MMS significantly outperformed the other methods in region candidate segmentation. As to the DDSM BCRP, we could not evaluate the candidate segmentation quantitatively, since there is no accurate contour of the lesion in the annotation. Instead, we present some examples of the region candidates segmented at the abnormal area in Figure 8. These examples include characteristic lesions such as spiculated masses and architectural distortions. Among the region candidates shown in Figure 8, the MMS is generally effective in segmenting lesions with a visible focal area. However, MMS still has limitations in segmenting subtle architectural distortions without a definable central mass. In Table 5, eighty percent of the masses with a subtlety level of 1 are architectural distortions, which leads to a lower detection sensitivity on lesions at this subtlety level. Breast architectural distortions are very characteristic and can be overlooked by CAD systems designed to detect masses. There have been several methods specially designed to identify breast architectural distortions (Rangayyan et al., 2010, Rangayyan and Ayres, 2006, Ben-Ari et al., 2017, Liu et al., 2016).

Table 7. Region candidate segmentation comparison between the proposed multi-scale morphological sifting and several other methods. The average best dice similarity index (DSI) is the average DSI between the lesions and the lesions' best matching region candidates.

Region candidate segmentation methods	Average best DSI of the lesions
Multi-scale morphological sifting	0.89 ± 0.07
Single-scale morphological sifting	0.86 ± 0.12
Multi-level thresholding	0.79 ± 0.15
(Min et al., 2017)	0.78 ± 0.11



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



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

Figure 8. Examples of the region candidate generation on DDSM BCRP. The black lines represent the annotations and the blue lines represent the region candidates segmented at the lesion area. AD stands for architectural distortion.

The proposed CasRFs is an adaptively growing ensemble learning machine with minimal need of tuning. With the default setting, the only parameter that needs to be specified for training is the sensitivity threshold ST . Since random forests have inbuilt random cross validation, instead of splitting the training data manually into training and validation sets as in (Dhungel et al., 2017), the training can be ran on all of the training data to gain richer training information. Compared with the CasRFs described in our previous work (Min et al., 2017), the modified version in this work uses a different initialization approach. Here, all the negative samples are used to train the default random forest to generate the probability ranking at the first layer instead of a random selected subset of

negative samples as in (Min et al., 2017). Moreover, the CasRFs are trained on the training samples from all scales instead of being trained on each individual scale as in (Min et al., 2017).

Both CasRFs and RUSboost are designed to use an under-sampling technique to balance the training data for each base classifier. However, unlike RUSboost which under-samples the majority class randomly, CasRFs uses a probability ranking guided under-sampling approach. When the training set is highly skewed, it could be unsafe to oversample the minority class to balance the training data (Dhungel et al., 2017), since the artificially generated samples may not correspond to real masses (Bria et al., 2014, Kang and Cho, 2006, Bria et al., 2016). The ensemble learning methods can handle high class imbalance while using only samples that correspond to real breast patterns. Figure 9 shows the comparison between the CasRFs, RUSboost, and some other region candidate classification methods explored in previous publications, the support vector machine (SVM) (Martins et al., 2009, Görgel et al., 2013), random under-sampling SVM (Klement et al., 2014) and SMOTE (synthetic minority over-sampling technique) SVM (Liu and Zeng, 2015, Klement et al., 2014), in terms of FROC curves. Note that both the under-sampling SVM and SMOTE SVM are designed to handle class imbalance. It can be observed that the CasRFs and RUSboost significantly outperform these methods on the imbalanced dataset.

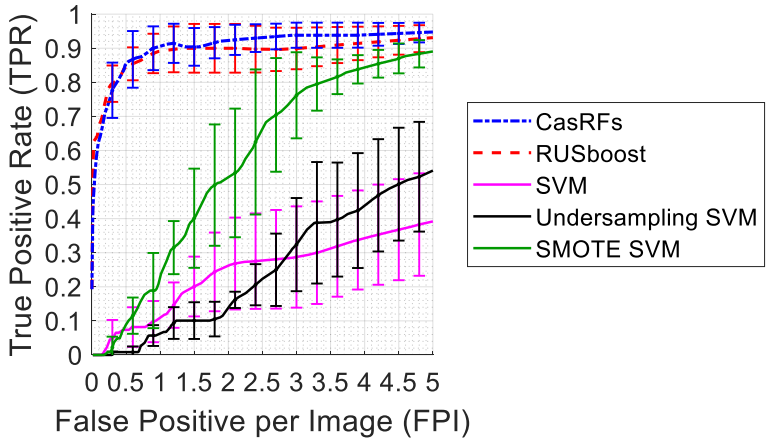
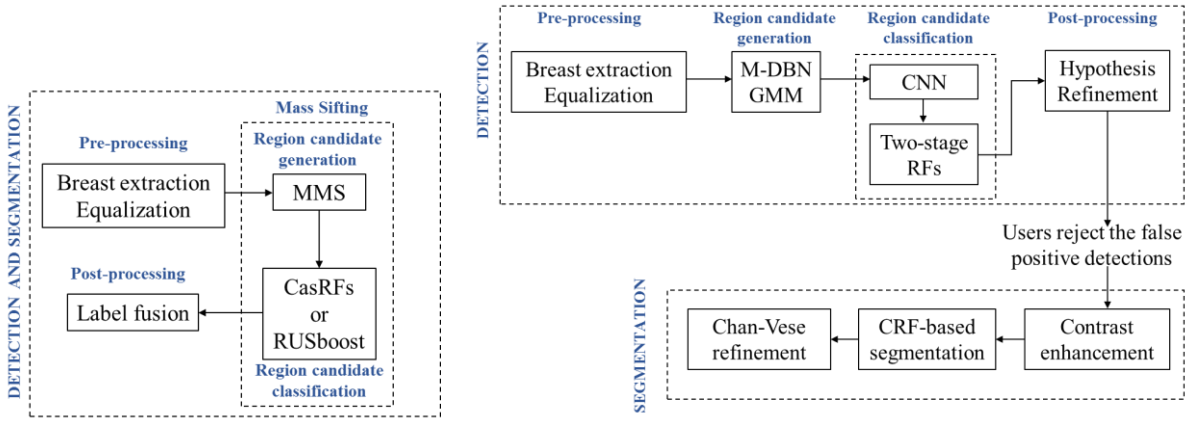


Figure 9. FROC curves of CasRFs, RUSboost, support vector machine (SVM), random under-sampling SVM and SMOTE SVM. SMOTE stands for synthetic minority over-sampling technique.

A cross-dataset evaluation is also carried out. The CasRFs model with the best performance among the five-fold validations generated on INbreast is utilized to detect masses in DDSM BCRP. It yields a 0.78 TPR at 4.2 FPs per image. The CasRFs model generated on the DDSM BCRP is also used for detection on mammograms from INbreast, and it yields a 0.82 TPR at 1.7 FPs per image. The cross-dataset testing performance is not as good as the testing performance within the same dataset. It is expected since these two datasets are different in the mode of image acquisition, resolution, and lesion types. In INbreast, architectural distortions are not considered as breast masses, while in DDSM BCRP, the lesions consist of both architectural distortions and spiculated masses.

Contrary to (Dhungel et al., 2017), the proposed scheme is an end-to-end system for both segmentation and detection with no user intervention that also generates a probability map. This allows for a simpler deployable, fully automated CAD system that supports easier interpretation of segmentation results. The proposed work also has a lower complexity as shown in Figure 10. It has fewer steps and consistently uses the same type of base classifier in an ensemble structure (RF for CasRFs and decision tree for RUSboost). Figure 10 (b) shows that (Dhungel et al., 2017) is a combination of a detection system (Dhungel et al., 2015a) and a segmentation system (Dhungel et al., 2015b). It adopted three different learning algorithms (CNN, RF and conditional random field) and refinement stages, which requires more stages of training and optimization (including millions of weights) than the proposed work. Moreover, the users need to reject the FP detections before the segmentation process in (Dhungel et al., 2017), while the proposed system can generate detection and segmentation simultaneously. The MMS method plays a significant role in achieving competitive detection and segmentation performance, since it generates region candidates that represent the true lesions explicitly. A potential limitation of the proposed work is that some of the parameters are determined empirically. In the future work, we would like to explore more adaptive approaches to determine the parameters, and the combination between the MMS and CNN to eliminate the need for hand-crafted features.



(a) (b)

Figure 10. Structure comparison between the proposed work and the state-of-the-art method (Dhungel et al., 2017). (a) The workflow of the proposed mass sifting method. MMS stands for multi-scale morphological sifting. (b) The workflow of the breast mass detection and segmentation method in (Dhungel et al., 2017). M-DBN stands for multi-scale deep belief nets, GMM stands for Gaussian mixture model, CNN stands for convolutional neural network, RF stands for random forest, and CRF stands for conditional random field.

5. Conclusion

In this work, we have presented a mammographic mass sifting scheme that detects and segments breast masses simultaneously. Evaluated on two public available datasets, the proposed method performs favorably to the state-of-the-art methods. This work introduces a novel mammographic mass sifting structure, where the mammographic patterns are sieved through a multi-scale morphological sifter and layers of base classifiers in the ensemble learning algorithms. The multi-scale morphological sifter is able to segment lesions from the background accurately. Two ensemble learning techniques, the self-grown CasRFs and RUSboost, are adopted to handle class imbalance. The CasRFs can adapt to severely skewed training data with minimal tuning. Both ensemble learning techniques are capable of attaining satisfactory classification performance while facing high class imbalance. In general, the proposed system achieves promising results in both detection and segmentation on digital and screen-film mammograms.

Acknowledgements

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.

Reference

DoD BCRP Spiculated Mass Detection Evaluation Data. University of South Florida.

Al-antari, M. A., Al-masni, M. A., Choi, M.-T., Han, S.-M. & Kim, T.-S. 2018. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *International Journal of Medical Informatics*, 117**,** 44-54.

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

Ball, J. E. & Bruce, L. M. Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation. Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, 2007. IEEE, 4973-4978.

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.

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

Bowyer, K., Kopans, D., Kegelmeyer, W., Moore, R., Sallam, M., Chang, K. & Woods, K. The digital database for screening mammography. Third international workshop on digital mammography, 1996. 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. *Medical Image Analysis,* 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., 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 Transactions on Nuclear Science,* 53**,** 2827-2833.

Chakraborty, D. P. 2008. Validation and Statistical Power Comparison of Methods for Analyzing Free-response Observer Performance Studies. *Academic Radiology,* 15**,** 1554-1566.

Cherry, K. M., Wang, S., Turkbey, E. B. & Summers, R. M. Abdominal lymphadenopathy detection using random forest. Medical Imaging 2014: Computer-Aided Diagnosis, 2014. International Society for Optics and Photonics, 90351G.

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. *Medical Physics,* 42**,** 3859-3869.

Danala, G., Aghaei, F., Heidari, M., Wu, T., Patel, B. & Zheng, B. Computer-aided classification of breast masses using contrast-enhanced digital mammograms. Medical Imaging 2018: Computer-Aided Diagnosis, 2018. International Society for Optics and Photonics, 105752K.

Dhungel, N. 2016. *Automated detection, segmentation and classification of masses from mammograms using deep learning.*

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

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

Dhungel, N., Carneiro, G. & Bradley, A. P. The automated learning of deep features for breast mass classification from mammograms. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016. Springer, 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. *Medical Image Analysis,* 37**,** 114-128.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology,* 26**,** 297-302.

Dubitzky, W., Granzow, M. & Berrar, D. P. 2007. *Fundamentals of data mining in genomics and proteomics,* Springer Science & Business Media.

Ganesan, K., Acharya, U. R., Chua, C. K., Min, L. C., Abraham, K. T. & Ng, K.-H. 2013. Computer-aided breast cancer detection using mammograms: a review. *IEEE Reviews in Biomedical Engineering,* 6**,** 77-98.

García-Pedrero, A., Gonzalo-Martín, C. & Lillo-Saavedra, M. 2017. A machine learning approach for agricultural parcel delineation through agglomerative segmentation. *International Journal of Remote Sensing,* 38**,** 1809-1819.

Gonzalez, R. C. & Woods, R. E. 2008. *Digital image processing / Rafael C. Gonzalez, Richard E. Woods,* Harlow, Harlow : Pearson/Prentice Hall.

Görgel, P., Sertbas, A. & Ucan, O. N. 2013. Mammographical mass detection and classification using Local Seed Region Growing-Spherical Wavelet Transform (LSRG-SWT) hybrid scheme. *Computers in Biology and Medicine,* 43**,** 765-774.

Gromet, M. 2008. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. *American Journal of Roentgenology,* 190**,** 854-859.

Hu, Q., Paisitkriangkrai, S., Shen, C., van den Hengel, A. & Porikli, F. 2016. Fast detection of multiple objects in traffic scenes with a common detection framework. *IEEE Transactions on Intelligent Transportation Systems,* 17**,** 1002-1014.

Jaiantilal, A. 2013. *Randomforest-matlab* [Online]. Available: https://github.com/jrderuiter/randomforest-matlab [Accessed].

Jalalian, A., Mashohor, S., Mahmud, R., Karasfi, B., Saripan, M. I. B. & Ramli, A. R. B. 2017. Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection. *EXCLI journal,* 16**,** 113.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E. & Forman, D. 2011. Global cancer statistics. *CA: a cancer journal for clinicians,* 61**,** 69-90.

Kang, P. & Cho, S. EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems. International Conference on Neural Information Processing, 2006. Springer, 837-846.

Karssemeijer, N. & te Brake, G. M. 1996. Detection of stellate distortions in mammograms. *IEEE Transactions on Medical Imaging,* 15**,** 611-619.

Klement, R. J., Allgäuer, M., Appold, S., Dieckmann, K., Ernst, I., Ganswindt, U., Holy, R., Nestle, U., Nevinny-Stickel, M., Semrau, S., Sterzing, F., Wittig, A., Andratschke, N. & Guckenberger, M. 2014. Support Vector Machine-Based Prediction of Local Tumor Control After Stereotactic Body Radiation Therapy for Early-Stage Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology*Biology*Physics,* 88**,** 732-738.

Kolb, T. M., Lichy, J. & Newhouse, J. H. 2002. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology,* 225**,** 165-175.

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A. & Karssemeijer, N. 2017. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis,* 35**,** 303-312.

Kozegar, E., Soryani, M., Minaei, B. & Domingues, I. 2013. Assessment of a novel mass detection algorithm in mammograms. *Journal of cancer research and therapeutics,* 9**,** 592.

Li, H., Wang, Y., Liu, K. R., Lo, S.-C. & Freedman, M. T. 2001a. Computerized radiographic mass detection. I. Lesion site selection by morphological enhancement and contextual segmentation. *IEEE Transactions on Medical Imaging,* 20**,** 289-301.

Li, L., Zheng, Y., Zhang, L. & Clark, R. A. 2001b. False-positive reduction in CAD mass detection using a competitive classification strategy. *Medical physics,* 28**,** 250-258.

Liao, P.-S., Chen, T.-S. & Chung, P.-C. 2001. A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng.,* 17**,** 713-727.

Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. *R news,* 2**,** 18-22.

Liu, L., Li, J. & Wang, Y. Breast mass detection with kernelized supervised hashing. 2015 8th International Conference on Biomedical Engineering and Informatics (BMEI), 2015. IEEE, 79-84.

Liu, X. & Zeng, Z. 2015. A new automatic mass detection method for breast cancer with false positive reduction. *Neurocomputing,* 152**,** 388-402.

Liu, X., Zhai, L. & Zhu, T. Recognition of architectural distortion in mammographic images with transfer learning. Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on, 2016. IEEE, 494-498.

Martins, L. d. O., Cardoso de Paiva, A., Corrêa Silva, A., Braz Junior, G. & Gattass, M. 2009. Detection of Masses in Digital Mammograms using K-Means and Support Vector Machine. *ELCVIA. Electronic letters on computer vision and image analysis,* 8**,** 39-50.

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

Mitra, J., Bourgeat, P., Fripp, J., Ghose, S., Rose, S., Salvado, O., Connelly, A., Campbell, B., Palmer, S., Sharma, G., Christensen, S. & Carey, L. 2014. Lesion segmentation from multimodal MRI using random forest following ischemic stroke. *NeuroImage,* 98**,** 324-335.

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

Mounce, S., Ellis, K., Edwards, J., Speight, V., Jakomis, N. & Boxall, J. 2017. Ensemble decision tree models using rusboost for estimating risk of iron failure in drinking water distribution systems. *Water Resources Management,* 31**,** 1575-1589.

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.

Neto, O. P. S., Silva, A. C., Paiva, A. C. & Gattass, M. 2017. Automatic mass detection in mammography images using particle swarm optimization and functional diversity indexes. *Multimedia Tools and Applications,* 76**,** 19263-19289.

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

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

Paisitkriangkrai, S., Shen, C. & van den Hengel, A. Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features. 2014 Cham. Springer International Publishing, 546-561.

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. & Ayres, F. J. 2006. Gabor filters and phase portraits for the detection of architectural distortion in mammograms. *Medical and biological engineering and computing,* 44**,** 883-894.

Rangayyan, R. M., Banik, S. & Desautels, J. L. 2010. Computer-aided detection of architectural distortion in prior mammograms of interval cancer. *Journal of Digital 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. *Annual Review of Biomedical Engineering,* 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.

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

te Brake, G. M., Karssemeijer, N. & Hendriks, J. H. 2000. An automatic method to discriminate malignant masses from normal tissue in digital mammograms1. *Physics in Medicine and Biology,* 45**,** 2843.

Varela, C., Tahoces, P. G., Méndez, A. J., Souto, M. & Vidal, J. J. 2007. Computerized detection of breast masses in digitized mammograms. *Computers in Biology and Medicine,* 37**,** 214-226.

Villamizar, M., Sanfeliu, A. & Andrade-Cetto, J. Computation of rotation local invariant features using the integral image for real time object detection. Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, 2006. IEEE, 81-85.

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