Journal of Cancer Research and Therapeutics

JCrT

Impact Factor® for 2012
0.761

PubMed
Indexed

Wolters Kluwer
Health | Medknow

Assessment of a novel mass detection algorithm in mammograms

ABSTRACT

Context: Mammography is the most effective procedure for an early detection of the breast abnormalities. Masses are a type of abnormality, which are very difficult to be visually detected on mammograms.

Aims: In this paper an efficient method for detection of masses in mammograms is implemented.

Settings and Design: The proposed mass detector consists of two major steps. In the first step, several suspicious regions are extracted from the mammograms using an adaptive thresholding technique. In the second step, false positives originating by the previous stage are reduced by a machine learning approach.

Materials and Methods: All modules of the mass detector were assessed on mini-MIAS database. In addition, the algorithm was tested on INBreast database for more validation.

Results: According to FROC analysis, our mass detection algorithm outperforms other competing methods.

Conclusions: We should not just insist on sensitivity in the segmentation phase because if we forgot FP rate, and our goal was just higher sensitivity, then the learning algorithm would be biased more toward false positives and the sensitivity would decrease dramatically in the false positive reduction phase. Therefore, we should consider the mass detection problem as a cost sensitive problem because misclassification costs are not the same in this type of problems.

KEY WORDS: Classification, detection, FROC analysis, mammograms, masses

Ehsan Kozegar[1], Mohsen Soryani[1], Behrouz Minaei[1], Inês Domingues[2]

Department of Computer Engineering, Iran University of Science and Technology (IUST), Tehran, Iran, [2]INESC TEC, Faculty of Engineering, University of Porto, Portugal

For correspondence:
Ehsan Kozegar,
Department of Computer Engineering, Iran University of Science and Technology (IUST), Narmak, Tehran, Iran
E-mail: Ehsan. koozegar@gmail.com

INTRODUCTION

Breast cancer is one of the most lethal diseases in various countries especially Western countries. Several reports about the outbreak and severity of breast cancer are published by different organizations.[1] According to some reports, breast cancer is the second most common disease after lung cancer (10.9% of cancer incidence in both men and women) and the fifth most common cause of cancer death.[2] The National Breast Cancer Foundation has estimated that 200,000 people suffer from the disease and 20,000 die every year. Furthermore, according to American National Cancer Institute, every 3 minutes one woman is diagnosed with a cancerous case and every 13 minutes, one woman is killed by the disease.[2]

Among various modalities, mammography is the most popular method to detect different abnormalities in breasts. During the past two decades, many scientists have been attempting to help radiologists in the detection and diagnosis of these anomalies. It is, however, important to note that Computer Aided Diagnosis (CAD) systems are designed to assist radiologist only as a second interpretation and never as a substitute.

Masses and microcalcifications (MCCs) are the two most frequent findings in mammograms. Detecting masses is more difficult than detecting MCCs because mass features can be ambiguous or similar to breasts' parenchyma. Masses are usually located in the dense regions of the breast. Furthermore, they have smoother boundaries than MCCs and more various shapes as well. These factors make mass detection a challenging problem both for humans (radiologists) and machines (CAD systems). It is reported that most abnormalities missed by radiologists are related to cancerous masses.[3] Most of the available commercial CAD systems to detect MCCs have reached 100% of detection rate, but the detection rate of masses is still below 90%.

Mass detection has a vital role in full CAD systems, and many studies have been made during the past two decades. Some recent studies are now reviewed. A well-known filter called Density-Weighted Contrast Enhancement (DWCE) was proposed by Petrick et al.[4] This filter is designed to remove background structures and enhance potential signals in which two nonlinear filters are applied to the original mammogram consequently. After applying DWCE filter, the

Laplacian of Gaussian (LoG) filter is used to detect potential edges of masses. Another well-known filter, which is based on edge detection, is known as iris filter and was proposed by Kobatake *et al.* and then developed by Varela *et al.*[5,6] Iris is used to enhance the rounded lesions in mammograms. Then, suspicious regions are localized by a simple adaptive thresholding whose threshold is considered as a proportion of the Cumulative Distributive Function (CDF). Kom *et al.* proposed a linear filter to enhance mammograms in the first step, and then designed a local adaptive thresholding in which a large window and a small window are located around a pixel to calculate a threshold value for the pixel.[7] This process is applied for all pixels in the image. Various approaches can be used instead of a rectangular neighborhood, for example, Fauci *et al.* proposed a method called Region Of Interest (ROI) hunter to detect massive lesions in mammograms in which concentric rings are considered as the neighborhood of a local maximum within a cell.[8] Eltonsey *et al.* proposed a new method to detect cancerous masses in which numerous gray levels in mammograms are diminished to fewer numbers called granule levels.[9] Martins *et al.* developed a method based on combining a clustering algorithm called growing neural gas (GNG) and Ripley's K function.[10] Nguyen *et al.* used a fixed rounded template to locate seeds of suspicious regions.[11] Then the classic region growing was applied to the obtained seeds. Luchanambal *et al.* proposed a method based on LoG filter to enhance edges in the image, and then an increasing template, which was proposed by Lai *et al.*, was used.[12,13] In[14] a modified phase portrait analysis method was introduced, based on the eigenvalue condition number and an eigenvalue intensity map. The method uses an iterative and tissue density-adaptive segmentation procedure with extraction of geometric features. False-positive reduction is accomplished using a fuzzy inference-based classifier. Recently, a technique based on bootstrap and morphological operators was proposed.[15] At the initial stage, median filter is applied to remove the noise, and unsharp masking techniques are used to enhance the quality of the mammograms. Then, the Expectation Maximization Bootstrap Subgroup is employed to detect suspicious pixels. In bootstrap technique, the pixel values are considered as universal population. Therefore, each pixel is taken into consideration to detect suspicious pixels. Finally, binary morphological operators and eight-connected component labeling methods are employed to reconstruct the shape, to remove isolated pixels, and to segment the suspicious regions.

The main contributions of this work are:
- The proposal of new segmentation method, inspired by binary search;
- The inclusion of a sampling block that greatly reduces the bias of the classifiers;
- A feature selection methodology that has never applied in the mass detection field;
- The classifier as an ensemble of three powerful ensembles. It is shown in results section that, when combined, the performance increases;
- The system outperforms other published methods.

The remaining of the paper is organized as follows: Materials and Methods section details the Material and Methods, in particular the proposed mass segmentation algorithm, the sampling methodology, set of features extracted, feature selection scheme, and classification ensemble. The paper proceeds with the Experiments and Results in Results section, where details on the used database are given and each one of the above steps are illustrated. The paper concludes in Discussion section with some final remarks and future lines of research.

MATERIALS AND METHODS

The proposed mass detector consists of six major steps as shown in Figure 1.

The main objective of preprocessing is contrast enhancement. In this step, all images are first resized to 512×512 pixels by cubic interpolation. Then, a 3×3 median and a Contrast Limited Adaptive Histogram Equalization (CLAHE) filter are used for impulse noise reduction and contrast enhancement, respectively. In the segmentation step, several suspicious regions are obtained from each image, which may contain a mass. The segmentation algorithm returns several False Positives (FPs) that need to be discarded. In order to do that, a machine learning approach was used where we start by extracting some region descriptors. These descriptors are feed into a classification algorithm that gives the final assessment. The sampling block deals with the imbalance problem caused by the use of an over-segmentation algorithm. The remainder of this section will expand upon the main steps.

Global thresholding techniques are rarely used in mammograms because there are many structures in mammograms that might be a mass but cannot be revealed with only one threshold. Generally, global thresholding techniques have poor performance in comparison with local techniques. On this basis, a local adaptive thresholding is proposed in this section, which is inspired by binary search to determine an appropriate threshold related to each local region (called cell). The flowchart of the mammogram segmentation algorithm
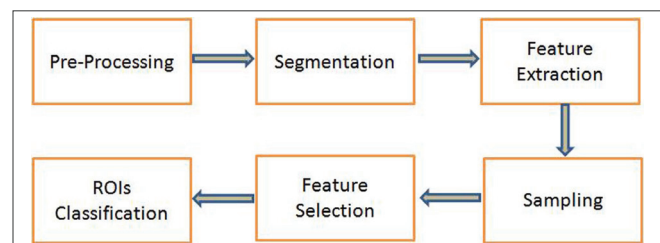

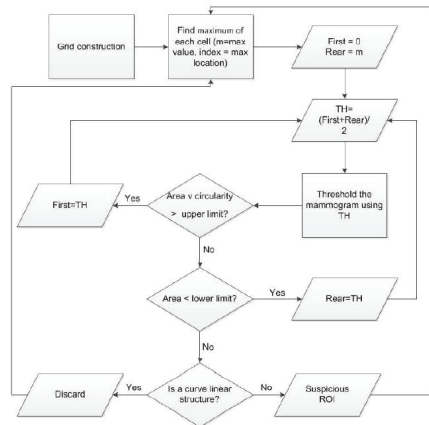**Figure 1:** Block diagram of the proposed mass detection algorithm

Figure 2: Flowchart of the mammogram segmentation algorithm

Figure 3: Segmentation example. Left: original mammogram; right: six regions obtained after segmenting the original mammogram

(applied to each cell of the grid respectively) is shown in Figure 2.

In the first step, each image is divided into equal non-overlapping cells (a grid). In each cell of the grid, the pixel with maximum gray level is found. The location of the maximum pixel is shown as Index and its value named m. As mentioned above, we are seeking an appropriate threshold inspired by binary search. First and Rear are the bounds of the range, which is being explored and TH is the proper threshold. First and Rear are initialized with 0 and m, respectively. In the first iteration, TH is assigned with the middle value of the range [0;m] and then the threshold is applied to the whole mammogram. After that, the circularity measure is extracted from the region that contains index:

$$\text{Circularity} = P^2/4\pi A \qquad (1)$$

In Equation (1), P is the circumference of the region and A is the area. In this equation, maximum circularity is 1 and the less circular the region, the smaller the circularity value will be. If the area or the circularity of that region exceeds the corresponding upper limit ($Area_{max}$, $Circ_{max}$) then we should search within the upper half of the previous range (i.e., [TH, Rear]). Else, if the area of the region is lower than a threshold ($Area_{min}$), then the TH is too high and we should search the proper TH in range [First; TH]. These instructions are iterated until a region with the area between $Area_{min}$ and $Area_{max}$ and also less circularity than $Circ_{max}$ is found (if it exists). In fact, masses generally have a radius between a lower limit and an upper limit and are also not very irregular. Although spiculated masses are irregular in shape, their circularity can be lower than a predefined limit and the irregularity occurs in the margins of those masses and have a fairly small effect on the whole circularity of those regions (in fact, the circularity of a mass cannot be more than 50). In this paper, $Area_{max}$ and $Area_{min}$ are set to 8000 and 155 pixels, respectively, because there is no mass that has a radius greater than50 pixels or smaller than 7 pixels. In addition, $Circ_{max}$ was empirically set to 7. All
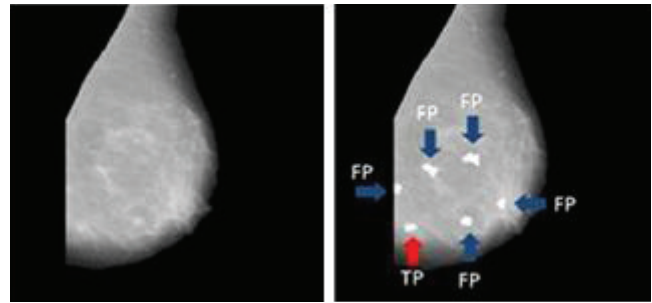
of the mentioned parameters were set on scaled images (512 × 512 pixels).

Moreover, we defined another measure as Area/Circularity to filter curve-linear structures such as blood vessels and milk ducts in mammograms. If a region has a value lower than a predefined threshold, it is considered as a curve-linear object and discarded. Most of the regions extracted in the previous step are in fact normal regions that are mistakenly detected as masses. An example of the segmentation algorithm is shown in Figure 3.

The fact that the number of FPs is almost 20 times the number of true positives (TPs), makes the training set highly imbalanced. Generally, machine learning techniques have difficulties with imbalanced databases, producing classifiers biased toward the major class. It is intended in this section to reduce the bias by sampling strategies. Sampling strategies are divided into two general categories: (1) under-sampling (2) over-sampling. In the first category, samples of the major class were decreased and in the second one, samples of the minor class were increased. In this paper, we use an over-sampling technique called Synthetic Minority Oversampling Technique (SMOTE) proposed by Chawla *et al.*[16] In this technique, depending upon the amount of over-sampling, neighbors of the K nearest neighbor are randomly chosen and the feature vector of the minor class samples is partially changed with regard to their K nearest neighbors. In this way, some synthetic samples are created and added to the database with the same label as the minor class.

As shown in the previous section, the segmentation algorithm returns a significant number of FP. In order to reduce them, some features were extracted from each region. These features can be categorized into five groups: (1) intensity features; (2) ranklet features; (3) fractal dimension; (4) Gray Level Cooccurrence Matrix (GLCM) features; and (5) Local Binary Pattern (LBP) features.

Intensity features
Given a ROI, average intensity, standard deviation, kurtosis, skewness, and entropy are calculated.

Ranklet features

The Ranklet transform was introduced for texture classification by Masotti and Campanini.[17] By analyzing four resolutions, 2, 4, 8, and 14, and three orientations (horizontal, vertical, and diagonal), 12 ranklet images result from each region. Moreover, two additional features called mean convergence (MC) and code variance (CV) are extracted from each ranklet image according to the method by Masotti and Campanini.[17]

Fractal dimension

The fractal dimension is calculated from equations of upper blanket and lower blanket that have been proposed by Yang *et al.*[18] If there are sudden gray-level changes such as edges in an image, fractal dimension will be smaller. However, if there are a few changes, the fractal value will be larger.

GLCM features

GLCM has been widely used in texture analysis. A unique matrix is constructed with respect to a displacement factor (d) and a rotation angle (θ) corresponding to each image. The displacement parameter is set to 1 pixel and four directions (0, 45, 90, and 135 degrees) are used to extract features from each ROI.

As shown in Table 1, a list of 22 features include 14 Haralick *et al.* features, 5 features proposed by Xu *et al.*, plus 3 features that exist in MATLAB 2010.[19,20] Note that some features are the same with respect to their names in Table 1 but they are different with each other with respect to their definitions.

LBP features

The last group of features is based on LBP operator. The LBP operator is a theoretically simple and powerful operator to analyze textures. In this step, we use the $LBP^{riu2}_{P,R}$ operator, which is invariant to gray scale and rotation. $LBP^{riu2}_{P,R}$ has P + 2 outputs and is implemented with a 2^p lookup table. If the histogram of these operators is calculated for each ROI, then a histogram with P + 2 bins will be obtained where each bin of the histogram can be a new feature. In this paper, three histograms are obtained related to the pairs (P, R): (8, 1), (16, 2), and (24, 3). In this way, 54 features are obtained from LBPs

calculated for each ROI and added to the total feature vector.

After the feature extraction step, d = 1446 features were obtained for each detected region. In the feature selection step, we want to find the k features that give us the most information and discard the other (d − k) features. There are several reasons why we are interested in reducing dimensionality:

- In most of the machine learning algorithms, complexity of the algorithms depends on the dimensionality (d), as well as on the size of data samples (N). Therefore, decreasing d results in decreasing the time and space complexity of that learning algorithm.
- When a feature is identified as a redundant feature, we do not need to extract it.
- Simpler models are more robust on small datasets, that is, they will be less sensitive to noises and outliers.
- Irrelevant features decrease the performance of the classifiers.

In this work, information gain (IG) was used. This measure is based on entropy and is defined as:

$$GAIN_{spit} = Entropy(P) - (\sum_{i=1}^{K} \frac{n_i}{n} Entropy(i)) \tag{2}$$

Where P is the feature set, K is the number of intervals each feature is divided into (K = 2 in the current application), n_i is the number of samples in each child, n is number of samples and

$$Entropy(t) = -\sum_{jp} (j \mid t) \log (p(j \mid t) \tag{3}$$

Although mass detection is naturally a cost sensitive problem, almost all methods that use pattern recognition algorithms for FP reduction ignore that fact. In this way, if a real mass is classified as a normal tissue, the cost is much higher than when a normal region is classified as a mass. In the first condition, the patient would not undergo treatment and the mass will gradually grow and spread. In the second one, the patient would be referred to pathology and the suspicious region would be diagnosed as a normal tissue after biopsy. These notions can be translated into a cost matrix like the one in Table 2.

If the classifier's prediction is correct, there will be no cost, otherwise there will be some cost that is described by C1 and C2, where C1 is greater than C2.

A meta-classifier called MetaCost is used to convert the classifier to a cost sensitive classifier using a predefined cost matrix. MetaCost uses a kind of Bagging ensemble to estimate

Table 1: Features extracted from the GLCM matrix

Haralick *et al.*[19]	Xu *et al.*[20]
Contrast	Autocorrelation
Correlation	Cluster prominence
Difference entropy	Cluster shade
Difference variance	Dissimilarity
Energy	Max probability
Entropy	
Homogeneity	
Information measure of correlation 2	**MATLAB**
Information measure of correlation 1	Contrast
Max correlation coefficient	Correlation
Sum average	Homogeneity
Sum entropy	
Sum variance	
Variance	

Table 2: Cost matrix

		Predicted class	
		Mass	**Nonmass**
True class	Mass	0	C1
	Nonmass	C2	0

the class probability of each sample in training, so that the conditional risk is minimized. The conditional risk R (i│x) is the expected cost of predicting that x belongs to class i:

$$R(i\,|\,x)=\sum_j P(j\,|\,x)C(i,j) \qquad (4)$$

Where p (j│x) is the probability that sample x belongs to class j (calculated according to bagging procedure in meta-cost) and C (i, j) is the element of cost matrix.

In summary, the MetaCost procedure works in this way: multiple bootstraps are replicated from training set, and a classifier is trained on each training set (this training step is very similar to bagging), and then the class probability (p (j│x) in Equation (4) for each sample is estimated by votes of base classifiers in the ensemble; each example of the training set is re-labeled using the equation above. These labels are class labels which are estimated as optimal classes using conditional risk equation. Finally, the classifier is reapplied to the re-labeled training set. One of the most important advantages of MetaCost is that if the cost matrix changes, only the final learning step must be repeated.

Our proposed classifier as the base classifier of MetaCost is a powerful ensemble classifier that combines three ensembles (AdaBoost, Bagging and Rotation Forest- RF) using a weighted majority vote strategy. The base classifier used both in Rotation Forest and Bagging is decision tree and in AdaBoost is Naive Bayes with the default WEKA settings. We constructed an ensemble of ensembles to make it cost sensitive [Figure 4].

Therefore, we have an ensemble of ensembles that is wrapped by MetaCost. We can change the proportion C1/C2 to reach our desired FP rate against sensitivity. If this proportion is increased, the sensitivity of the mass detector will be increased (a positive point) and the FP rate will be increased too (a negative point). Nevertheless, there is a trade-off between these two measures, which are dependent on the radiologists' point of view.

RESULTS

In this step, all modules of our mass detector are going to be assessed on mini-MIAS and INBreast databases. mini-MIAS consists of 330 images so that every pixel is of size 1024 ×

Figure 4: Proposed classifier to reduce FPs

1024.[21] These 330 images include 209 normal images, 56 images with at least one mass, and the remaining have other types of anomalies. One of the images (mdb059) was discarded in our experiments because there is no information about the center of the mass present in the mammogram. INBReast database has a total of 115 cases (410 images) of which 90 cases are from women with both breasts (4 images per case) and 25 cases are from mastectomy patients (2 images per case).[22] Several types of lesions (masses, calcifications, asymmetries, and distortions) are included. In this work 107 images were used (all images with at least one mass) with a total of 116 masses. Note that, while mini-MIAS is a well-known database, with the advantage of being already used in several published works, it is a small database of digitized mammograms having only the center and radius information about the findings location. INBreast, however, is a recent database having the disadvantage of not being used by many works yet making it more difficult to compare among different algorithms. It has, as advantages, the fact all the images are Full-field digital mammograms and it has accurate information on the form of detailed contours on the shape and location of every finding. We used mini-MIAS dataset to compare our results with other competing methods. Moreover, we evaluated our detection algorithm on INBreast dataset to show our detector's robustness and effectiveness.

All experiments were performed in a Dell 1520 laptop with a 2.66 GHz CPU and 4 GB of RAM.

The CLAHE contract enhancement effect is shown in Figure 5.

Researchers normally use three different rules to label each region as TP or FP. If a loose rule is used to tag each region, many regions are considered as a TP, but in fact, they are not well-matched to radiologist-drawn regions. A bounding box is considered around the ROI to tag a region in the first step. (LX, LY) is supposed to be length and width of the bounding box and (X_{cad}, Y_{cad}) is the region's center of gravity and $(X_b,$
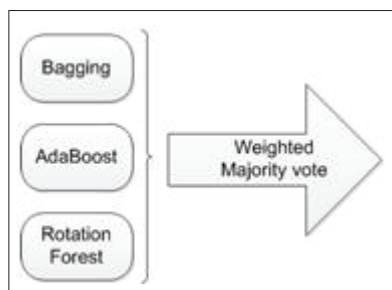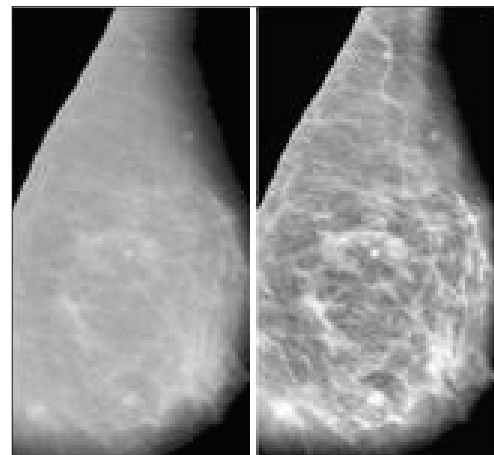

Figure 5: Enhancement results. Left: original mammogram (mdb312); right: result after applying CLAHE filter
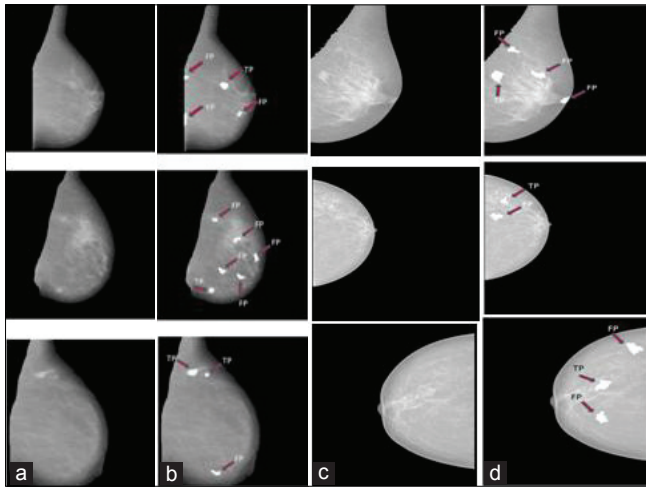
Figure 6: Segmentation examples (a) original mammograms from mini_MIAS, (c) original mammograms from sssinbreast, (b) and (d) regions obtained after segmenting the original mammograms

Y_b),R_b is center and radius of biopsy-proven mass. We combine three typical rules to create a very strong tagging rule. These rules are as follows:

- c_1: if $|Y_b - Y_{cad}| < \max (R_b, LY/2)$ and $|X_b - X_{cad}| < \max (R_b, LX/2)$
- c_2: if $(X_{cad} - X_b)^2 + (Y_{cad} - Y_b)^2 \leq R_b^2$
- c_3: if exists 50% overlap between biopsy proven mass and suspicious region

An extracted ROI would be labeled as TP if all above rules were true. Otherwise, even if one of them were false, that ROI is labeled as FP, as illustrated in Figure 6.

In the segmentation step, there is a trade-off between sensitivity and FP rate. However, we consider that sensitivity is more important than FP rate because we can reduce the FP rate in the classification phase. In Table 3 some segmentation methods were implemented for comparison in which the first, second, and last methods were implemented by us and the remainder were implemented by Oliver *et al.*[23] There, seven mass detectors were compared: (a1) based on a detection of concentric layers, (b1) a Laplacian edge detector approach, (c1) thresholding, (c2) Iris filter, (c3) a Difference of Gaussians, (d1) a pattern matching approach, and (d2) a classifier approach.[23]

Only the sensitivity of DWCE filter is comparable with the sensitivity of our mass segmentation. But, with almost the same sensitivity, our method results in fewer FPs per image (4.77FP per image) than DWCE filter (12 FP per image).

The proposed methods by Eltonsey *et al.*, Kom *et al.* and DoG filter have lower sensitivity and also more FPs in comparison with our method.[7,9,23] Although Nguyen's method results in less FP, its sensitivity is very low, making it unreliable and thus not suitable to assist radiologists.[11] Nguyen *et al.* used a rounded fixed template to detect different masses (which

Table 3: Results of different segmentation algorithms tested on mini-MIAS dataset

Segmentation method	Sensitivity	FP rate
Template matching[11]	0.38	2.90
Region growing[9]	<0.60	>8.50
Adaptive thresholding[7]	<0.65	>9.00
DWCE filter[4]	<0.90	>12.00
DoG filter[23]	<0.72	>10.50
Proposed segmentation	0.91	4.77

Table 4: Sampling effect in improving the performance of the learner

	Precision	Recall	F-measure	j-coef
No sampling	0.099	0.123	0.103	0.056
Sampling	0.167	0.316	0.218	0.159

have different shapes and size) and we believe this is the reason the method failed. By using the proposed method, the segmentation of 261 mammograms lasted 388 minutes, that is, 1 minute and 48 seconds for each image.

As far as we know, this work is the first published work presenting mass detection results on the INBreast database. We have achieved a sensitivity on INBreast of 87% (the algorithm missed 15 masses) and FP rate per image is 3.67 in the segmentation phase.

There are many measures in the literature to evaluate learning models, which focus on the classifier predictability. One of the most frequently used measures is accuracy but we cannot use it to evaluate our models because it is not adequate for imbalanced problems. Given a two class dataset with 9990 nonmass and 10 mass, suppose a classifier predicts every sample as nonmass. It is obvious that the classifier performs poorly. But, according to that measure, the classifier has an accuracy of 99.9%. As accuracy is a misleading measure, we have decided to use cost sensitive measures like precision, recall, and F-measure. We use F-measure as our principal method to evaluate our models because it combines precision and recall in one equation. Moreover, we use another measure called j-coefficient as the second principal method (it combines all four elements of the confusion matrix). WEKA was used to analyze different machine learning algorithms.[24]

As mentioned in Section 2.2, we used the SMOTE technique to increase samples of the minor class and reduce the bias of the classifier toward the major class. The sampling rate was set to 2000 so that the number of samples in both classes is equivalent. Note that even if number of samples is the same in both classes, the cost of miss-classifications is different. Therefore, we use a fixed cost matrix in which the proportion of C1/C2 is set to 5. As shown in Table 4, the performance of the classifier more than doubled after applying the oversampling method.

As explained in Section 2.4, five groups of features were extracted from each suspicious ROI. In this step, 10-fold

Table 5: Discriminate ability for each group of features separately and when combined

Features	Precision	Recall	F-measure	j-coef
Intensity	0.051	0.842	0.097	0.060
GLCM	0.051	0.912	0.097	0.069
Fractal dimension	0.045	0.982	0.087	0.033
LBP	0.065	0.684	0.119	0.097
Ranklet	0.137	0.246	0.176	0.133
All together	0.167	0.316	0.218	0.159

Table 7: Performance of the proposed ensembles when used individually or combined

	Precision	Recall	F-measure	j-coef
AdaBoost	0.180	0.300	0.225	0.187
Bagging	0.223	0.193	0.207	0.174
RF	0.254	0.123	0.165	0.151
Combined	0.307	0.222	0.257	0.233

Table 6: Impact of feature selection

	Precision	Recall	F-measure	j-coef
IG	0.180	0.386	0.246	0.215

Table 8: Operating points of some mass detectors reported by Oliver *et al.*[23]

	Sensitivity	FPs per image
B1	0.790	4.1
	0.910	8.1
C2	0.806	3.8
	0.895	7.3
D1	0.754	6.0
D2	0.823	6.3

cross validation is used to evaluate the discriminate ability of each group separately. The capability of each group for discriminating mass and nonmass regions is shown in Table 5.

Although the intensity features, GLCM features and fractal dimension have a high recall rate, their precision is low, which results in a low F-measure. In contrast, the ranklet features have the opposite effect: they have less recall but higher precision than the other groups. As shown in the last row of Table 5, a higher performance was obtained by combining features due to the higher diversity in the feature space.

When examining the discriminate ability of ranklet features, we encountered memory limitations, even when assigning the maximum space that a Windows 32 bit can allocate to a process. For this reason, we run our experiments on a Windows 64 bit machine with 16 GB RAM. Therefore, we used feature selection methods to reduce time and space complexity and also improve the classifier performance. Some of the methods tested were Best First Search (BFS), Correlation Feature Selection (CFS), Genetic Algorithms (GA), IG, and Principal Component Analysis (PCA). Due to space constrictions, only the best one is presented in Table 6.

By comparing the last row of Tables 5 and 6, we find that after using IG feature selection, the results improve and the performance measures are increased due to the removal of irrelevant features, which degrades the classifier predictability.

We now address the performance of the proposed ensemble classifiers when used together. We apply 10-fold cross-validation three times and results are reported according to the average and standard deviation of the performance measures [Table 7]. Experimental results show that each training phase takes about 40 minutes. Note that the radiologist does not have to wait 40 minutes for testing each image because the training phase can be done in offline mode, that is, the system

can be trained once and used many times. When the database becomes large enough, the system can be trained again. It should be mentioned that the classification of each image takes at most 1 minute in testing phase. Therefore, when a new mammogram is given to the system, it takes at most 3 minutes to detect masses (2 minutes for segmentation and 1 minute for feature extraction and classification).

When using AdaBoost alone, the recall is fairly high but the precision is lower. In contrast, when using RF alone, the precision is high but the recall is low. To improve the performance, RF, Bagging, and AdaBoost are combined. In this way, F-measure increases. We must take into account that diversity is a critical criterion for ensemble construction. It is thus better to combine classifiers that have low correlation. In this work, RF was included due to its capacity to recognize FPs. In contrast, AdaBoost is a classifier that recognizes TPs. It is important to note that other classification methodologies, like Support Vector Machines (SVMs), Multilayer perceptron and Decorate, were tested having the here proposed ensemble of ensembles the best performance. In particular, when SVM was added to the ensemble, the F-measure degraded. When the SVM is used with any classifier in the ensemble (AdaBoost, Bagging, or RF) the performance was degraded. The same behavior was observed when instead of SVM, Multilayer perceptron or Decorate were used.

In the last step, our mass detector is compared with other mass detectors using a Free-response Receiver Operating Characteristic (FROC) curve. The parameter for constructing the FROC curve is the C1/C2 ratio in the cost matrix of MetaCost classifier. When the C1/C2 ratio increases, the classifier will be biased toward TPs and therefore its sensitivity will increase too. In contrast, its FP rate will increase. The trade-off between sensitivity and FP rate can be changed according to the radiologist's point of view. By changing this ratio, the FROC curve is constructed. In 2010 Oliver implemented several mass detectors and tested them on mini-MIAS dataset.[23] According to this survey, researchers can compare their mass detectors with some well-known detectors. He performed a 10-fold cross validation three times on 261 mammograms including normal and mass-contained images. Three FROC curves were
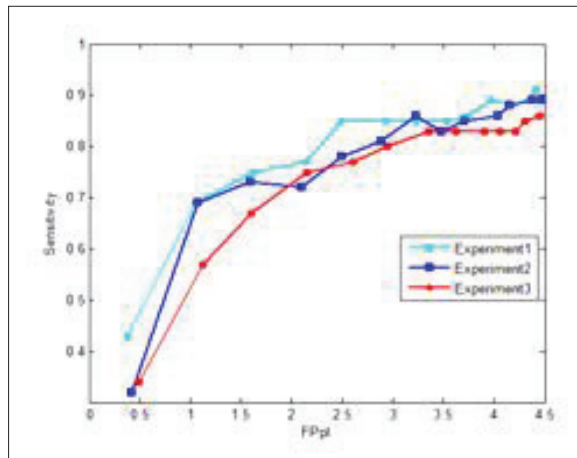
Figure 7: FROC curves obtained in three runs of 10-fold cross validation
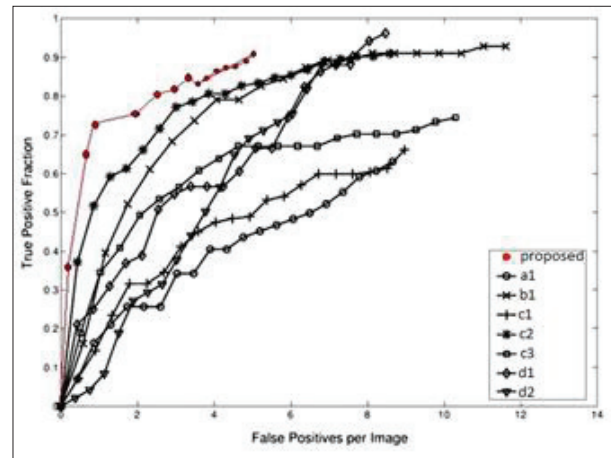

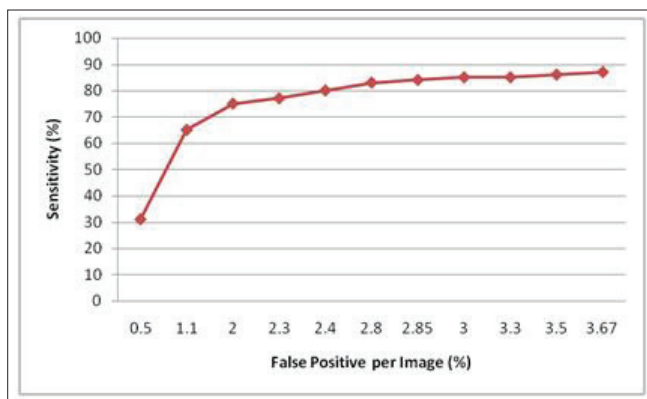**Figure 8:** Comparison between Oliver FROC curves and ours

Figure 9: FROC curve of our mass detection algorithm obtained from INBreast database

constructed and the average of these curves is the ultimate FROC curve, which can be compared with others. Our three FROC curves are shown in Figure 7.

The FROC curves obtained by Oliver are shown in Figure 8. The red curve is the average FROC curve of our mass detector.

As can be seen in Figure 8, the sensitivity of a1, c1, and c3 barely reach 70% while our detector can obtain 90% sensitivity. Moreover, our mass detector has 90% sensitivity with 4.5 FPs per image on average, while the other methods have at least 8 FPs at their maximum sensitivity. d2 at its maximum sensitivity (nearly 85%) has 8 FPs per image. In summary, our detector has greater sensitivity and a lower FP rate than the others. We can also compare FROC curves according to their best operating point, which is a qualitative point that may be different in different experts' point of view. Some operating points are reported by Oliver [Table 8]. Note that our mass detector leads to fewer FP taken the same sensitivity. For example, c2 lead to 3.8 FP in 80% sensitivity but our detector has 2.5 FPs with the same sensitivity. In summary, our detector results in fewer FP independent of the sensitivity. In addition, the FROC curve on INBreast database is shown in Figure 9.

DISCUSSION

In this paper, we proposed a mass detection algorithm that outperforms other implemented methods according to FROC analysis. First, an effective iterated segmentation method was proposed so that a high sensitivity of 91% with small false positive rate of 4.8 FPs per image were obtained. In addition, we proved that a good selection method is necessary for false positive reduction because irrelevant features degrade the performance of classifiers. Furthermore, almost all machine learning approaches are faced with imbalance dataset problems in false positive reduction phase. When classifiers encounter imbalance datasets, they could be biased toward major class. We improved our classifier's performance by a sampling method called SMOTE. Finally, we propose to ensemble classifiers instead of using individual classifiers because of their better performance.

In future work we will study other texture features, such as Ripley's K function. To apply the False Positive reduction module proposed in this paper on BI-RADS classification is foreseen.

REFERENCES

1. De Oliviera J, De Albuquerque A, Deserno TM. Content-based image retrieval applied to BI-RADS tissue classification in screening mammography. World J Radiol 2011;3:24-31.
2. Oliveira ML, Braz JG, Cardoso P, Gattass M. Detection of masses in digital mammograms using K-means and support vector machine. ELCVIA 2009;8:39-50.
3. Malagelada AO. Automatic mass segmentation in mammographic images, Ph.D. dissertation, Department of Electronics-Computer Science and Automatic Control. University of Girona, Spain; 2007.
4. Petrick N, Chan HP, Sahiner B, Wei D. An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection. IEEE Trans Med Imaging 1996;15:59-67.
5. Kobatake H, Murakami M, Takeo H, Nawan S. Computerized detection of malignant tumors on digital mammograms. IEEE Trans Med Imaging 1999;18:369-78.
6. Varela C, Tahoces PG, Mendez AJ, Souto M, Vidal JJ. Computerized detection of breast masses in digitized mammograms. Comput Biol Med 2007;37:214-26.

7. Kom G, Tiedeu A, Kom M. Automated detection of masses in mammograms by local adaptive thresholding. Comput Biol Med 2007;37:37-48.

8. Fauci F, Raso G, Magro R, Forni G, Lauria A, Bagnasco S, *et al.* A massive lesion detection algorithm in mammography. Phys Med 2005;21:23-30.

9. Eltonsy NH, Tourassi GD, Elmaghraby AS. A concentric morphology model for the detection of masses in mammography. IEEE Trans Med Imaging 2007;26:880-9.

10. Martins LD, Paiva AC, Gattass M. Detection of breast masses in mammogram images using growing neural gas algorithm and ripleys k function. J Signal Process Syst 2009;55:77-90.

11. Nguyen VD, Nguyen DT, Nguyen TD, Thi N, Tran DH. A program for locating possible breast masses on mammograms. In: Proceeding of the 3rd International Conference on the Development of BME, Springer, Vietnam, 2010. p. 11-4.

12. Lochanambal KP, Karnan N, Sivakumar R. Identifying masses in mammograms using template matching. In: Second International Conference on Communication Software and Networks, IEEE, Singapore 2010.

13. Lai SM, Li X, Bischof WF. On techniques for detecting circumscribed masses in mammograms. IEEE Trans Med Imaging 1989;8:377-86.

14. Mencattini A, Salmeri M. Breast masses detection using phase portrait analysis and fuzzy inference systems. IJCARS 2011;7:573-83.

15. Singh WJ, Nagarajan B. Automatic diagnosis of mammographic abnormalities based on hybrid features with learning classifier. Comput Methods Biomech Biomed Engin 2012.

16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-57.

17. Masotti M, Campanini R. Texture classification using invariant ranklet features. Pattern Recognit Lett 2008;29:1980-6.

18. Yang SC, Wang CM, Chung YN, Hsu GC, Lee SK, Chung PC, *et al.* A computer-aided system for mass detection and classification in digitized mammograms. Biomed Eng-App Bas C 2005;17:215-28.

19. Haralick RM, Shanmugan K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern 1973;3:610-21.

20. Xu R, Zhao X, Li X, Kwan C, Chang C. Target detection with improved image texture feature coding method and support vector machine. IJIT 2006;1:47-56.

21. Suckling J. The mammographic image analysis society digital mammogram database. In: Exerpta Medica International Congress Series, Elsevier, New York 1994;1069. p. 375-8.

22. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: Towards a full field digital mammographic database. Acad Radiol 2012;19:236-48.

23. Oliver A, Freixenet J, Martí J, Pérez E, Pont J, Denton ER, et al. A review of automatic mass detection and segmentation in mammographic images. Med Image Anal 2010;14:87-110.

24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software. SIGKDD Explorations 2009;11:10-8.