Iranian Journal of Radiology
Volume 18 · Issue 3 · e106717

OPEN PEER REVIEW

An Efficient Method for Automated Breast Mass Segmentation and Classification in Digital Mammograms

Behrouz Niroomand Fam, Alireza Nikravanshalmani, Madjid Khalilian

SUBMITTED	REVISED	ACCEPTED	DOI
22 Jun 2020	11 May 2021	15 May 2021	10.5812/iranjradiol.106717

About this report

This open peer review (OPR) report shows the public comments made on this article during peer review, by the reviewers, the associate editor and the editor-in-chief, for every revision round, together with the authors' replies. Comments appear only with the consent of the people who wrote them. Reviewers who chose to sign their report are named; the others are shown as "Referee". Comments are edited for language only, and confidential remarks to the editors are not included.

More about open peer review: brieflands.com/knowledgebase/opr

Contents

1 **Original submission** Referee 1, Editor-in-Chief 2
2 **Revision 1** Referee 1, Referee 2, Editor-in-Chief 4
Authors' reply to reviewers 5
Reviewer and editor comments 16
3 **Revision 2** Referee 2, Associate Editor, Editor-in-Chief 19
Authors' reply to reviewers 20
Reviewer and editor comments 32
4 **Revision 3** Editor-in-Chief 33
Authors' reply to reviewers 34
Reviewer and editor comments 36

1 Original submission

2 comments · 10 Aug 2020 – 15 Aug 2020

Referee 1 Reviewer

MAJOR REVISION 10 Aug 2020

1-Please add a background section to the abstract according to the journal format. In addition, please add an objective section to the paper.
2-In line 34, “issues” seems to mean tissues. Please correct.
3-Did the authors use a radiologist's opinion and drawing as the gold standard? It is unclear what the gold standard is.
4-The mammography BIRADS lexicon considers characteristics of shape, margin, and density for scoring a breast mass. Here, the authors concentrate only on shape. If the authors process, for example, the edge and margin characteristics with appropriate tools such as wavelets, they can extract more discriminating features for masses. Why did the authors concentrate only on the mass shape?
5-A mammogram should be taken under standard conditions in which the radiation is set based on the size of the breast. If the standard conditions have not been followed, the image will be taken with excessive or insufficient radiation. Can this affect the efficacy of the method described in the paper? [I mean that if the radiation dose is higher or lower than the correct dose, it could affect the diagnostic precision and efficacy of the method.]
6-In experiment 1, the authors have mentioned Sensitivity = (16), FPPI = (17), Precision = (18), and Recall = (19), but 16, 17, 18, and 19 are unclear. Please clarify.
7-It is unknown whether the diagnostic efficacy indices for mass detection are calculated for 1592 masses, benign masses, or malignant masses. Please clarify.
8-In figure 7, the authors mentioned a sensitivity of 1 in part C, but the sensitivity in this part is lower than 1. Please explain and correct. [There are some false-negative pixels in the figure.]
9-829 benign + 764 malignant equals 1593, not 1592.
10-Please add the formulas for calculating FPPI, recall, and precision per image.
11-For mass detection, we can calculate sensitivity, FPPI, recall, and precision per image. The authors mentioned “The results showed that the sensitivity, FPPI, Precision and Recall of the suspicious region localization is 100%, 6.4, 72% and 56%, respectively”. If the authors have calculated these indices per image, then for all 1592 images, we should calculate the mean and standard deviation [SD] of the indices across the entire population. Please explain this in the paper and report the indices as mean ± SD [minimum-maximum].
12-It could be interesting to know the diagnostic indices of mass detection for benign and malignant masses. Therefore, I recommend that the authors calculate the sensitivity, FPPI, recall, and precision in benign and malignant cases separately. In this case, the authors can statistically compare these indices between benign and malignant cases [to show whether there is a significant difference between these two groups]. If the findings differ statistically between benign and malignant cases, it would be interesting if the authors added an appropriate discussion of this to the paper.
13-The authors mentioned C ≥ 80, which I think is wrong. It seems this should be C ≥ 0.8.
14-In experiment 2, how did the authors determine the cutoff points of 0.6, 0.7, and 0.8 for classifying the different groups? [benign, probably benign, probably malignant, and malignant]
15-It is unclear how the authors calculated the sensitivity and specificity in experiment 2 when there are 4 groups [benign, probably benign, probably malignant, and malignant]. In fact, which groups are considered benign and malignant? I mean that the status of the probably benign and probably malignant groups is unclear. Is the probably benign group considered benign or malignant? Please clarify.
16-For calculating the sensitivity and specificity in experiment 2 [differentiating malignant and benign masses], we can consider different cutoff points; for each cutoff point, we can calculate a different set of

sensitivity and specificity values. [First cutoff point: benign cases are considered benign, and all others are considered malignant; second cutoff point: benign and probably benign cases are considered benign, and probably malignant and malignant cases are considered malignant; third cutoff point: benign, probably benign, and probably malignant cases are considered benign, and malignant cases are considered malignant.] For each condition, we can calculate a different set of sensitivity and specificity values. The authors did not specify the cutoff point. In addition, it would be better if they considered all three cutoff points separately and calculated the diagnostic indices for each one separately. In this case, we can perform the ROC analysis and report the AUC, too. This would be a useful adjunct to the paper.
17-In the discussion, the authors have explained the indices used in the paper. It would be better if they added some discussion to the paper and compared the findings with those of similar studies and with other image processing tools introduced earlier for this purpose.
18-Is it possible that the indices would be better or worse if the masses contained microcalcifications? It would be interesting if the authors discussed this. [I know the scope of the study is limited to mass detection, not microcalcification. However, microcalcification is an important diagnostic component of a mammogram, and incorporating data on it would increase the diagnostic indices of the method for discriminating malignant masses.]

Editor-in-Chief

MAJOR REVISION 15 Aug 2020

Dear authors,
Your article needs Major Revision.
Regards

2

Revision 1

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

September 7, 2020

Associate Editor
The Iranian Journal of Radiology

Manuscript ID: 106717

Manuscript Title: Mammogram image analysis, breast mass segmentation and classification

Dear Editor:

Thank you very much for reviewing our manuscript. We also greatly appreciate the reviewers for their complimentary comments and suggestions. Those comments have all been very helpful for revising and improving our paper as well as providing important guidance for our research. We have studied the comments carefully and have made corrections, which we hope meet with approval. The major revised parts are highlighted in red color for your convenience of re-reviewing. The detailed corrections and the responses to your comments are listed below point by point.

Sincerely,

Alireza Nikravanshalmani, Ph.D.
Professor
Department of Computer.
Karaj Branch, Islamic Azad University, Karaj, Iran.

Reviewer 1:

We appreciate that the reviewer's comments. The followings are our point-by-point responses:

1) Please add background part to abstract according to journal format. In addition, please add objective part to the paper.

Response 1: As suggested by the reviewer, we have added the background part to the abstract and the objective part to the revised manuscript, and now it reads as follows:

"**Background:** Masses are the most common signs of breast cancer. Correctly classifying masses in mammogram as malignant or benign is still a big challenge. Automatically detecting and classifying mammographic masses can assist radiologists to diagnosis breast cancer.

2. Objectives

In this paper, we propose a method based on the density of wavelet coefficients and morphological features classifier for automatic mass detection in digital mammograms that aims to help the radiologist in interpreting mammograms."

2) In the line 34, "issues" seems to be tissues. Please correct.

Response 2: We apologize that we did not write this correctly. We corrected it in the revised manuscript.

3) Did the authors used radiologist opinion and drawing as gold standard? It is unclear what is the gold standard?

Response 3: Thank you for your comments. We apologize that the description of this part was not clear. The ground truth available with each mammogram is used for measuring the classification rate. In CBIS-DDSM database, the suspicious region is marked by the trained mammographer are considered as ground truth. These marked regions only are required for the analysis as well as for interpretation and to be considered as ROI. We also rewrote the "4.1" section, and now it reads as follows: "In this study, we have employed a new updated and standardized version of the DDSM database [19], i.e., Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [20] which contains images with standard DICOM format. The CBIS-DDSM database reviewed and annotated by trained mammographers. We employ all mass cases of this database, which contain 1593 mammograms (829 benign and 764 malignant) of 892 cases and a total of 1698 masses with ground truth (GT) masks (i.e., annotations). In addition, The CBIS-DDSM contains pixelwise annotations for the ROIs, e.g., mass shape, BI-RADS assessment (rating (0, 2-5)), as well as lesion's pathology (2 annotations: benign or malignant). In this study, each view was used as a separate image and all images were converted into PNG format and resized to 1024×1024 pixels using bicubic interpolation."

4) The mammography BIRADS lexicon considers characteristics of shape, margin and density for scoring a breast mass. Here, the authors concentrate only on shape. If the authors process for example the edge and margin characteristics with appropriate tools such as wavelet, they can extract more discriminating features for masses. Why did the authors concentrate only on the mass shape?

Response 4: Thank you for your comments. In our view, any mammogram image feature has both advantages and disadvantages for breast mass classification in mammographic

images. Although the number of features selected for breast mass classification reported in literature varies with different CAD approaches employed, it is desirable to use an optimum number of features that will still allow the CAD system to achieve the desired target performance. Moreover, the use of a large number of features will increases computational complexity and memory requirements. In this study, since in the first stage the suspicious mass regions robustly identified and the boundary of this regions are clearly defined, we believe that morphological features (especially compactness feature) are more discriminative and can achieve higher classification accuracy than other features (i.e., margin and density features). The analysis of the shape and border of breast masses is also essential for radiologist in interpreting mammograms. Moreover, the compactness feature is independent of linear transformations, rotation, starting point, and size of the given contour and measures the degree of roughness of a region. We also rewrote the "3.3." section, and now it reads as follows: "When a mass is localized, it is quite difficult to determine whether it is benign or malignant. The masses are characterized by their shape, density and margin properties. A mass shapes as defined by BI-RADS are round, oval, lobular, irregular or have architectural distortion. While mass margins could be circumscribed, obscured, microlobulated and ill-defined. The features extracted from shape and margin of mass are widely used in CAD system in discriminating benign from malignant tumor [12]–[14]. These features are the most significant features with high classification accuracy compared to the other sets of features [15].

Although, many different features have been employed to classify breast masses in the literature, it is desirable to use an optimum number of features that will still allow the CAD system to achieve the desired target performance. Moreover, the use of a large number of features will increases computational complexity and memory requirements.

Based on the above motivations, in this study, we use the mass shape characteristics in order to achieve increased accuracy of measurement results. Round mass that is typically smooth and distinct is likely to be associated with benign breast disease, whereas a malignant tumor usually has a rough boundary and is irregular in shape. The roundness of the mass is calculated using the following equation [16]:

$$C = 1- \frac{4 \times \pi \times A}{P^2} \quad (15)$$

Where C is compactness, P is total number of pixels in the edge or margin of the mass and A its number of pixels or the area of the mass. If the C = 0 then the mass is a circle, otherwise it of other shapes. The compactness is used to measure shape characteristics. A high C value indicates a large perimeter enclosing a small area. The compactness index increase value as the shape becomes more complex and rougher. Moreover, the compactness is independent of linear transformations, rotation, starting point, and size of the given contour and measures the degree of roughness of a region. Therefore, typical malignant masses are expected to have higher values of compactness compared to typical benign tumors. In order to discriminate between different mass shapes and according to radiologists' suggestions, we used specific threshold measures to mass classifying. The threshold is determined by comparisons between different mass shapes in the CBIS-

DDSM dataset based on subjective assessment of mass classification and will be described in the next section."

5) A mammography should be taken in a standard condition in which the radiation is set based on the size of breast. If the standard conditions have not been followed, the image will be taken in over- or under-radiation conditions. Can this affect the efficacy of the method described in the paper? [I mean if the radiation dose is higher or lower than the correct radiation, it could affect the diagnostic precision and efficacy of the method].

Response 5: Thank you for your comments. The proposed method includes two major stages: suspicious region localization and classify these regions. The acquired mammogram from imaging device can directly affect the finding suspicious regions (ROIs), and may change the diagnostic precision and efficacy of the proposed method. In order to make more accurate decisions and obtain better visual interpretation, analysis and classification of mammogram masses, as part of preprocessing, the optimal combination of various enhancement methods (e.g., CLAHE, Guided, etc.) and noise removal are applied on mammogram images. (see Fig. 2, 4 (a-f) and "3.1. Pre-processing" section).

6) In the experiment 1, the authors have mentioned the Sensitivity = (16), FPPI = (17), Precision = (18) and Recall = (19) while the 16, 17, 18 and 19 are unclear. Please clarify.

Response: Thank you for your comments. The formulas were written correctly in the original manuscript on the relevant page. Unfortunately, the formulas were corrupted during the building of the PDF file for submission. We have added the formulas in the revised manuscript.

$$\text{Sensitivity} == \frac{Number(TP)}{Number(TP)+number(FN)} \text{ or } \frac{Number((> \bigcap ROI) \neq \varnothing)}{Number(>)} \quad (16)$$

$$\text{FPPI} = \frac{number\ of\ false\ positive}{number\ of\ images} \quad (17)$$

$$\text{Precision} = \frac{TP}{TP+FP} \text{ or } \frac{ROI\ ¿}{ROI} \quad (18)$$

$$\text{Recall} = \frac{TP}{TP+FN} \text{ or } \frac{ROI\ ¿}{¿} \quad (19)$$

7) It is unknown if the diagnostic efficacy indices for mass detection is calculated on 1592 masses or on the benign masses or malignant masses? Please clarify.

Response 7: We are sorry for the confusion. As explained in the manuscript, we employ all mass cases of CBIS-DDSM database, in total, there are 829 benign and 764 malignant masses in 1593 images. This has been summarized in a new Table 2 in the revised manuscript.

Table 2. Confusion matrix of the proposed method

Total	Malignant	Benign	
829	81	748	**Benign**

	9.8%	90.2%	
764	621 81.3%	143 18.7%	**Malignant**
1593	702	891	**Total**

8) In the figure 7, the authors mentioned sensitivity of 1 in part C while the sensitivity in this part is lower than 1. Please explain and correct. [there are some false negative pixels in the figure]

Response 8: In this study, a novel breast mass segmentation technique was proposed in order to find suspicious regions (SR) on mammograms. The key basis to authenticate or validate the results of the proposed segmentation method is the information provided by radiologists and their opinions, along with the similarity measure between obtain results with ground truth. The purpose of the segmentation stage is to detect all suspicious regions on mammograms, so if a lesion (grand truth) is not present in any of the suspicious regions (no coverage of ground truth, see Fig. 7 (a)), the lesion is lost and therefore it is considered as “False Negative” (sensitivity is obtained 0). In other hand, only the regions detected by segmentation method (full coverage of ground truth (Fig. 7 (b)) or partial coverage of ground truth (Fig. 7 (c))) are further analyzed and classified into four categories. As a result, sensitivity or probability of detection of SR is defined as follows:

$$\text{Sensitivity} == \frac{Number(TP)}{Number(TP)+number(FN)} \text{ or } \frac{Number((>\bigcap ROI)\neq\varnothing)}{Number(>)} \quad (16)$$

Therefore, since in both figure (Fig. 7(b) and Fig. 7(c)) the suspicious region contains part or all of a grand truth the sensitivity of each figure is calculated 1 using Eq. (16). We also rewrote the "4.2" section and now it reads as follows: "Suspicious region localization (SRL) is shown in Fig. 4, 5. The sensitivity, average false positives per image (FPPI), precision, and recall which are used to validate the performance of the proposed segmentation method and compare it with stat-of-the-art methods. True Positives (TP), False Positives (FP), and False Negatives (FN) are defined in Fig. 6. For simplicity, the same notation is used for a region and its number of pixels. For instance, FN refers to the false negative region or to its number of pixels, depending on the context. ROI denotes the suspicious region found by the algorithm. The information provided by radiologists and their opinions, along with the similarity measure between suspicious regions with ground truth (see Fig. 7), is the main basis of validation in this work.

The purpose of the segmentation stage is to extract all suspicious regions (SR) from mammograms, so if a lesion (grand truth) is not present in any of the SRs (Fig. 7 (a)), the lesion is lost and therefore it is considered as “False Negative”. In other hand, only the regions detected by segmentation method (these regions are considered as “True Positive”, Fig. 7 (b, c)) are further analyzed and classified into four categories. As a result, we defined sensitivity or probability of detection of SR as follows:"

9) 829 benign +764 malignant equals to 1593 not 1592.

Response 9: We apologize for the confusion. We corrected it in the revised manuscript.

10) Please add the formulas for calculating FPPI, recall and precision per image.

Response 10: Thank you for your comments. We have added the formulas in the revised manuscript.

11) For mass detection, we can calculate sensitivity, FPPI, recall and precision per image. The authors mentioned “The results showed that the sensitivity, FPPI, Precision and Recall of the suspicious region localization is 100%, 6.4, 72% and 56%, respectively” if the authors have calculated these indices per image, then for all 1592 images, we should calculate the mean and standard deviation [SD] of the indices on all population group. Please explain in this regard in the paper and mention the indices as mean±SD [minimum-maximum].

Response 11: Thank you very much for your suggestion. We rewrote the statistical indices as follows: "The results showed that the sensitivity, FPPI, Precision and Recall of the suspicious region localization is 100%, 6/4 ± 5/3, 0.71 ± 0.02 and 0.56 ± 0.01, respectively"

12) It could be interesting if we know the diagnostic indices of mass detection in terms of benign and malignant masses. So I recommend the authors calculate the sensitivity, FPPI, recall and precision in benign and malignant cases separately. In this case, the authors can compare these indices between benign and malignant cases statistically [to show if there is significant difference between these two groups]. If the findings between benign and malignant cases are different statistically, it would be interesting if the authors add appropriate discussion in this regard in the paper.

Response 12: Thanks for these comments. As suggested by the reviewer, we have repeated the experiment #2 and have calculated the statistical indices for diagnosing benign and malignant cases. We have added a table (table 2) to show how our method is able to distinguish between benign and malignant classes. We also have added a table (table 3) to summarizes the performance of the proposed CAD system in terms of sensitivity, specificity, precision, AUC, and overall accuracy.

Table 2. Confusion matrix of the proposed method.

Total	Malignant	Benign	
829	81 9.8%	748 90.2%	**Benign**
764	621 81.3%	143 18.7%	**Malignant**
1593	702	891	**Total**

Table 3. Classification performance of the .proposed CAD system in CBIS-DDSM dataset

(%)	Index
81.3%	Sensitivity
90.2%	Specificity
88.5%	Precision

82.2%	AUC
85.9%	Overall Accuracy

13) the authors mentioned C ≥ 80 that I think it is wrong. It seems this should be C ≥ 0.8.

Response 13: We are sorry for the confusion. We corrected it in the revised manuscript.

14) In experiment 2, how did the authors determined the cut off points of 0.6, 0.7 and 0.8 for classifying different groups? [benign, probably benign, probably malignant and malignant]

Response 14: Thank you for your comments. Please refer to the response to “Response 4” above. (i.e. last paragraph of the Mass Classification, "In order to discriminate …".)

15) It is unclear how did the authors calculate the sensitivity and specificity in experiment 2 while there are 4 groups [benign, probably benign, probably malignant and malignant]. In fact, what are the groups benign and malignant? I mean the situation of groups probably benign and probably malignant are unclear. Are probably benign group is considered as benign or considered as malignant? please clarify.

Response 15: Thank you for your comments. We apologize that the description of this part was not clear. In this experiment, a prediction of benign and probably benign is considered a positive prediction (benign), and malignant and probably malignant is negative prediction (malignant). Therefore, the results of diagnosing a mass are set as follows:

- True positive (TP): malignant masses are correctly diagnosed malignant.
- False positive (FP): benign masses are incorrectly diagnosed malignant.
- True negative (TN): benign masses are correctly diagnosed benign.
- False negative (FN): malignant masses are incorrectly diagnosed benign.

These have been added in the revised manuscript.

16) For calculating the sensitivity and specificity in experiment 2 [differentiating malignant and benign masses], we can consider different cut off points; for each cut off point, we can calculate a different set of sensitivity and specificity. [first cut off point: benigns are considered as benign and all others are malignant, second cut off point: benigns and probably benigns are considered as benign and probably malignants and malignants are considered as malignant; third cut off points: benigns, probably benigns and probably malignants are considered as benigna and malignants are considered as malignant]. for each condition, we can a different set of sensitivity and specificity. The authors did not determine what is the cut off point? In addition, it is better they consider all three cut off points separately and calculate the diagnostic indices for each one separately. In this case, we can implement the ROC analysis and represent the AUC, too. This would be a useful adjunct for the paper.

Response 16: We appreciate the reviewer's comments and suggestions. We have discussed the definition of benign and malignant groups in Section 4.3. Please refer to the response to “comment 15” above. We have followed this suggestion and performed several experiments with different cut-off points. The statistical indices showed that different cut-off points (e.g., benign is considered negative (benign) and all others are positive (malignant)) were not clinically significant, which, due to the use of the shape

and margin features for mass classification, has predictable. In the revised manuscript, the "4.3" section has been rewritten almost. The part of mass classification has been changed completely according to the reviewer comments and a new table (Table 2) has been added for Confusion matrix of our CAD system. It is clearly shown that the benign cases are correctly classified with 90.2%, while the malignant cases with 81.3%. The false positive represents the 128 malignant cases that are incorrectly classified and negatively affect the specificity. We also have added a table (table 3) to summarizes the performance of the proposed CAD system in terms of sensitivity, specificity, precision, AUC, and overall accuracy. A new figure (Fig. 9) has been added to show the area under ROC curve of our proposed system.

The revised part was as following: "We used objective measures to evaluate the performance of our mass classification method. Confusion matrix is utilized to show how our method is able to distinguish between benign and malignant classes. The proposed method performance is evaluated with Curve of receiver operator characteristic (ROC) and Area under curve (AUC). ROC Curve definition is as follows [19],

$$\text{Sensitivity} = \frac{Number(TP)}{Number(TP)+Number(FN)} \quad (20)$$

$$\text{Specificity} = \frac{Number(TN)}{Number(FN)+Number(FP)} \quad (21)$$

In this experiment, a prediction of benign and probably benign is considered a negative prediction (benign), and malignant and probably malignant is positive prediction (malignant). Therefore, the results of diagnosing a mass are set as follows:

- True positive (TP): malignant masses are correctly diagnosed malignant.
- False positive (FP): benign masses are incorrectly diagnosed malignant.
- True negative (TN): benign masses are correctly diagnosed benign.
- False negative (FN): malignant masses are incorrectly diagnosed benign.

The overall classification accuracy of the proposed system is calculated as,

$$\text{Overall accuracy} = \frac{TP+TN}{TP+FN+TN+FP} \quad (22)$$

As referred in Table 2, the benign cases are correctly classified with 90.2%, while the malignant cases with 81.3%. Table 3 summarizes the performance of the proposed method. The results showed that proposed method distinguishes between benign and malignant lesions with an overall accuracy of 85.9%. Fig. 9 shows the ROC curve of our proposed method with AUC of 82.2%."

.Table 2. Confusion matrix of the proposed method

Total	**Malignant**	**Benign**	
829	81 9.8%	748 90.2%	**Benign**
764	621 81.3%	143 18.7%	**Malignant**
1593	702	891	**Total**

Table 3. Classification performance of the proposed .CAD system in CBIS-DDSM dataset

(%)	Index
81.3%	Sensitivity
90.2%	Specificity
88.5%	Precision
82.2%	AUC
85.9%	Overall Accuracy

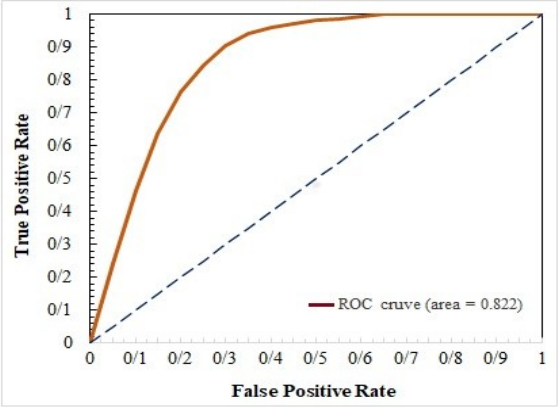


Figure 9. ROC curves of the proposed CAD system.

17) In the discussion, the authors have explained regarding the indices used in the paper. it is better they add some discussions to the paper and compare the findings with similar studies and with other image processing tools introduced earlier for this purpose.

Response 17: Thank you very much for your suggestion. A new discussion has been added in the revised manuscript and now it reads as follows:

"In recent years, many studies have been conducted on the development of CAD systems for breast cancer detection. Most of these studies are thoroughly dependent on segmentation and hand-crafted feature, which are not optimal in general. In some studies, wavelet coefficients and features extracted from wavelet coefficients have been used to detect breast abnormalities in mammograms [22], [23]. However, in these methods it is necessary to define hand-crafted features to detect and classify the mass regions in mammogram. As a result, the performance of these methods is limited, and in addition to high computational cost [24], also have resulted in a rather significant number of false positives [25]. More recently, deep learning-based approaches especially Convolutional Neural Networks (CNNs) have been introduced to overcome these limitations, but they need large annotated datasets for training which are expensive and time-consuming to acquire. Moreover, training a CNN requires a massive amount of data, high computing power and large memory resources.

In this study, Density of wavelet coefficients was employed to develop automatic mass detection in mammograms, without the need for feature extraction. Unlike traditional methods, our proposed method is derived from all structures of the mammogram image and, in addition to the simplicity of implementation, does not depend on complex mathematical and Fourier transforms."

18) Is it possible if the masses contain microcalcifications, the indices would be better or worse? It would be interesting if the authors discuss. [I know the scope of the study is only on the mass detection; not on microcalcification. However, microcalcification is an important diagnostic part of mammogram and combining its data would increase the diagnostic indices of method in discriminating malignant masses].

Response 18: Thank you for your constructive comments. We agree with your opinion regarding the diagnosis of microcalcifications can increase the diagnostic indices of method in discriminating malignant masses. However, in this study, suspicious region localization (SRL) focuses on finding suspicious mass regions. Hence, the small regions (or holes) found in the SRL process are removed or filled (see Figure 3). Since microcalcifications diameter is mostly less than 1mm and appear as bright little spots on mammograms, these regions (suspicious microcalcification regions) are usually removed (or filled) in proposed method. We have added discussion of these to the revised manuscript, and now it reads as follows: "A breast mass is the most common type of abnormality in mammogram. Occasionally, the masses contain microcalcifications [24] that diagnosis of these abnormalities can increase the diagnostic sensitivity for malignancy masses. However, since in this study, suspicious region localization (SRL) stage focuses on finding suspicious mass regions, the small regions (or holes) found in the SRL process are removed (see Figure 3). Since microcalcifications diameter is mostly less than 1mm, these regions (suspicious microcalcification regions) are usually removed in the proposed method."

Reviewer 2:

We appreciate that the reviewer’s comments. The followings are our point-by-point responses:

1) The English writing needs minor revision

Response 1: Thanks for these comments. We have followed this suggestion and the text was revised by an expert in English.

2 Revision 1

3 comments · 8 Oct 2020 – 28 Oct 2020

Referee 1 Reviewer

MAJOR REVISION 8 Oct 2020

The authors have corrected many parts, but there are some other corrections and clarifications that should be made:

It seems the paper contains two major parts: the first is mass detection [benign or malignant; the nature of the mass is not important], which is implemented using Discrete Wavelet Transform (DWT) [part 3.2. Suspicious region localization]. The second part is determining the nature of a mass [malignant vs. benign], which is performed using shape analysis and the compactness index [C] [part 3.3. Mass Classification]. The goal of the first part is the detection of a mass in mammography [locating a mass as accurately as possible and ensuring that its location corresponds to the actual mass in the breast], and the goal of the second part is determining whether a mass is malignant. The analysis of mass detection is pixel-based [for each image, we calculate the diagnostic indices], and the analysis of malignancy detection is patient-based. If I am right, please clarify this point in the paper, as I think it is confusing. For example, the gold standard for the first part is the mass borders drawn and annotated by the radiologists, and the gold standard for the second part is histopathological assessment, which is not mentioned in the paper and should be mentioned as part of the information in the CBIS-DDSM dataset database. We can calculate sensitivity in both of these parts, and each sensitivity has its own implication and meaning [pixel-based analyses for mass detection yield a result for each image, and this is why I requested that the authors calculate the mean and SD for their indices, and they responded: "The results showed that the sensitivity, FPPI, Precision and Recall of the SRL is 100%, 6/4 ± 4/5, 0.71 ± 0.02 and 0.56 ± 0.01, respectively." Patient-based analysis yields the sensitivity of shape analysis in determining whether a mass is malignant]. As I mentioned, the second goal is patient-based analysis, and the authors have mentioned the results in tables 2 and 3. These two approaches should be emphasized and clarified in the paper. [For example, two sensitivity formulas could be confusing. It would be better for the authors to mention that the first is pixel-based—formulas 16-19—and the second is patient-based—formulas 20, 21.] In addition, some other corrections are required:
I-The authors should arrange the tables sequentially. Table 1 is mentioned in the discussion.
II-Table 3, which is patient-based, needs two other indices, PPV and NPV, which could be calculated easily from table 2. In addition, the authors can calculate PLR and NLR based on this table.
III-One important point in the paper is that the authors add the following in the objective section: "in this paper, we propose a method based on the density of wavelet coefficients and morphological features classifier for automatic mass detection in digital mammograms that aims to help the radiologist in interpreting mammograms." That covers only the first goal and does not address the second part, which is the diagnosis of malignancy in a mass. In addition, the title could be adjusted to match both parts of the paper.
IV-There is a very important point in the paper that is not clarified at all or is confusing: is the compactness analysis [for calculating the C index and determining whether the mass is malignant] performed on the results of DWT analysis or on the ROI regions annotated by the radiologists in the database? This is important because, when a DWT output is used for shape analysis, it is probable that the FP pixels and FN pixels change the shape analysis result, and this can completely change the malignancy detection output compared with the results obtained when the radiologists' annotations are the input image for the "C" calculation. The authors mentioned, "Based on the above motivations, in this study, we use the mass shape characteristics in order to achieve increased accuracy of measurement results." But they did not mention which shapes are considered: hand-drawn ROI annotations or the output of wavelet analysis?
V-Considering the above-mentioned point, they can use the ROIs determined by DWT and the ROIs annotated

by the radiologists for the C calculation; then they can present experiment 2 under both conditions, compare the results of malignancy detection under these two conditions, and discuss them. This could be an interesting part if we want to combine two image analysis methods and see whether the accuracy of malignancy detection is lowered when we use DWT outputs.
V-Some parts added in red in the results belong to the methods section, not the results, and I request that the authors transfer all parts that are [in fact] study methods [not results] to the methods section. [Perhaps the formulas and their explanations.]

Some other questions could be mentioned as follows:
1-The authors mentioned, "The CBIS-DDSM database reviewed and annotated by trained mammographer". It would be good if the authors explained whether this annotation means that the radiologists drew the lesion borders on the images and thus identified the lesions on the images.
2-In response to question 6, the authors mentioned the sensitivity and recall formulas, which are identical [formulas 16 and 19]. Why are these two formulas identical? [TP/TP+FN]
3-In the figure 7 legend, both sensitivities in parts b and c are 1. If we consider pixel-based accuracy calculation, then the sensitivity of part b would be 1, and the sensitivity of part c would be lower than 1, as we have some missed pixels [FN pixels]. Please describe.
4-The authors mentioned: "The proposed method performance is evaluated with Curve of receiver operator characteristic (ROC) and Area under curve (AUC). ROC Curve definition is as follows [19]", while in the following part they mentioned the formulas for calculating Sen and spec in a 2 by 2 cross table, which are not formulas for ROC. They should correct this part. In addition, I requested that they change the cutoff point [in a patient-based analysis for malignancy detection] so that they can implement ROC, but their explanations in the paper do not address this point, and the part related to ROC is confusing. I request that the authors complete and clarify this part. In addition, if we have only three cutoff points in the patient-based analysis, then the ROC curve would not be so smooth, and it would be a continuous broken line with three ties. This is confusing.
5-The sensitivity of 100% [without SD] in pixel-based analysis [for mass detection] means that all ROIs of the DWT have included the masses annotated by the radiologists on the images, and the FN pixels are 0 for all 1593 images. Am I right? This is somewhat strange. In addition, in this case, why is the recall not 1 [recall relates to FN]? Please explain.
6-If the C thresholds of 0.6, 0.7 and 0.8 are changed, the UAC and diagnostic indices of malignancy detection [patient-based analysis] will differ. It would be better if the authors added a brief discussion in this regard. In addition, they mentioned that they chose the numbers based on the radiologists' opinions, which is vague. Did the radiologists have any previous experience in this regard? [This is a new image analysis approach, and I wonder how the radiologists could determine these thresholds.]

Referee 2 Reviewer

MAJOR REVISION 20 Oct 2020

Dear associate editor,
The article needs revision.
1) Please define the abbreviations at their first occurrence in the article.
2) Please write the references according to the IJR format.
For example, [2]-[4] should be written as (2-4), etc.
3) Please compare the results with other similar works in the literature and discuss accordingly in the discussion section.
4) Please write the limitations of the study before the conclusion.
5) There is no need to mention the abstract of the work in the conclusion section. Please delete the first

paragraph and only write the final conclusion.
Regards

Editor-in-Chief

MAJOR REVISION 28 Oct 2020

Dear authors,
Your article needs Major Revision.
Regards

3

Revision 2

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

November 17, 2020
Associate Editor
The Iranian Journal of Radiology

Manuscript ID: 106717

Manuscript Title: An efficient method for automated breast mass segmentation and classification in digital mammograms.

Dear Editor:

Thanks for your comments concerning our manuscript. Those comments have all been very helpful for revising and improving our paper as well as providing important guidance for our research. Based on the suggestions and comments of the reviewers, the paper has been made major revision carefully and the title of manuscript has been revised to be “An efficient method for automated breast mass segmentation and classification in digital mammograms”. We hope it will be able to meet with your approval. The major revised parts are highlighted in red color for your convenience of re-reviewing. The detailed corrections and the responses to your comments are listed below point by point.

Best wishes!

Yours sincerely

Alireza Nikravanshalmani, Ph.D.
Professor
Department of Computer.
Karaj Branch, Islamic Azad University, Karaj, Iran.

We appreciate that the reviewer's comments. The followings are our point-by-point responses.

Response to Reviewer 2:

1) Please define the abbreviations for the first time in the article.
   Response 1: We have made this change.
2) Please write the references according to IJR format. For example [2]-[4] should be written as: (2-4), ..., etc.
   Response 2: We have made this change.
3) Please compare your results with other similar works in the literature and discuss accordingly in the discussion section.
   Response 3: We appreciate your suggestions. According to this, we have made major revision and rewrote Background and discussion section.
4) Please write the limitations of your study before conclusion.
   Response 4: Thank you for your constructive comments. We have added limitations of current study in the revised manuscript and now it reads as follows:
   There are several limitations in this study. Firstly, this study is a small study with a relatively small number of mass lesions and the results must therefore be interpreted with caution. A larger sample of mammogram images with lesion annotated is required, to adequately assess the proposed method. Another limitation in the present study is that Compactness characteristic is used to classify masses. However, the utility of compactness is limited due to its sensitivity to noise. To minimize this influence, several combinations of image enhancement algorithms were used in order to enhance the performance of mammogram.
5) There is no need to mention the abstract of the work in the conclusion section. Please delete the first paragraph and only write the final conclusion.
   Response 5: Thanks for this comment. In the revised manuscript, the Conclusion section has been rewritten almost, and now it reads as follows:
   In this paper we have introduced a novel approach to mammogram analysis that is able to find suspicious mass regions and classifying these regions into four categories. It can be concluded that the suspicious regions were successfully segmented and the classifier succeeded to discriminate between masses. The proposed method in this study can be used in breast cancer screening and training of medical staff. In addition, the proposed mass segmentation method can be applied to various studies, and potentially improve the performance of CAD systems. Although the proposed method could segment the suspicious regions and their boundary more accurately, the false positives of the proposed method are considerable. Thus, it is necessary to optimize the size of the ROI and the number of these regions. We investigate it as a future work.

Response to Reviewer 3:

We appreciate that the reviewer's comments. The followings are our point-by-point responses.

1) The authors should arrange the tables sequentially. Table 1 is mentioned in discussion.
Response 1: We apologize for the mistake in the table sequentially, as you can see in manuscript, the correct sequence of tables was presented.

2) The table 3 that is patient based needs two other indices that are PPV and NPV that could be calculated easily from the table 2. In addition, they can calculate PLR and NLR based on this table.
Response 2: Thank you for your suggestion. As suggested by the reviewer, we calculated the two indicators PPV and NPV and added them to Table 2.

4) One important point in the paper is that the authors add in objective part: "in this paper, we propose a method based on the density of wavelet coefficients and morphological features classifier for automatic mass detection in digital mammograms that aims to help the radiologist in interpreting mammograms." That covers only the first goal and does not denote the second part that is the malignancy diagnosis in a mass. In addition, the title could be matched for both parts of the paper.

Response 4: Thank you very much for your suggestions. We rewrote the objective part as follows: "In this work, for the purpose of accurate and efficient mass detection in digital mammograms, we propose a DWC-based method for localizing the suspicious lesions on mammograms. The masses that appear in mammograms are further analyzed and classified into four groups based on morphological features." The manuscript title, also rewrote as follows: "An efficient method for automated breast mass segmentation and classification in digital mammograms".

5) There is a very important point in the paper that is not clarified at all or is confusing: it is unclear if the compactness analysis [for calculating C index and determining if the mass is malignant] is performed on the results of DWT analysis or on the ROI regions annotated by the radiologists in the data base? This is important as when a DWT output is used for shape analysis, it is probable that the FP pixels and FN pixels change the shape analysis result and this can completely change the malignancy detection output when we want to compare results with the situation that the radiologist annotations are the feed image of "C" calculation. The authors mentioned "Based on the above motivations, in this study, we use the mass shape characteristics in order to achieve increased accuracy of measurement results." But they did not mentioned which shape are considered, annotated handy ROI drawings or the output of wavelet analysis?
Response 5: We are sorry for the confusion. In order to discriminate between different mass shapes, we used CI index and certain thresholds. The threshold measures were determined after conducting comprehensive comparisons between different mass shapes from the CBIS-DDSM dataset. These certain thresholds are used to classify the suspicious regions obtained by DWT analysis. Therefore, once a suspicious region is extracted, the subsequent task is to classify it as benign, probably benign, malignant and probably malignant. This is done by using morphological feature analysis and certain thresholds for each suspicious region. In

the revised manuscript, the "3.3. Mass Classification" section has been rewritten almost. The revised part of “3.3. Mass Classification” is as following.

3.3. Mass Classification

The classification stage is performed using shape features classifier. The aim of this step is to distinguish between benign and malignant masses. Although, several morphological features have been employed for the classification of benign and malignant masses in the literature (12–14,18), it is desirable to use an optimum number of features that will still allow the CAD system to achieve the desired target performance. The border of benign tumors is relatively smooth, while the malignant have much more irregular borders. According to this hypothesis, the Compactness of the tumor is calculated using the following equation (19):

$$CI = 1- \frac{4 \times \pi \times A}{P^2} \qquad (15)$$

Where CI is compactness index, P is total number of pixels in the edge or margin of the tumor and A its number of pixels or the area of the tumor. If the CI = 0 then the mass is a circle, otherwise it of other shapes. The compactness is used to measure shape characteristics. A high CI value indicates a large perimeter enclosing a small area. As the shape becomes more complex and rougher, the value of the CI index increases. Moreover, the CI is independent of linear transformations, rotation, starting point, and size of the given contour and measures the degree of roughness of a region. Therefore, typical malignant masses are expected to have higher values of CI compared to typical benign tumors. In order to discriminate between different mass shapes, we used certain thresholds to mass classifying. The thresholds are determined by comprehensive comparisons between different mass shapes that annotated by the radiologists in the CBIS-DDSM dataset as follows: First, we calculate parameter CI for each mass in CBIS-DDSM dataset, if the mass is benign then CI < 60% and for malignant masses is CI ≥ 80%. If CI $\in$ [60%, 70%) the mass is probably benign and otherwise the mass is considered probably malignant (see Fig. 5). The suspicious regions detected from the previous stage will be analyzed based on these thresholds and are classified into four categories.

6) Considering the above-mentioned point, they can use the ROIs determined by DWT and ROIs annotated by the radiologists for the C calculation; then they can represent the experiment 2 in both two situations and compare the results of malignancy detection in these two situations and discuss. This could be an interesting part if we want to combine two image analysis methods together and see if the accuracy of malignancy detection is lowered when we use DWT outputs or not?

Response 6: Thank you for your comments. Calculating and rigorous evaluation of CI index and threshold values require data that is labeled with Ground Truth. Therefore, as mentioned in Section 5, the threshold values are calculated based on Ground Truth from CBIS-DDSM database which is annotated by experts (please refer to Mass Classification and Database section in manuscript). However, to calculate threshold values based on extracted suspicious regions, we were unable to find suitable annotations for these regions. We used the obtained threshold from CBIS-DDSM dataset values to analysis and to predict the type of lesions of the extracted suspicious regions. Please refer to the response 5 above.

7) Some parts added in red part in results are belong to method section not results and I request the authors to transfer all parts that are [in fact] study method [not result] to the method section. [may be the formulas and their explanations]

Response 7: Thanks for these comments. In the revised manuscript, the Result section has been rewritten almost. Some sentences have been removed to the Discussion and Method sections. In addition, a new section "3.5 Evaluation protocol" has been added to describe the evaluation criteria.

The revised parts of “Results” and ".5 Evaluation protocol" are as following.

3.5 Evaluation protocol

Several indicators were used to evaluate the performance of proposed method in this study. The ROI denotes the suspicious region found by the algorithm and region GT stands for the ground truth. For simplicity, the same notation is used for a region and its number of pixels.

The sensitivity, average false positives per image (FPPI), precision, and Jaccard index (IoU) were used to evaluate the performance of the proposed segmentation method. Sensitivity is used to measure the proportion of actual positives which are correctly identified. The sensitivity is also known as true-positive rate or probability of detection (see Fig. 6), and is defined as follows:

$$= \frac{Number\ of\ TP\ regions}{Number\ of\ TP\ regions + Number\ of\ FN\ regions} = \text{Sensitivity}$$

$$\frac{Number\left(\left(> \bigcap ROI\right) \neq \varnothing\right)}{Number\left(>\right)} \tag{16}$$

In a CAD system, we want to allow high sensitivity in the finding suspicious regions, even at high FP costs. FP rates can be reduced through classification. false positive per image (FFPI) is defined as follows:

$$\frac{number\ of\ false\ positives}{number\ of\ images} = \text{FPPI} \tag{17}$$

The Precision and Jaccard index (IoU), defined as:

$$\frac{ROI\ ¿}{ROI}\ \text{Precision} = \frac{TP}{TP + FP}\ \text{or} \tag{18}$$

$$\frac{ROI\ ¿}{ROI\ ¿} = \frac{Areao\ f\ overlap}{Area\ of \cup ¿\ ¿} = \text{Jaccard Index} \tag{19}$$

The Jaccard index is used to measure how much suspicious region boundary overlaps with the GT boundary.

To evaluate classification method performance, we utilized receiver operator characteristic (ROC) curve with its area under the curve (AUC) and overall accuracy. An

ROC curve illustrates the diagnostic ability of a classifier system and is defined based on sensitivity and specificity as follows (14):

$$\text{Sensitivity} = \frac{Number(TP)}{Number(TP)+Number(FN)} \quad (20)$$

$$\text{Specificity} = \frac{Number(TN)}{Number(FN)+Number(FP)} \quad (21)$$

The overall classification accuracy is calculated as:

$$\text{Overall accuracy} = \frac{TP+TN}{TP+FN+TN+FP} \quad (22)$$

The negative predictive value and the positive predictive value are also defined as follows:

$$\text{NPV} = \frac{True\ negative}{True\ negative+False\ negative} \quad (23)$$

$$\text{PPV} = \frac{True\ positive}{True\ positive+False\ positive} \quad (24)$$

4. Results

4.1. Experiment #1: Suspicious region localization (SRL) is illustrated in Fig. 4, 5. True Positives (TP), False Positives (FP), and False Negatives (FN) are defined in Fig. 6. The information provided by radiologists and their opinions in CBIS-DDSM dataset, along with the similarity measure between suspicious regions with ground truth (see Fig. 6), was the main basis of validation in this work.

The obtained results show that the sensitivity, FPPI, Precision and IOU of the segmentation method was 100%, 6/4 ± 4/5, 0.71 ± 7.6% and 0.68 ± 0.06%, respectively. The comparison results in the same dataset are shown in Table 1. In comparison to other previous methods. Our method had a superior performance based on the sensitivity results.

4.2. Experiment #2: For each suspicious mass region, we computed the CI index as described in Sec 3.3. Confusion matrix was utilized to show how our method was able to distinguish between benign and malignant tumors. The overall performance of the diagnostic system was also evaluated with ROC curve analysis. In this study, ROC curve analysis was performed to classify benign and malignant in the extracted suspicious regions. Therefore, a prediction of benign and probably benign is considered a negative prediction (benign), and malignant and probably malignant is positive prediction (malignant). The results of diagnosing a mass are set as follows:

- True positive (TP): malignant masses are correctly diagnosed malignant.
- False positive (FP): benign masses are incorrectly diagnosed malignant.
- True negative (TN): benign masses are correctly diagnosed benign.
- False negative (FN): malignant masses are incorrectly diagnosed benign.

In our experiment, the benign cases are correctly classified with 90.2%, while the malignant cases with 81.3%. We list the experimental result in Table 2 and Fig. 9 shows the

ROC analysis. The results showed that proposed method overall accuracy is 85.9%, the sensitivity is 81.3%, the specificity is 90.2%, the positive predictive value is 88.5%, and the negative predictive value is 83.9%. These indices appear in Table 3. According to the Table 3 and Fig. 9, we believe that the classification ability for breast tumors with shape information is satisfactory.

8) The authors mentioned “The CBIS-DDSM database reviewed and annotated by trained mammographer” it would be good if the authors explain this annotation means they drew the lesion borders on the images and so determined the lesions on the images.

Response 8: We appreciate your suggestions. We corrected it in the revised manuscript. The revised part of "Dataset" is as following.

3.4. Dataset

In this study, we have employed a new updated and standardized version of the DDSM database (20), i.e., Curated Breast Imaging Subset of DDSM (CBIS-DDSM) (21) which contains images with standard DICOM format. The CBIS-DDSM database reviewed and annotated by trained mammographers. We employ all mass cases of this database, which contain 1593 mammograms (829 benign and 764 malignant) of 892 cases and a total of 1698 masses with Ground Truth (GT) masks (i.e., annotations). In addition, The CBIS-DDSM database contains pixelwise annotations for the ROIs, e.g., mass shape, BI-RADS assessment (rating (0, 2-5)), as well as lesion’s pathology (2 annotations: benign or malignant). The region of interest segmentation, bounding boxes and pathologic ground truth about the diagnosis are also included in this dataset. In this study, each view was used as a separate image and all images were resized to 1024×1024 pixels using bicubic interpolation. For our experiments, the ROIs are extracted by using the binary segmentation masks provided in the CBIS-DDSM dataset.

9) In response to question 6, the authors mentioned sensitivity and recall formula that these formulas are identical [formulas 16 and 19]. Why these two formulas are identical? [TP/TP+FN]

Response 9: We are sorry for the confusion. We agree with your opinion regarding the Recall formula is not clear. We had used the Recall index as statistical comparison parameter to evaluate the overlapping of the Ground Truth (GT) and the ROI. This index is based on pixels and show the true positive percentage. However, in order to avoid confusing the readers, we deleted this indicator. To evaluate the performance of proposed segmentation method, the standard Jaccard similarity index (or UoI) is adopted. The index is calculated by comparing the suspicious region segmented by DWT with GT region outlined by the CBIS-DDSM dataset. The value of UoI lies between 0 and 1. It would be 1 if suspicious region and GT are completely overlapping, while when these two regions are entirely different, it would be 0. In the revised manuscript, the "4.2. Experiment #1" section has been rewritten almost. We have also added some discussions to the section of Discussion.

The revised “4.2. Experiment #1” and "5. Discussion" are as following.

4.1. Experiment #1: Suspicious region localization (SRL) is illustrated in Fig. 4, 5. True Positives (TP), False Positives (FP), and False Negatives (FN) are defined in Fig. 6. The information provided by radiologists and their opinions in CBIS-DDSM dataset, along with

the similarity measure between suspicious regions with ground truth (see Fig. 6), was the main basis of validation in this work.

The obtained results show that the sensitivity, FPPI, Precision and IOU of the segmentation method was 100%, 6/4 ± 4/5, 0.71 ± 7.6% and 0.68 ± 0.06%, respectively. The comparison results in the same dataset are shown in Table 1. In comparison to other previous methods. Our method had a superior performance based on the sensitivity results.

5. Discussion

Although in the experimental results we have shown effectiveness of the proposed method and discussed the possible applications, there are some technical essentials need to be noted are as follows.

In this study, we proposed a new method to mass detection in digital mammograms. The method consists of two major stages, ie, finding suspicious regions and classify these regions by morphological features. In the first stage, Density of wavelet coefficients was proposed to develop automatic mass detection in mammograms, without the need for feature extraction. Unlike traditional methods, our proposed method is derived from all structures of the mammogram image and, in addition to the simplicity of implementation, does not depend on complex mathematical and Fourier transforms.

One of the challenges of SRL is the usually high number of false positive (FP) rates. As shown in Table 1, the proposed segmentation method has better sensitivity than other algorithms in detecting suspicious mass regions. High sensitivity and relatively low FPPI of the proposed method indicate the good performance of our method. Therefore, the proposed method can be used to suspicious region localization in mass screening programs or in breast CAD systems.

The purpose of the segmentation stage is to extract all suspicious regions from mammograms, so if a mass is not present in any of the ROIs, *i.e.*, $¿ \bigcap ROI = \varnothing$ (Fig. 6 (a)), the mass is lost and therefore it is considered as “False Negative” region. In other hand, only the regions detected by segmentation method (these regions are considered as “True Positive” regions) are further analyzed in next stage. As illustrated in Figure 6, in both sections (b, c), the mass is covered by the ROI and as a result there is a true positive region, i.e., $GT \bigcap ROI \neq \emptyset$.

We have used the Precision and Jaccard index as statistical comparison parameters to evaluate the overlapping of the GT and the ROI. The Precision Index informs about the percentage of ROI pixels which overlapping with the GT pixels. The value of Precision index lies between 0 and 1. It would be 1 if one of ROI and GT are completely overlapping (see Fig. 6 (b)), while when these two regions are entirely different, it would be 0 (see Fig. 6 (a)). In addition, we used the Jaccard similarity index to measure the similarity between the GT and ROI regions.

A mass on a mammography image can be either malignant or benign, depending on its shape. The compactness characteristic has been identified as an important feature in distinguishing benign masses from malignant masses. We used the mass compactness index and the threshold values to distinguish between benign and malignant masses. As shown in

Fig. 5, the threshold measures are determined on the basis of comprehensive comparisons between different mass shapes and based on the trad-off between sensitivity and specificity from the CBIS-DDSM dataset. Based on this, there cut-off points were chosen, 60%, 70% and 80%. Masses were then stratified into four groups, benign, malignant, probably benign, and probably malignant (see Fig. 9). Once the suspicious regions are extracted from the mammograms, the CI index is calculated to these regions. Subsequently, based on the obtained thresholds, the masses are classified into four categories. The obtained results show that the proposed method exhibits an Overall Accuracy of 85.9% with AUC of 90.1%. The recognition results by proposed system were shown in Table 3. We believe that the outputs of this study and the proposed measurement can help radiologists in malignancy diagnosis of breast masses. In addition, the proposed suspicious region localization method may provide a valuable reference for abnormalities segmentation in other studies.

A breast mass is the most common type of abnormality in mammogram. Occasionally, the masses contain microcalcifications (22) that diagnosis of these abnormalities can increase the diagnostic sensitivity for malignancy masses. However, in this study, SRL stage focuses on finding suspicious mass regions, therefore, the small regions (or holes) found in the SRL process are suppressed (see Fig. 3). Since microcalcifications diameter is mostly less than 1mm, these regions (suspicious microcalcification regions) are usually suppressed in the SRL process.

There are several limitations in this study. Firstly, this study is a small study with a relatively small number of mass lesions and the results must therefore be interpreted with caution. A larger sample of mammogram images with lesion annotated is required, to adequately assess the proposed method. Another limitation in the present study is that Compactness characteristic is used to classify masses. However, the utility of compactness is limited due to its sensitivity to noise. To minimize this influence, several combinations of image enhancement algorithms were used in order to enhance the performance of mammogram.

10) In the figure 7 legend, both sensitivities in parts b and c are 1. If we think of pixel-based accuracy calculation, then the sensitivity of part b would be 1 and the sensitivity of part c would be lower than 1 as we have some missed pixels [FN pixels]. Please describe.
Response 10: Thank you for your comments. We apologize that the description of this part was not clear. In the sensitivity definition, the sensitivity is represented, not by pixels, but by regions, As shown in Equation 16.

$$\text{Sensitivity} = \frac{Number\ of\ TP\ regions}{Number\ of\ TP\ regions + Number\ of\ FN\ regions} \text{ or } \frac{Number\left(\left(> \bigcap ROI\right) \neq \varnothing\right)}{Number\left(>\right)} \quad (16)$$

The aim of this index is to show true positive rate. It would be 1 if suspicious region and GT are completely or partially overlapping (i.e., ¿$\bigcap ROI \neq \varnothing$), while when these two regions do not overlap (i.e., ¿$\bigcap ROI = \varnothing$), it would be 0.

However, in order to avoid confusing the readers, in the revised manuscript, the "4.2. Experiment #1" and "5. Discussion" section has been rewritten almost. Please refer to the response 9 above.

11) The authors mentioned: “The proposed method performance is evaluated with Curve of receiver operator characteristic (ROC) and Area under curve (AUC). ROC Curve definition is as follows [19]” while in the following part they mentioned the formula of calculating Sen and spec in a 2 by 2 cross table that is not the formula for ROC. They should correct this part. In addition, I requested them to change the cut of point [in a patient based analysis for malignancy detection] and so they can implement ROC while they explanations in the paper does not denote this point and the part related to ROC is confusing. I request the author to complete and clarify this part. in addition, if we have only three cutoff point in the patient based analysis, then THE ROC curve would not be so soft and it would be a continuous broken line with three ties. This is confusing.

Response 11: Thank you for your comments. We apologize that the description of this part was not clear. The ROC curve is defined based on sensitivity and specificity. In other word, the ROC curve presents a trade-off between sensitivity and specificity. In the revised manuscript, Experiment #2 has been rewritten. We also recalculated the cut-off points with optimal sensitivity and specificity, and repeated ROC Curve analyses using these cut-off points (Figure 9). Please refer to the response 7 above.

The revised part of "4.3. Experiment #2" is as following.

4.2. Experiment #2: For each suspicious mass region, we computed the CI index as described in Sec 3.3. Confusion matrix was utilized to show how our method was able to distinguish between benign and malignant tumors. The overall performance of the diagnostic system was also evaluated with ROC curve analysis. In this study, ROC curve analysis was performed to classify benign and malignant in the extracted suspicious regions. Therefore, a prediction of benign and probably benign is considered a negative prediction (benign), and malignant and probably malignant is positive prediction (malignant). The results of diagnosing a mass are set as follows:

- True positive (TP): malignant masses are correctly diagnosed malignant.
- False positive (FP): benign masses are incorrectly diagnosed malignant.
- True negative (TN): benign masses are correctly diagnosed benign.
- False negative (FN): malignant masses are incorrectly diagnosed benign.

In our experiment, the benign cases are correctly classified with 90.2%, while the malignant cases with 81.3%. We list the experimental result in Table 2 and Fig. 9 shows the ROC analysis. The results showed that proposed method overall accuracy is 85.9%, the sensitivity is 81.3%, the specificity is 90.2%, the positive predictive value is 88.5%, and the negative predictive value is 83.9%. These indices appear in Table 3. According to the Table 3 and Fig. 9, we believe that the classification ability for breast tumors with shape information is satisfactory.

Table 2. Confusion matrix of the proposed method.

Total	Malignant	Benign	
829	81 9.8%	748 90.2%	**Benign**
764	621 81.3%	143 18.7%	**Malignant**
1593	702	891	**Total**

Table 3. Classification performance for the proposed CAD system in CBIS-DDSM dataset.

(%)	Index
81.3	Sensitivity
90.2	Specificity
90.1	AUC
83.9	NPV
88.5	PPV
85.9	Overall Accuracy

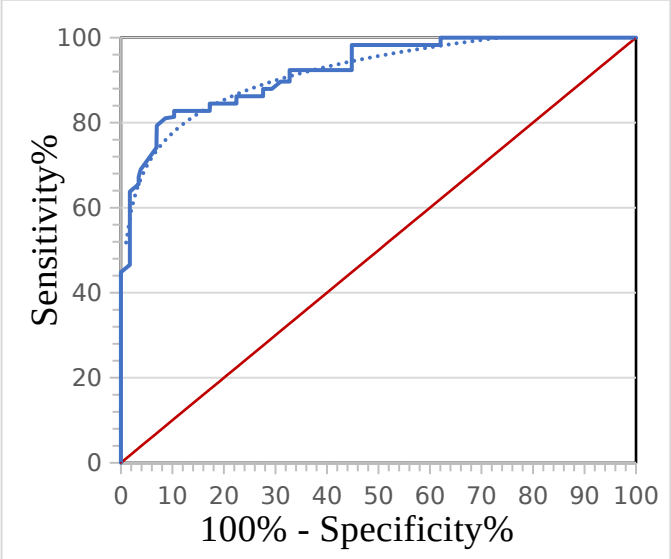


Figure 9. ROC curves of the proposed CAD system.

12) The sensitivity of 100% [without SD] in pixel based analysis [for mass detection] means all ROIs of the DWT have included the masses annotated by the radiologists on the images and the FN pixels are 0 for all 1593 images. Am I right? This is somehow strange. In addition, In this case, why the recall is not 1 [recall relates to FN]? Please explain.

Response 12: We are sorry for the confusion. As explained in the manuscript and mentioned in the response 10,11, the sensitivity or true positive rate is defined based on the region as follows:

$$\text{sensitivity} = \frac{Number\ of\ TP\ regions}{Number(TP) + number(FN)} \text{ or } \frac{Number((>\bigcap ROI) \neq \varnothing)}{Number(>)} \quad (16)$$

The sensitivity of 100% means that all masses in mammogram image are correctly detected as suspicious region. Please refer to the response to “Response 10 and 11” above.

13) If the C thresholds of 0.6, 0.7 and 0.8 are changed, the UAC and diagnostic indices of malignancy detection [patient based analysis] will differ. It is better if the authors add a brief discussion in this regard. In addition, they mentioned they chose the numbers based on the radiologist opinion that is vague. Did the radiologists have any previous experience in this regard? [This is a new image analysis approach and I wonder how the radiologists could determine these thresholds?].

Response 13: Thank you for your comments. According to this, we have added some discussions to the section of Discussion. In addition, As mentioned in Response 5 and 6, the CI index is calculated by CAD system for the extracted suspicious regions, not by the radiologists. The threshold measures are determined on the basis of comprehensive comparisons between different mass shapes and based on the trad-off between sensitivity and specificity from the CBIS-DDSM dataset (see Fig. 5). We believe that the proposed measurement may provide a valuable reference of malignancy diagnosis of breast masses to radiologists. Please refer to the response 5, 6 and 9 above.

3 Revision 2

3 comments · 24 Nov 2020 – 8 May 2021

Referee 2 Reviewer

ACCEPT 24 Nov 2020

Dear associate editor,
The authors incorporated my comments into the article.
Best regards

Associate Editor

MINOR REVISION 7 May 2021

1-It is better to report the AUCs as decimal numbers [between 0-1] rather than percentages.
2-Please correct: in the passage "…. and increasing survival rates by early ……" "increasing" should be "increase".
3-Please add the full terms for abbreviations in the abstract and main paper separately. For example: CAD in the main paper in line 36 should be "Computer-aided diagnosis (CAD)". DWC in the objective should also be given in full, followed by the abbreviation.
4-Please correct: "new deep deep learning framework" to "new deep learning framework".
5-This part is unclear: "0.812 TPR@2.53 FPI on CBIS-DDSM database". Please clarify.
6-In line 74, "The first phase is" should be "in the first phase".
7-Please arrange the figures in the paper in numerical order. Figure 4 should not be mentioned before figure 3.
8-The formulas are cut off at the top and are not presented correctly. For example, the first formula in line 90.
9-The passage "and A its number" seems to be "and A is number".
10-The passage "otherwise it of other shapes" seems grammatically incorrect. It could be: "otherwise it has other shapes".
11-Please use a consistent style of writing. The authors have written: 6/4 ± 4/5, 0.71 ± 7.6%. It seems that .71 ± 7.6% is a better style.
12-This sentence is unclear: "In our experiment, the benign cases are correctly classified with 90.2%, while the malignant cases with 81.3%". With 90.2 what? With 81.3% what?
13-Please correct: "new method to mass detection" to "new method for mass detection".
14-Please correct: "there cut-off points were chosen" should be "three cut-off points were chosen".
15-Please delete "6. Conclusion" and replace it with "in conclusion".
16-Please add the full terms for abbreviations in tables as footnotes to the tables.
17-Please add the full terms for abbreviations in figure legends in parentheses.
18-In figure 6, part C, is the sensitivity equal to 1 here? In part B, with complete coverage of the ground truth, sensitivity is 1, but with partial coverage, sensitivity is less than 1.

Editor-in-Chief

MINOR REVISION 8 May 2021

Dear authors,
Your article needs Minor Revision.
Regards

4

Revision 3

AUTHORS' RESPONSE TO REVIEWERS

Authors' reply to reviewers

The next pages contain the authors' point-by-point reply to the comments of this revision round, as submitted with the revised manuscript. It is published as the authors wrote it.

May. 9, 22

Dear Professor Kavous Firouznia

First, I want to thank you and the reviewers for the extremely helpful comments provided for our paper. In the paper, we've addressed all comments from the reviewer as follows. We hope it will be able to meet with your approval.

Response to Reviewer 2:

1)It is better the AUCs report based on a decimal number [between 0-1] not based on percent.
Authors' Response 1: We have made this change.

2)Please correct: in part "…. and increasing survival rates by early ……" "increasing" should be "increase".
Authors' Response 2: We have made this change.

3) Please add full term of abbreviations in abstract and main paper separately. For example: CAD in main paper in line 36 should be "Computer-aided diagnosis (CAD)". Or DWC in objective should be mentioned in full and abbreviation.
Authors' Response 3: We have made this change.

4)Please correct: "new deep deep learning framework" to "new deep learning framework".
Authors' Response 4: Thanks for this comment. We have made this change.

5)This part is unclear: "0.812 TPR@2.53 FPI on CBIS-DDSM database" please clarify.
Authors' Response 5: Thank you, we have revised this sentence for clarity: "They achieved the best performance of 0.812 true positive rate (TPR) at 2.53 false positive per image (FPI) on CBIS-DDSM database."

6)In line 74, "The first phase is" should be "in the first phase"
Authors' Response 6: We have made this change.

7)Please arrange the figures in paper based on sequential numbers. The figure 4 should not be mentioned before figure 3.
Authors' Response 7: We have made this change.

8)The formulas are cut at the upper line and are not mentioned in correct form. For example first formula in line 90.
Authors' Response 8: Thank you for your comments. The formulas were written correctly in the original manuscript on the relevant page. Unfortunately, the formulas are corrupted during the building of the PDF file for submission. A figure has been added in the revised manuscript to show all formulas in the correct form.

9)The part: "and A its number" seems to be "and A is number".
Authors' Response 9: We have made this change.

10)Part: "otherwise it of other shapes" seems wrong based on writing. It could be: "otherwise it has other shapes".
Authors' Response 10: We have made this change.

11)Please follow unified method of writing. The authors have written: 6/4 ± 4/5, 0.71 ± 7.6%. it seems .71 ± 7.6% has better style

Authors' Response 11: We have made this change.

12)This sentence is unclear: “In our experiment, the benign cases are correctly classified with 90.2%, while the malignant cases with 81.3%” with 90.2 what? with 81.3% what?

Authors' Response 12: Thank you for your comments. We apologize that the description of this part was not clear. In order to avoid confusing the readers, in the revised manuscript, this part has been rewritten and now it reads as follows: "Confusion matrix of the proposed CAD system is presented in Table 2 and Figure 9 shows the ROC analysis. The results showed that proposed method overall accuracy is 85.9%, the sensitivity is 81.3%, the specificity is 90.2%, the positive predictive value is 88.5%, and the negative predictive value is 83.9%. These indices appear in Table 3. According to the Table 3 and Figure 9, we believe that the classification ability for breast tumors with shape information is satisfactory."

13)Please correct: “new method to mass detection” to “new method for mass detection”

Authors' Response 13: We have made this change.

14)Please correct: “there cut-off points were chosen” should be “three cut-off points were chosen”

Authors' Response 14: We have made this change.

15)Please delete “6. Conclusion” and replace it by “in conclusion”

Authors' Response 15: We have made this change.

16)Please add full term of abbreviations in tables ass footnote of the table.

Authors' Response 16: We have made this change.

17)Please add full term of abbreviations in figure legends in parenthesis.

Authors' Response 17: We have made this change.

18)At the figure 6 part C, is the sensitivity equal to 1 here? In part B with complete coverage of ground truth, sensitivity is 1 but in partial coverage, sensitivity is less than 1.

Authors' Response 18: In the first stage (i.e., suspicious region localization), a suspicious region (SR) is considered as a true positive if it overlaps with GT region (i.e., $GT \bigcap SR \neq \emptyset$), otherwise, it is considered as a false positive, as shown in figure 6. In order to avoid confusing the readers, in the revised manuscript, figure 6 has been redesigned and now as follows:

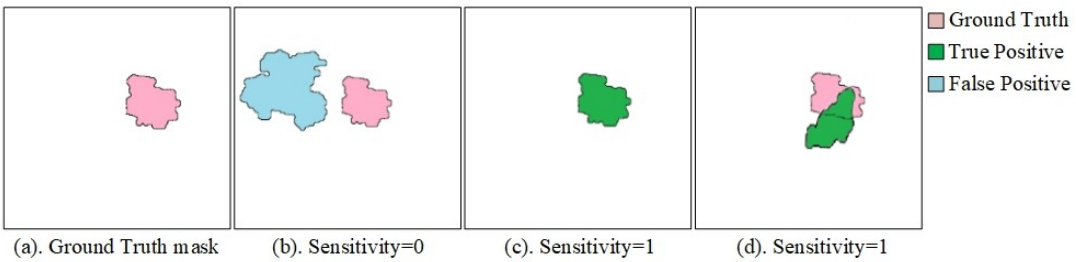


Figure 6. Examples of sensitivity calculations for suspicious region (SR) extraction. (a) Ground truth, (b) no coverage of Ground Truth (i.e., $\bigcap SR = \varnothing$), Sensitivity=0, (c) One of SRs overlaps completely the Ground Truth (complete coverage of Ground Truth, $\bigcap SR \neq \varnothing$), Sensitivity=1, (d) One of SRs cover portion of the Ground Truth (partial coverage of Ground Truth, $\bigcap SR \neq \varnothing$), Sensitivity=1.

4 Revision 3

1 comment · 15 May 2021

Editor-in-Chief

ACCEPT 15 May 2021

Dear authors,
Your article is acceptable for publication in IJR.
Regards