Improvement of Sample Selection: A Cascade-Based Approach for Lesion Automatic Detection

Shofwatul 'Uyun[1,a], M. Didik R Wahyudi[1,c]
[1]Department of Informatics, Faculty of Science and Technology, State Islamic University Sunan Kalijaga, Yogyakarta, Indonesia

Lina Choridah[2,b]
[2]Department of Radiology, Faculty of Medicine, Gadjah Mada University, Yogyakarta, Indonesia

***Abstract*—Computer-Aided Detection (CADe) system has a significant role as a preventative effort in the early detection of breast cancer. There are some phases in developing the pattern recognition on the CADe system, including the availability of a large number of data, feature extraction, selection and use of features, and the selection of the appropriate classification method. Haar cascade classifier has been successfully developed to detect the faces in the multimedia image automatically and quickly. The success of the face detection system must not be separated from the availability of the training data in the large numbers. However, it is not easy to implement on a medical image because of some reasons, including its low quality, the very little gray-value differences, and the limited number of the patches for the examples of the positive data. Therefore, this research proposes an algorithm to overcome the limitation of the number of patches on the region of interest to detect whether the lesion exists or not on the mammogram images based on the Haar cascade classifier. This research uses the mammogram and ultrasonography images from the breast imaging of 60 probands and patients in the Clinic of Oncology, Yogyakarta. The testing of the CADe system is done by comparing the reading result of that system with the mammography reading result validated with the reading of the ultrasonography image by the Radiologist. The testing result of the k-fold cross validation demonstrates that the use of the algorithm for the multiplication of intersection rectangle may improve the system performance with accuracy, sensitivity, and specificity of 76%, 89%, and 63%, respectively.**

***Keywords*—CADe system; haar; cascade classifier; mammogram**

I. INTRODUCTION

Breast Cancer is the second most common forms of cancer in the world and has the first position as the most common form of cancer among women (World Cancer Research Fund International). One of eight women are under the risk of being diagnosed with breast cancer during their lifetime (WHO). The exact causes of the emergence of the cancer cells are not yet known. Therefore, a preventive action by performing the breast screening has a very significant role in reducing the number of victims [1]. The recommended imaging technology is mammography because it has more advantages than the other imaging. Furthermore, the Radiologist will give an assessment on the mammogram image. The interpretation screening on a mammogram is a challenge for the Radiologist because there are often some difficulties in finding the abnormal parts (the disorders) on the mammogram, which may happen because of many factors [2]. The researchers have developed some techniques to improve the Radiologist's performance, one of which is developing a system of computer-aided cancer detection on the mammogram commonly called the Computer-Aided Detection (CADe) System [3].

The CADe system is required to reduce the errors and to improve the Radiologist's ability in making the interpretation on mammography. Some CADe commercial systems have been used by the Radiologist widely. However, those cannot often function optimally yet (there is a positive phase case in the true positive case) [4]. The researchers are still trying to optimize the performance of the CADe system on using the mammogram image shown in the recent literature. In general, the CADe system developed by the researchers is divided into two classes with some kinds of variations of the class type, including: normal and abnormal [5]; mass and non-mass [6] and the finding of microcalcification and not [4][7].

There are some phases in developing the CADe system, including pre-treatment, the determination of RoI (Region of Interest), feature extraction, feature selection, and classification and the testing. The process of determining the RoI is under the direction and guidance of the Radiologist by conducting the cropping on the RoI to obtain some patches and the extraction and the feature selection. The use and selection of features are by viewing the purpose of the classification model development itself. Generally, previous researches, for the CADe system, use the mammography that is developed based on the three features on the mammogram image, those are the features of color, texture and shape. [8] use the color feature, while [9] use the shape feature and [6] combine the shape and texture on their research. Among both features, the last one is most widely used for mammography in previous researches [10][11][12][13] and [14]. The method of feature extraction widely used for the classification of RoI in statistic way is efficient and optimal and may describe the texture of the mammogram itself. Some previous researches that have developed the RoI classification system into two classes RoI (mass and a nonmass) mostly conduct the feature extraction statistically [6] and [12]. However, the determination of the part of RoI is still done manually; the system can only determine the class of RoI.

The next phase for the CADe system is a classification process that serves to classify the RoI predetermined by its feature similarity. Some classification methods commonly used in the field of pattern recognition are the artificial neural networks (ANN) [15], the support vector machine (SVM) [16], and the adaptive network-based fuzzy inference system [17]. [18] have successfully developed a classification method for detecting the face image called the cascade classifier. After that, a lot of researches in the field of computer vision use and develop a cascade classifier for some purposes. Cascade classifier, previously drilled, has been proved to quickly detect the objects that have previously been drilled and successful in some multimedia images. Some problems often emerge in developing the CADe system with the supervised learning for the mammogram image. [7] states that there are two problems: the number of image pixels analysed in the large size and the vast areas of microcalcification (the positive area) that are not greater than the negative area commonly called the class imbalance; the limitations of RoI (positive) is also discussed by [19]. Besides the problem of the RoI limitation for positive samples (learning based cascade classifier), the rectangle of the cropping result of the Radiologist for the same area shows that there are inconsistencies on the part of the Radiologist in providing the markers on the RoI for the same area. Therefore, [19] propose three filtering strategies: sum, mean, and max. The testing result using the Jaccard coefficient has proved that a filter using the max operator has the best performance. The similar problem becomes the concern of [20] that is associated with the class imbalance, in which there are two proposed algorithm: the majority level uses the fuzzy membership function of Gaussian and alpha-cut types to reduce the size of data, while the minority class uses the diffusion membership function of mega- trend to generate some examples for the minority class. The two algorithms are proposed by [20] to cover the two classes that do not have the balanced amount. In this case, the data used is numerical and not in the image.

Therefore, this research develops the use of the cascade classifier concept with the aim to detect whether there are the lesions or not on the mammography. The mammography quality with multimedia images is very different; the mammography has a quality that is very far from the 'ideal' one, which results in difficulties for the Radiologist in identifying the abnormalities in the sought part. This research conducts the feature extraction on the mammography in wavelet by using the Haar feature and the integral image. The limited number of patches on the positive sample for the training data is one obstacle in developing the CADe system using the cascade classifier. Therefore, an algorithm is required to multiply the number of patches as a positive sample on the CADe system to improve its performance.

II. Research Method

The research is carried out in seven phases as follows.

A. Mammografic Image Acquisition and Ultrasonography

The image acquisition process is produced by the mammography and ultrasonography imaging technology from 60 Probands and patients in the Clinic of Oncology Kotabaru Yogyakarta. The breast imaging with mammography is conducted in two views: mediolateral-oblique (MLO) and craniocaudal (CC).

B. Annotations and Cropping Region of Interest (RoI)

After obtaining the image for each of these categories, the Radiologist gives an assessment on the mammography to provide an annotation on the part that is considered a disorder of the breast tissue. Furthermore, the given annotation ROI is classified into two categories: lesions and non-lesions. In determining an annotation on the mammogram image, the Radiologist also interprets the ultrasonography image to convince the truth of interpretations conducted visually on the mammography.

C. Multiplication of intersection (rectangle)

Based on the RoI cropping result by the Radiologist, there are some rectangles intersecting one another. Besides, the limited number of patches as the positive RoI samples (lesions) becomes an obstacle to the process of training using the Haar cascade classifier and may affect the recognition accuracy level. Therefore, to increase the RoI (the intersected rectangles), algorithm 1 and algorithm 2 are proposed, from now on called the algorithm for the multiplication of intersection rectangle. Algorithm 1 will work as long as rectangle 1 (X_1; Y_1; W_1; H_1) is not equal to rectangle 2 (X_2; Y_2; W_2; H_2).

```
Algorithm 1
WHILE ($X_1$< $X_2$) OR ($Y_1$< $Y_2$) OR ($W_1$> $W_2$) OR ($H_1$> $H_2$) //
as long as rectangle 1 ≠ rectangle 2
DO
     IF ($X_1$< $X_2$)
        $X_1$ += 1;
     IF ($Y_1$< $Y_2$) AND ($X_1$ = $X_2$)
        $Y_1$ += 1;
     IF ($W_1$> $W_2$)
        $W_1$ -= 1;
     IF $H_1$> $H_2$ AND $X_1$ = $X_2$
        $H_1$ -= 1;
     DRAW RECTANGLE( $X_1$, $Y_1$, $W_1$, $H_1$ );
ENDWHILE
```

```
Algorithm 2
IF ( R1 ∩ R2 ) //Check if the Rectangles intersect
between one another
BEGIN
     R2 = R1 ∩ R2 //newRect
ENDIF
     Algorithm 1; // use algorithm 1
```

D. Feature Extraction

Conducting the feature extraction of the disorders or RoI grouped into lesions and non-lesions by the Radiologists uses the Haar feature and the integral image to be able to represent the disorder feature on the mammography. The basis of classification on detecting the object lies in the use of some features of Haar-like. Some of these features are represented by the intensity values of the pixels by calculating the value difference between the light-colored pixel area and the dark one. Some of the Haar features can easily do the scaling either being raised or reduced in size to detect the image with various size (Viola and Jones, 2001). There are four basic types that can be used because it is easy to calculate the difference between the white area and the black one using the formula (1)

$$f_i = Sum(r_{i,white}) - Sum(r_{i,black}) \tag{1}$$

$$h_i(x) = \begin{cases} 1, \; jika \; f_i > threshold \\ -1, \; jika \; f_i < threshold \end{cases}$$

E. Feature Selection

Conducting the feature selection by modifying the procedure of Ada Boost further is stated in the points of discussion. The use of the appropriate features greatly affects the accuracy of the system. The features are from the use of Haar-like feature and integral image. The next process is the selection of the best features that will be the basis for the classification in the next process. The algorithm used to select the best feature is the boosting algorithm. AdaBoost training algorithm is used to improve the performance of classification with the simple training algorithms. The feature selection process is by calculating the weight for each feature that is calculated using the formula (2) (Viola and Jones, 2001).

$$F = sign(w_1h_1 + w_2h_2 + \cdots + w_nh_{n)} \qquad (2)$$

$$in\,which, h_i(x) = \begin{cases} 1, \; if\; f_i > \theta_i \\ -1, \; if\; f_i < \theta_i \end{cases}$$

F. Classification of lesions and non-lesions

Conducting the classification between lesions and non-lesions using a cascade classifier further is stated in the points of discussion. The performance scheme of the cascade classifier conducts the classification of RoI based on the features gradually used as shown in Figure 2. The calculation of the classification result has the greatest weight calculated based on the formula (3). The CADe system for the purpose of detecting the RoI consists of two classes: objects and undesired object. Viola and Jones (2001) have developed and tested the classification algorithm to detect whether the frame captured from the image is the form of a face object or not.

$$strong\ classifier = (\alpha_1h_1 + \alpha_2h_2) + (\alpha_Th_T) < Threshold \qquad (3)$$

G. Performance evaluation of CADe system

The testing scheme on the proposed phases of the algorithm in CADe system consists of two types, first, to test the results of training on the training data using the k-fold cross validation; second, to calculate the level of accuracy, sensitivity and specificity of the CADe systems with the results of the assessment and observations of the Radiologist on the mammogram and ultrasonography images. The assessment result of the Radiologist may become the preference in assessing the results of detecting the CADe system. A general description of each phase is shown in Figure 1.

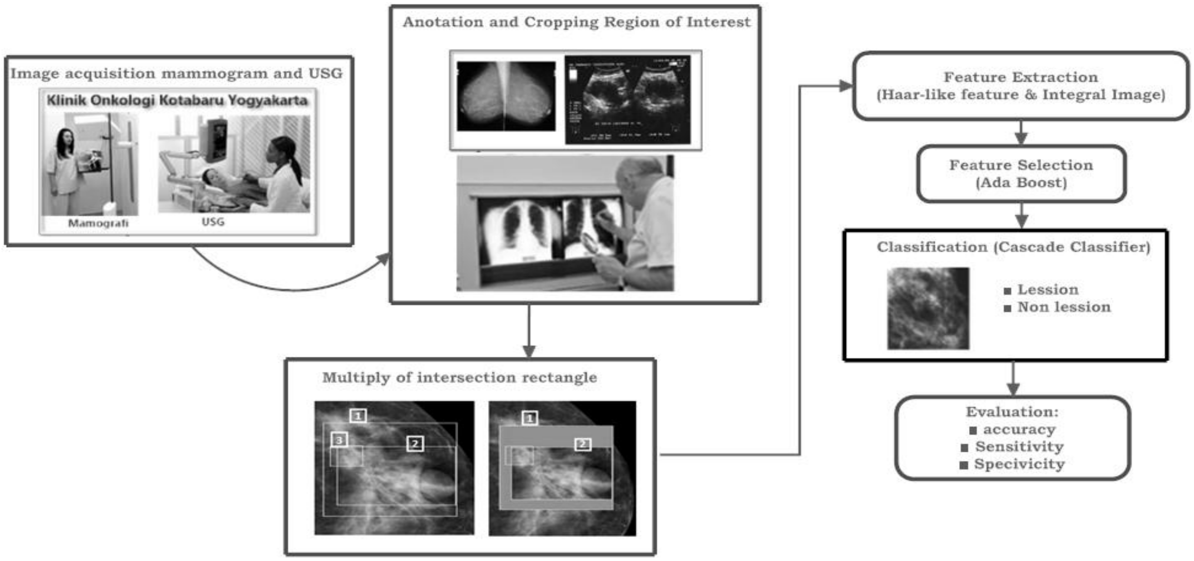


Image acquisition mammogram and USG
Klinik Onkologi Kotabaru Yogyakarta
Mamografi
USG
Anotation and Cropping Region of Interest
Multiply of intersection rectangle
1
3
2
1
2
Feature Extraction
(Haar-like feature & Integral Image)
Feature Selection
(Ada Boost)
Classification (Cascade Classifier)
■ Lession
■ Non lession
Evaluation:
■ accuracy
■ Sensitivity
■ Specivicity

Fig. 1. General description of the research phase

III. Results and Analysis

The first phase conducted is the data acquisition of the mammography and ultrasonography as the result of breast imaging of the Probands and Patients in the Clinic of Oncology. After that, the annotation is conducted on the mammogram image on any part considered as the disorder by the Radiologist with the assessment validation using the ultrasonography image. From the cropping result conducted by the Radiologist, it may be inferred that there is only one rectangle/RoI on a single image, but on some mammogram images, it is found that the Radiologists give the annotations more than once at the close/intersected locations, and some are at different locations. Figure 2 shows examples of the mammogram image with annotation by the Radiologist more than one rectangle is shown in Fig. 2. In general, the results of annotations on more than one RoI may be grouped into two types: between a rectangle with the other one is apart from each other or not intersect to each other, while it is also found that a rectangle and the other one is intersecting to each other.

After the cropping of the RoI (lesion) by the Radiologist, the CADe system is also able to show the results of the cropping. If the Radiologist finds lesions in more than one location as shown in Fig. 3, it is necessary to give the special treatment so that the area as the intersected result of the both rectangles may have the pixel shifting gradually.

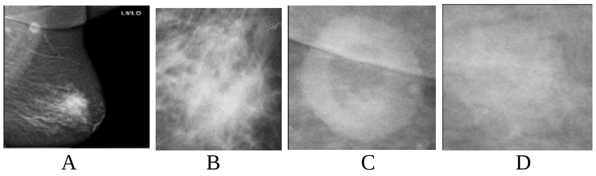


Fig. 2. Example of early image to do the cropping (a), on that image there are the lesions in 3 different locations (b,c,d)

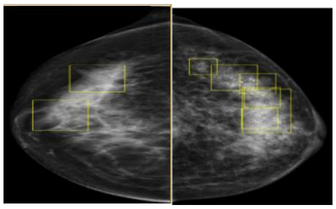


Fig. 3. The annotation result of the Radiologist: there is no intersection (left), and there is intersection (right)

Examples of the result of the assessment and the administration of annotations on the mammography with the view of MLO on the left and the right side with the RoI cropping results show that there are separated rectangles (not intersecting) as illustrated in Figure 4 and 5. Figure 6 shows the view of CC on the left and the right sides which rectangles are not intersecting. Figure 7 illustrates the example of an assessment result of mammography with the view of MLO on the left and the right sides and the RoI cropping result showing that there are intersecting rectangles. Figure 8 and 9 show the view of CC on the left and the right sides which rectangles are intersecting.

Therefore, this research proposes an algorithm for the multiplication of intersection rectangle to increase the intersecting rectangle patch by conducting the pixel shifting. Both algorithms are used before the RoI feature extraction. There are two algorithms, the first algorithm is used when there are two annotations (rectangles) that have different sizes, one is contained in the other. Figure 10 shows that rectangle 2 is in rectangle 1.

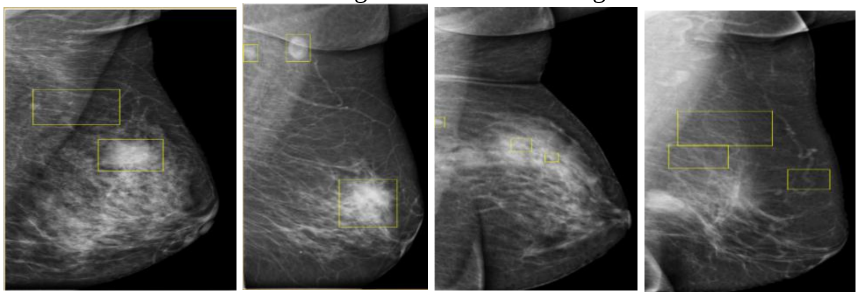


Image 8 Image 11 Image 74 Image 76

Fig. 4. The RoI *cropping* result with the view of MLO on the left side

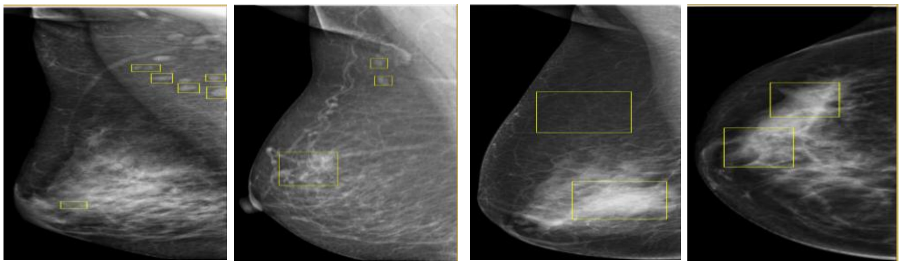


Image 20 Image 52 Image 38 Image 50

Fig. 5. The RoI cropping the view of MLO on the right side

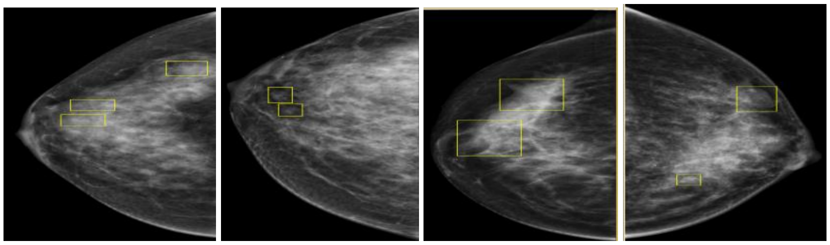


Image 19 Image 80 Image 50 Image 89

Fig. 6. The RoI cropping result with the view of CC on the right side (image 19, 80 and 50) and the left side (image 89)

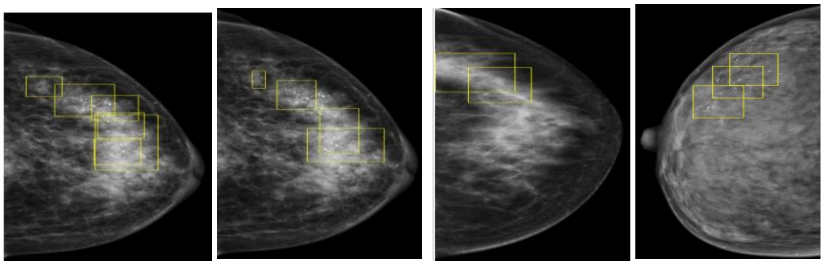


Image 15 Image 69 Image 48 Image 67

Fig. 7. The RoI cropping result is intersecting the view of CC on the left side (image 15, 69 and 48 and the right side (image 67)

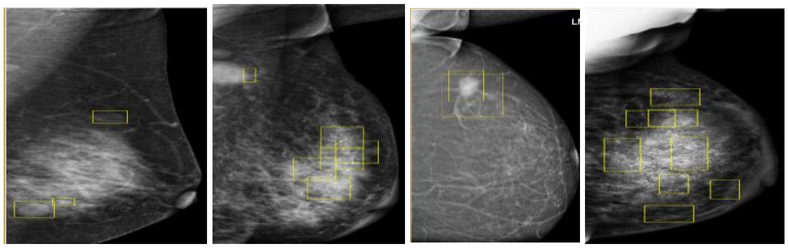


Image 78 Image 70 Image 47 Image 45

Fig. 8. The RoI cropping results with the view of MLO on the left side

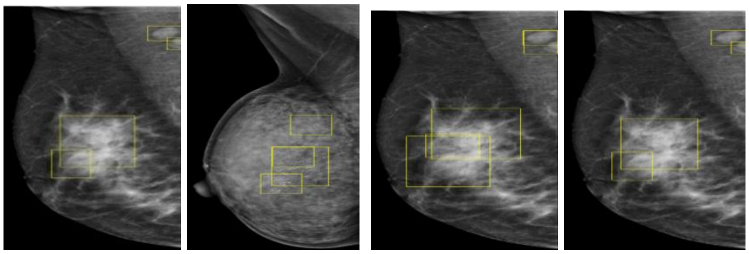


Image 27 Image 68 Image 51 Image 27

Fig. 9. The RoI cropping results with the view of MLO on the right side

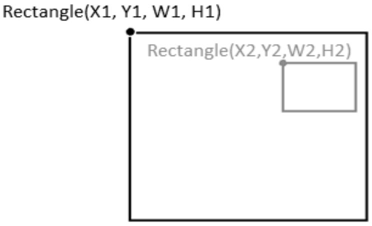


Fig. 10. Illustration of one rectangle is inside the other one

The aim is to make more RoI, or to make the difference between the two rectangles by making a positive sample from rectangle 1 to rectangle 2. The second condition, if there are two annotations (rectangles) one side of which (the line) intersects (tangent) with the side or the line of the other rectangle. The illustration can be shown as Fig. 11, where rectangle 1 intersects or has tangent with rectangle 2 (R1 ∩ R2), so the side that will have the pixel shifting (duplication) is the area included in the area of R1 and R2.

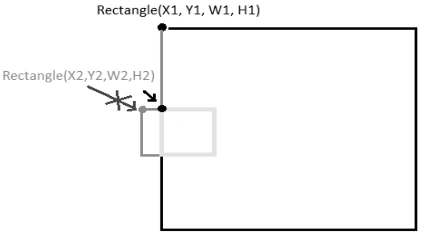
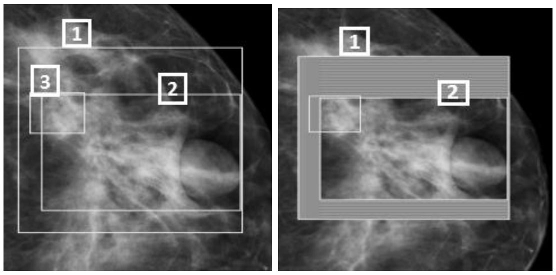


Fig. 11. Illustration of a rectangle intersecting or having tangent with another rectangle

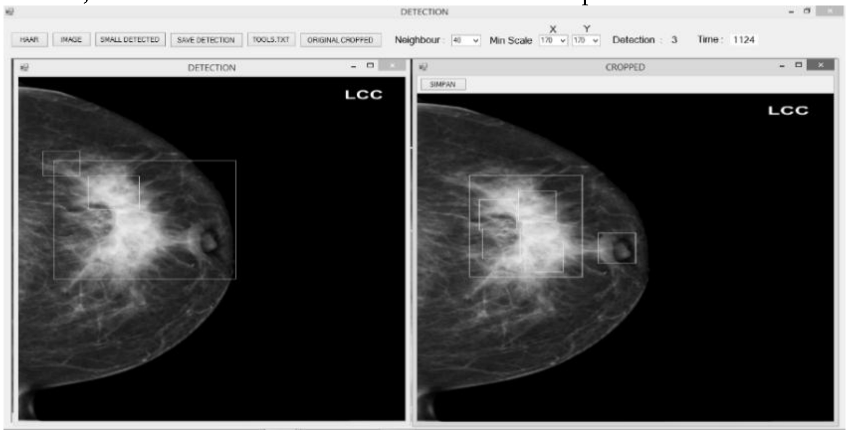
The second algorithm is to create more rectangles. The multiplied area is the one resulting from the intersection of both rectangles. The workings of the algorithm proposed are as follows: Previously some processing on the image of the cropping results (RoI or abnormality or lesion) are conducted by the Radiologist. The image that will have the multiplication of RoI certainly is the mammogram image that has been given more than one annotation by the Radiologist and is intersecting, which means that the Radiologist conducts the RoI cropping on one image with more than one rectangle. Furthermore, the RoI will have the image multiplication by processing the pixels sequentially and simultaneously starting from the largest RoI to the other intersection of the RoI. Figure 12 illustrates the RoI multiplication process with rectangle 1 and 2. The area on rectangle 2 will have the pixel shifting to rectangle 1 to make (X_1; Y_1; W_1; H_1) equal to (X_2; Y_2; W_2; H_2). The algorithm for the multiplication of intersection rectangle from the two rectangles will create a lot of rectangles with different positions and grayscale values at each pixel.



(a) (b)

Fig. 12. (a) The cropping result by the Radiologist (b) The multiplication result of RoI 1 and RoI 2.

There are some pre-treatments and trials before the training, including the process of creating the sample sized in 40 x 40 pixels is conducted before the image sample of the cropping result of the Radiologist whether there is an intersection of one another, or there is no intersection. Based on the experiment result with a few number of stages, finally, it obtains the best weight of the training experiment that is shown in the cascade classifier with eight stages. The bigger the stage value is, the more accurate the detection will be. However, the number of stages also should consider the number of the positive samples. In this research the determination of the number of stages is conducted by the trial error and the most optimal number of stages obtained is 8 stages. The maximum value of error (maximum false alarm) that may be received is 30%, which means that 30% of the negative sample may be detected as the positive ones. The greater this value is, the more inaccurate the detection will be. However, it cannot replace it with a value of 0% because the training process will not finish. There is an experiment on the neighbourhood value to obtain the best result in this research, and the best result is at a value between 30 and 40. The higher the neighborhood value is, the more accurate the detection will be, but if it is overdone it will not be able to detect anything on the mammogram images. While for the minimum scale used to find the pixels that is detectable, this research uses a scale of 170x170 after doing the trial and error to see data from the cropping result of the Radiologist on the file 'info.txt'. The larger the scale is, the less time it takes to process an image. The example of the result of the detection system using the algorithm for the intersecting rectangles to increase the RoI in the image, in which there is more than one intersecting rectangle, is the mammogram image as the annotation results by Radiologist, and the annotation resulted from the CADe system as shown in Figure 13.



Neighbour : 40 Min Scale X 170 Y 170 Detection : 3 Time : 1124

Fig. 13. The annotation is given by the Radiologist (right), the annotation as the detection result of the CADe system (left)

The testing result conducted in this research shows a significant difference between the CADe system that uses the algorithm for the multiplication of intersection rectangle with the one without using the algorithm with an accuracy of 44% and 76%, a sensitivity of 41% and 89% and a specificity of 48% and 63% as shown in Figure 14. The use of the algorithm for the multiplication of intersection rectangle may improve the accuracy of the CADe system on the mammogram image that has a number of positive rectangles for the positive class (there are lesions) that is very little, whereas the number of the training data largely determines the success of a system in the training process.

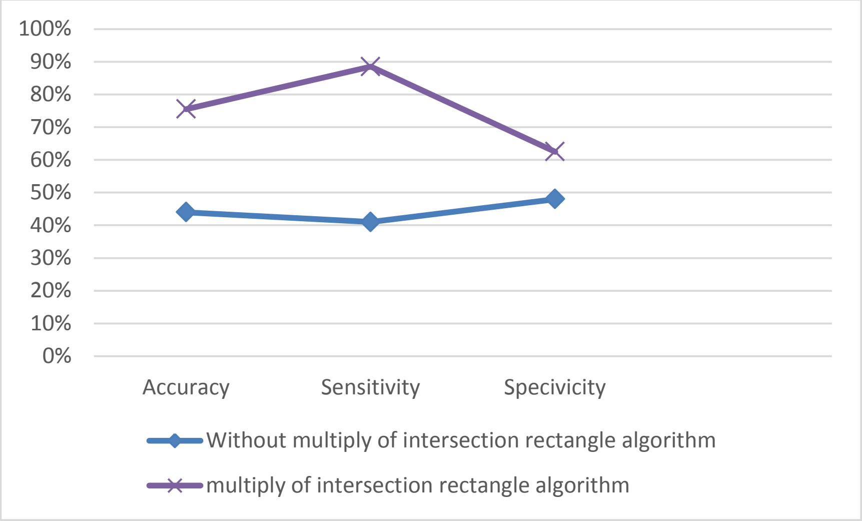


Fig. 14. The testing result with and without the multiplication of intersection rectangle algorithm

IV. Conclusion

The proposed algorithm for the improvement of selecting rectangle that aims to multiply the patch of RoI for a positive sample on the CADe system is proved to be able to improve the system performance. The testing result using the k-fold cross validation shows that the automatic detection of lesions using an approach based on the cascade classifier with the algorithm for the improvement selecting rectangle may detect the RoI much better with the level of accuracy, sensitivity, and specificity of 76%, 89%, and 63%, respectively. Meanwhile, without using the algorithm for the improvement of selecting rectangle it only has the level of accuracy, sensitivity and specificity of 44%, 41%, and 48%, respectively. However, it is required to develop other methods such as using the fuzzy systems to cope with the process of training on the CADe system with the very limited training data.

Acknowledgment

This research is supported by the Directorate Higher Education Islamic, the Directorate Geeral Islamic Education, the Minister of Religious Affairs of the Republic of Indonesia.

References

[1] A. Akhsan & T. Aryandono. Prognostic factors of locally advanced breast cancer patients receiving neoadjuvant and adjuvant chemotherapy. *Asian Pac J Cancer Prev*, 2010, Vol. 11, pp. 759-761.

[2] J.S. Drukteinis, E.C. Gombos, S. Raza, S.A. Chikarmane, A. Swami & R. L. Birdwell. MR imaging assessment of the breast after breast conservation therapy: distinguishing benign from malignant lesions. *Radiographics*, 2012, Vol. 32, No. 1, pp. 219-234.

[3] L.H. Eadie, P. Taylor & A. P. Gibson. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *European journal of radiology*, 2012, Vo. 81, No. 1, pp. e70-e76.

[4] C. Marrocco, M. Molinara, C. D'Elia & F. Tortorella. A computer-aided detection system for clustered microcalcifications. *Artificial intelligence in medicine*, 2010, Vol. 50, No. 1, pp. 23-32.

[5] C. C. Jen & S. S. Yu. Automatic detection of abnormal mammograms in mammographic images. *Expert Systems with Applications*, 2015, Vol. 42, No. 6, pp. 3048-3055.

[6] M. L. de Oliveira, G. B. Junior, A. C. Silva, A. C. de Paiva & M. Gattass. Detection of masses in digital mammograms using K-means and support vector machine. *Electronic Letters on Computer Vision and Image Analysis*, 2009, Vol. 8, No. 2, pp. 39-50.

[7] A. Bria, N. Karssemeijer & F. Tortorella. Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications, *Medical image analysis*, 2014, Vol. 18, No. 2, pp. 241-252.

[8] M. Langarizadeh & Mahmud. R. Breast Density Classification Using Histogram-Based Features. *Iranian Journal of Medical Informatics*, 2012, Vol. 1, No. 1, pp. 1-5.

[9] A. Vadivel & B. Surendiran. A fuzzy rule based approach for characterization of mammogram masses into BI-RADS shape categories. Computers in Biologyand Medicine, 2013, Vol. 43, No. 4, pp. 259-267.

[10] A. M. Khuzi, R. Besar, W. W. Zaki & N. N. Ahmad. Identification of masses in digital mammogram using gray level co-occurrence matrices. *Biomedical Imaging and Intervention Journal*, 2009, Vol. 5, No. 3.

[11] T. S. Subashini, V. Ramalingam & S. Palanivel. Automated assessment of breast tissue density in digital mammograms. *Computer Vision and Image Understanding*, 2010, Vol. 114, No. 1, pp. 33-43.

[12] I. K. Maitra, S. Nag & S. K. Bandyopadhyay. Identification of abnormal masses in digital mammography images. *International Journal of Computer Graphics*, 2011, Vol. 2, No. 1, pp. 17-29.

[13] M. A. Al Mutaz, S. Dress and N, Zaki. Detection of Masses in Digital Mammogram Using Second Order Statistics and Artificial Neural Network. *International Journal of Computer Science and Information Technology (IJCSIT)*, 2011, Vol. 3, pp. 176-186.

[14] S. 'Uyun, S. Hartati, A. Harjoko & Subanar, Selection mammogram texture descriptors based on statistics properties backpropagation structure, *International Journal of Computer Science and Information Security (IJCSIS)*, 2013, Vol. 11, No. 5, pp. 1-5.

[15] J. Jiang, P. Trundle & J. Ren. Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics*, 2010, Vol. 34, No. 8, pp. 617-631.

[16] S. W. Borges, D. E. Moraes, S. A. Correa, P. A. Cardoso & M. Gattass. Detection of masses in mammogram images using CNN, geostatistic functions and SVM. *Computers in Biology and Medicine*, 2011, Vol. 41, No. 8, pp. 653-664.

[17] F. C. Fernandes, L. M. Brasil, J. M. Lamas & R. Guadagnin. Breast cancer image assessment using an adaptative network-based fuzzy inference system. *Pattern Recognition and Image Analysis*, 2010, Vol. 20, No. 2, pp. 192-200.

[18] P. Viola, & M, Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001, Proceedings of the 2001 IEEE Computer Society Conference*, pp. I-511.

[19] E. Cheng, H. Ling, P. R. Bakic, A. D. Maidment & V. Megalooikonomou. Automatic detection of regions of interest in mammographic images. In *SPIE Medical Imaging*, 2011, pp. 79623J-79623J. International Society for Optics and Photonics.

[20] D. C. Li, C. W. Liu, & S. C. Hu. A learning method for the class imbalance problem with medical data sets. *Computers in biology and medicine*, Vol. 40, No. 5, 2010, pp. 509-518.