Improvement of Automated Detection Method for Clustered Microcalcification Based on Wavelet Transformation and Support Vector Machine

Kohei Arai, Indra Nugraha Abdullah, Hiroshi Okumura, Rie Kawakami
Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*—The main problem that corresponding with breast cancer is how to deal with small calcification part inside the breast called microcalcification (MC). A breast screening examination called mammogram is provided as preventive way. Mammogram image with a considerable amount of MC or called clustered MC has been a problem for the doctor and the radiologist. Particularly, when they should determine correctly the region of interest. This work is an improvement work from the previous work. It utilizes the Daubechies D4 wavelet as a feature extractor and the SVM classifier as an effective binary classifier. The escalating point shown with 84.44% of classification performance, 90% of sensitivity and 91.43% of specificity.**

Keywords-Automated Detection Method; Mammogram; Clustered Microcalcification;Wavelet; SVM; Standard Deviation.

I. INTRODUCTION

Breast cancer is the uncontrolled growth of breast cells caused by a genetic abnormality. Mostly breast cancer starts from lobules cells, glands or milk producer, and duct cells. Duct cells are parts that transporting milk from the lobules to the nipple. A tumor can be categorized into two main types. First is benign type with characteristic nearly similar with the normal one in appearance such as slow growth, will not spread to the other body parts. The second is a malignant type with vice versa characteristics from benign type.

Based on the Globocan, an international World Health Organization agency for cancer located in France, breast cancer is the most frightening cancer for women in the world, and become the most common cancer both in developing and developed regions. In 2008 estimated 1.38 million new cancer cases diagnosed, the proportion of breast cancer was 23% of all cancers.

TABLE I. SUMMARY OF BREAST CANCER INCIDENCE AND MORTALITY WORLDWIDE IN 2008

Region	Cases	Deaths
World	1384	458
Africa Region	68	37
American Region	320	82
East Mediterranean Region	61	31
Europe Region	450	139
South-East Asia Region	203	93
Western Pacific Region	279	73

In table 1, we can notice to all regions, the rates of mortality are extraordinarily high and obviously there is no region in the world that has not affected with this cancer. The most worrisome region is Europe region with the number of incidence case and mortality case are 450 and 139, respectively. That means the rate of mortality in this region is 0.308 and made this rate is similar to the rate of the world region, which is 0.331.

In order to overcome this problem, every woman needs to concern about their health through various continuous tests. Breast cancer tests covering screening tests, diagnostic tests, and monitoring tests. In this study, we will focus on the test in screening tests called Mammograms, the most valuable tool not only to screen the cancer, but also to diagnose and evaluate.

Mammogram can read any signs of abnormality such as asymmetry of shape, irregular areas, clusters of small microcalcification (MC) and area of skin thickening. Commonly, the radiologist also operates a Computer Aided Diagnosis (CAD) system. This system will analyze the digital format of mammogram, and the result is a mammogram with any markers in the suspicious areas. The difficulty for the system is to discover clustered extra small calcifications in the form of clusters called with clustered MC.

Abdallah et.al [2] reported the efficient technique to detect the ROI using multi-branch standard deviation analysis and resulting the promising result which more than 98% of true positive (TP) cases. Papadopoulos et.al [3] had proposed the work consists of three stages, first was the cluster detection stage, second stage was the feature extraction stage, and the final stage was the classification stage. In the final stage, they were comparing neural network (NN) and support vector machine (SVM). Accomplished that performance of the SVM was greater than the NN. The most current one is Tieudeu et.al [4] detected the clustered MC based on the analysis of the their texture. Selection process had done via labeling method of the image that obtained from subtraction the smoothing image from the contrast enhance image, and classification of features completed by neural network. This method was resulting superfine sensitivity equal with 100% and 87.7% of specificity with proper classification rate 89%.

We also had conducted the similar work. However, previous work's result on classification utilizing neural network was deficient [5]. This system needs a new method to achieve better performance of classification, as well as sensitivity and specificity results. Wavelet transform decompose an image to high and low frequency of data when the MC closely related with the high frequency data. Combination of wavelet transform with the SVM classifier as a powerful tool for classifying binary class can be a guaranteed point to obtain better results. Therefore, we propose Daubechies D4 wavelet coefficient's feature with the SVM.

II. Proposed Method

A. Dataset

Dataset comes from Japanese Society of Computer Aided Medical Imaging Technology. Each image has size 2510 x 2000 pixels with single pixel consist of 10 bits. This dataset has two types of images, one-side of breast and two-side of breast. The following images are shown as sample images from the dataset:

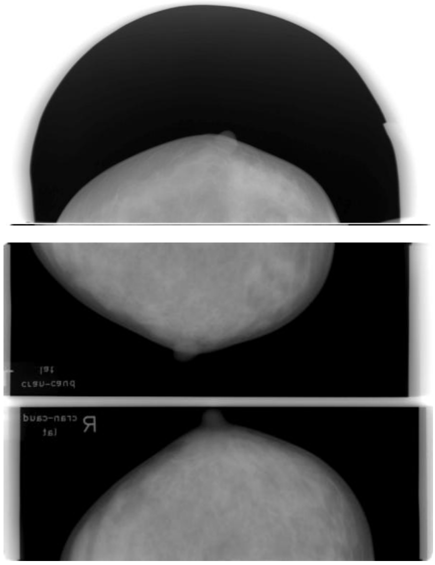


Fig. 1. Samples of mammogram image inside the dataset. One-side of breast (top), two-side of breast (bottom).

There are three categories inside the dataset, normal (N), calcification (C), and tumor (T). However, since the detection for tumor category is easy even using human perception. We only consider two categories that are N and C and inside the C class can be found some numbers of clustered MC. The number of images in C and N categories are 12 images, 33 images, respectively.

B. Detection of MC and Clustered MC

1) Breast Tissue Detection Based on Texture-based Analysis

In this study, we apply the method that has developed by Tieudeu et.al [4] with modification in one specified area. They are developed the main method by utilizing three methods. First is enhancing the contrast of the original image then produce an image called contrast enhance image (CI) and the way to get this image become a point of modification. The second is smoothing the original image then produce an image called with smoothed image (SI). The last is subtraction the smoothed image from enhanced image then called with difference image (DI).

This adoption motivated by clustered MC that allied with breast mass can be concluded as a benign or even premalignant cancer. Frequently, MC only associated with extra cell growth inside the breast. This method also can save time and memory processing for further process. Different with the previous study when forming the CI, we are using the histogram equalization method with the aim to spread the most frequent intensity values that make the lower contrast reach a higher contrast. The details are represented by the equation below:

$$prob_n = \frac{\#pixels\ intensity\ n}{\#pixels};\ \ n = 0, 1, 2 \dots L \tag{1}$$

$$M_i = floor\left((L)\sum_{n=0}^{i} prob_n\right) \tag{2}$$

Where $prob_n$ denotes the normalized histogram for each gray level value, n is gray level values, L is maximum gray level value and M is image matrix.

2) Multi-branch Standard Deviation Analysis

MC related with local maxima pixels in the image. This motivated us to find a correlation between the local maxima and its neighboring pixels. In this study, we conduct an analysis utilizing standard deviation method to find this correlation as reported by Abdallah et.al [2]. Develop a multi-branch point of view become basic needs. It because highly possible if we find a local maxima in one direction and after take a look in a different direction that point is not a local maxima. That critical point provides promising solution to find the clustered MC in one small area. The illustration provided as below:

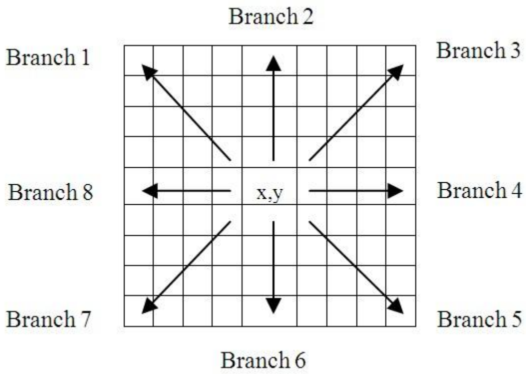


Fig. 2. Multi-branch standard deviation analysis to find MC.

Where x, y point is an ideal local maxima if from all branches seen as a local maxima. Dissimilar with Abdallah et.al work, we only calculate the standard deviation for the n-brightest pixel only.

At the time that we need to know one point is local maxima from one branch, the threshold value and the counter needed. While calculating the threshold between the central pixel and its neighbor pixels if the standard deviation greater than the threshold value the counter will be increasing by one, whereupon an ideal local maxima is the point that has a counter value equal with eight. Described with the following equation:

$$STD_i = \sqrt{\frac{\sum_{i=1}^{n}(Center - x_i)^2}{n}};\ \ i = 1, 2, \ldots, 8 \tag{3}$$

Where:

STD_i	=	Standard Deviation at branch *i*
Center	=	Cluster center
x_i	=	Gray level value at the specified position i
n	=	Number of pixels

As said before the counter will have a maximum value 8, that value is equal with a total of branches in this method.

The mammogram image is scanned under 0.1 mm x 0.1 mm, acquired the number of pixels is 10 x 10 pixels/mm. Regarding the size of single MC is under 1 mm, the 9 x 9 window size will be appropriate enough to detect MC. Region of interest (ROI) as a final result of this section has size 128 x 128 which matched with the most clustered MC's size. In this study, one mammogram image represented by one ROI although there is more than one clustered of MC can be found. It is because this system has a purpose as assistance to the doctor and the radiologist when they are facing the clustered MC. Even if only one representation of clustered MC is found still means the patient defined as calcification's patient and need further treatment. Moreover, selection criterion of ROI is the area with the highest number of suspicious local maxima pixels (MC).

C. Daubechies D4 Wavelet Transform

For $N \epsilon \mathbb{N}$, Daubechies wavelet of class D-*2N* is function $\psi = {}_N\psi \ \epsilon \ L^2(\mathbb{R})$ denoted by

$$\psi(x) := \sqrt{2}\sum_{k=0}^{2N-1}(-1)^k h_{2N-1-k}\varphi(2x-k), \tag{4}$$

where $h_0, \ldots, h_{2N-1} \ \epsilon \ \mathbb{R}$ are the constant filter coefficients that fulfilling the conditions

$$\sum_{k=0}^{N-1} h_{2k} = \frac{1}{\sqrt{2}} = \sum_{k=0}^{N-1} h_{2k+1}, \tag{5}$$

similarly, for $l = 0, 1, \ldots, N-1$,

$$\sum_{k=2l}^{2N-1+2l} h_k h_{k-2l} = \begin{cases} 1 & if\ l = 0, \\ 0 & if\ l \neq 0, \end{cases} \tag{6}$$

and where $\varphi = {}_N\varphi : \mathbb{R} \rightarrow \mathbb{R}$ is the scaling function, given by the recursive equation

$$\varphi(x) = \sqrt{2}\sum_{k=0}^{2N-1} h_k \varphi(2x-k) \tag{7}$$

Daubechies orthogonal wavelets of classes D2 - D20 (only even index numbers) are the wavelets that generally used [7]. The index number belongs to the number *2N* of coefficient. Single wavelet has a number of vanishing moments equal to half the number of coefficients. In this study we propose to use Daubechies D4 wavelets, it has two vanishing moments. With these vanishing moments D4 can encodes polynomial of two coefficients, for example constant and linear signal components. It will be suitable as a feature extractor for representing clustered MC.

D. Support Vector Machine

SVM is a powerful tool for data classification. The indicators are the easiness to apply and impose Structural Risk Minimization (SRM). SRM armed the SVM to have strong ability in generalization of data. Its function is to minimize an upper bound on the expected risk. In principle, SVM learns to obtain optimal boundary with maximum margin that able to separate set of objects with different class of membership.

In order to achieve the maximum margin classifier, we have two options. Hard margin and soft margin are the options that totally depend on linearity of the data. Hard margin SVM is applicable to a linearly separable dataset. However, often the data is not linearly separable. Soft margin SVM emerged as its solution [8]. The optimization problem for the soft margin SVM presented as below:

$$\min_{w,b}\frac{1}{2}\parallel w \parallel^2 + C\sum_{i=1}^{n}\xi_i$$

$$subject\ to:\quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0. \tag{8}$$

where $w,\ C,\ \xi,\ b$ are the weight vectors, the penalty of misclassification or margin errors, the margin error, the bias, respectively.

In (8) can lead us to efficient kernel methods approach. A kernel method is an algorithm that depends on the data only through kernel function, which computes a dot product in some possibly high dimensional data. Using the function ϕ training vector the input space x is mapped into higher dimensional space. $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is called kernel function. The degree of the polynomial kernel can control the flexibility of resulting classifier [9]. It will be appropriate with this research when we classify the clustered MC and non-clustered MC. Polynomial kernel is shown in following equation:

$$K(x_i, x_j) = (\gamma {x_i}^T x_j + r)^d, \gamma > 0. \tag{9}$$

where γ, r, d are kernel parameters, and i, j denote i^{th}, j^{th} vector in dataset.

In this research, we propose to use Sequential Minimal Optimization (SMO). SMO act as efficient solver of the optimization problem in training of support vector machines. SMO also solves the problems analytically by way of breaks the problems into a series of smallest possible problems.

Despite of this algorithm guaranteed to converge, it used heuristics to choose the pair of multipliers that able to accelerate the rate of converge.

III. Experiments

A. Detection of MC

Through the described method, we obtained all images called the CI, SI and DI. From below DI image, we can obviously see the breast tissue area. This area will be the main concern when finding the clustered MC. Through the DI also we obtained efficiency in memory and time processing. As an example, shown with the images below:

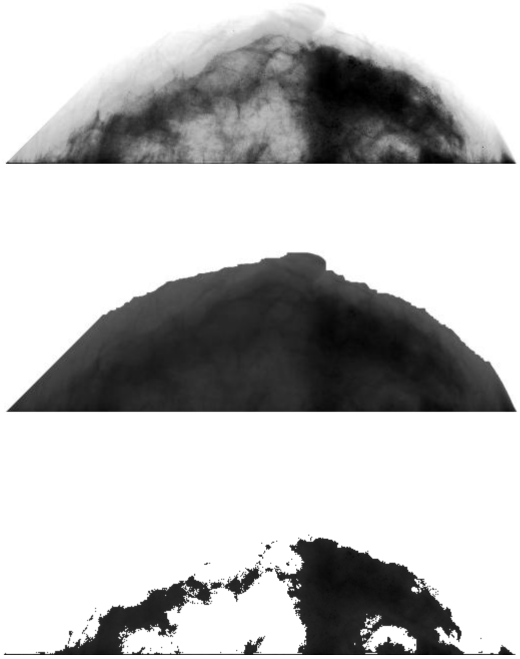
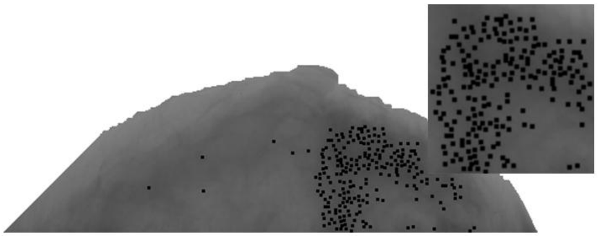


Fig. 3. Sample of the CI, SI and DI images.

Mostly the MCs are detected on this category. It clearly shows that multi-branch standard deviation method with window size 9 x 9 was suitable to detect the clustered MC. The following images show detected MC inside the mammogram image:



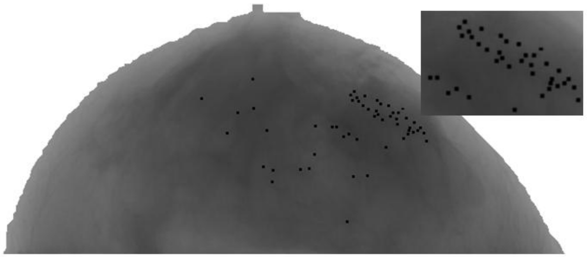


Fig. 4. The samples of detected MC on mammogram images.

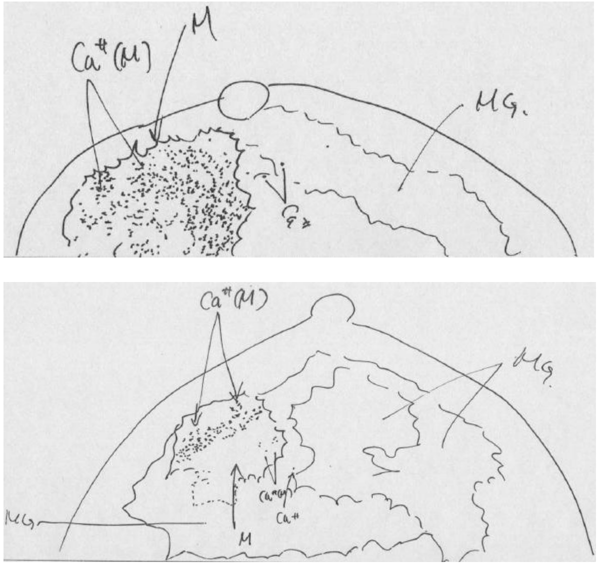


Fig. 5. The sketch images (reverse side with original image) of corresponding mammogram images in Fig. 4

The maximum n value was 200. It means we only calculate 200-brightest pixels inside the mammogram image. After the experiment, threshold value for MC detection equal with 8 was the maximum threshold value.

B. Detection of clustered MC

The size of mammogram image is large, make the efficiency in time and memory processing should be properly considered. As a solution, the ROI selection was not applied pixel-wise detector, but based on 128 x 128 pixels of window-wise detector and moved around the image. The window with the highest number of detected MC will be selected. According to the proposed method, resulting ROI images as presented below:

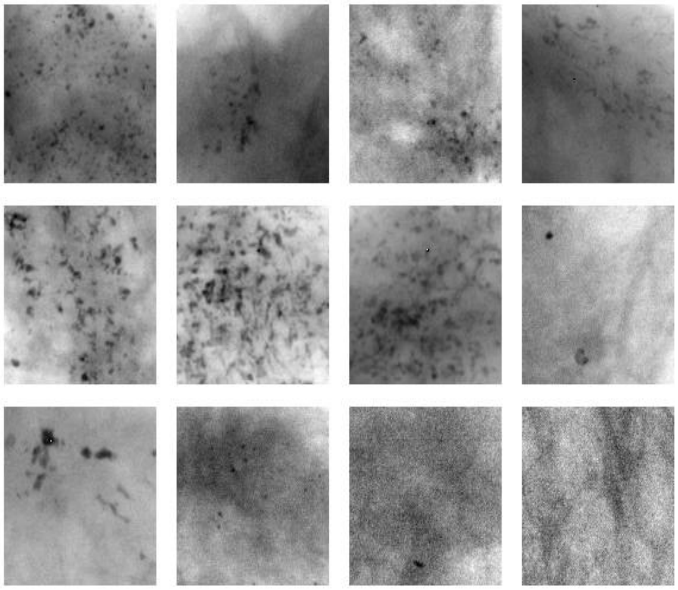


Fig. 6. The detected clustered MC region.

From above image, the system was resulting 12 of clustered MC images. The correct and false detected clustered MC were 9 and 3 images, respectively.

Precisely, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) are the options for diagnosis decision. TP means similarity clustered MC of judgment from an expert and system, TN means similarity a non-clustered MC judgment from an expert and system, FP means a non-clustered MC classified as clustered MC, and last is FN which means a clustered MC classified as a non-clustered MC. After the experiment, the results shown with the following table:

TABLE II. CONFUSION MATRIX

TP	**FP**
9	3
FN	**TN**
1	32

At the comparison process, in order to increase the accuracy of ideal clustered MC image, we were involving the sketch image from the doctor as the knowledge base as shown in Fig. 5.

Hereafter let we talk about other parameters that can indicate the system whether is acceptable or not which are sensitivity and specificity. Both parameters are shown as below:

Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)

Previously, we were working with T category as well as N and C categories. However, similar with aforementioned reason, the determination of T category is something effortless. The doctor will find out quickly the mammogram image that considered as a tumor or not. In this study, we were only working with N and C category. Acquired the sensitivity and specificity values are 90% and 91.43%, respectively.

C. Features Selection

Features selection is based on statistical theory consist of max, mean, variance, standard deviation, coefficient of variation, with two additional features, centroid and 7 Hu moments. These features capture from decomposition result of each detail, approximation, horizontal, vertical, and diagonal details, visualize as shown in Figure.

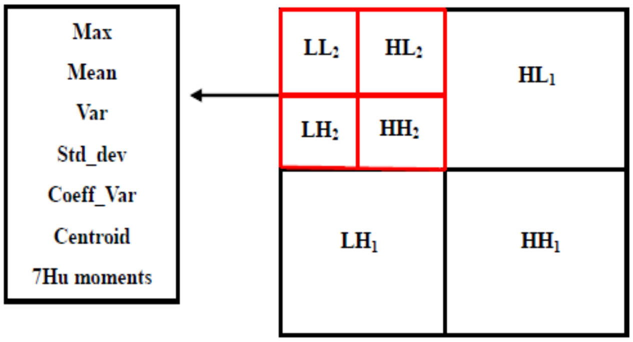


Fig. 7. Features selection from wavelet coefficients.

The mean, variance, and standard deviation are valuable because of their relationships to the normal curve. Variance distribution represents the high frequency data that related with MC. Likewise, max number of wavelet coefficient involved is an additional feature to represent high frequency data. Meanwhile, coefficient of variation shows the extended variability in relation to mean of the population.

The Hu moments are proved to be invariant to the scaling, rotation and reflection. We used centroid to measure contour's centroid of the clustered MC image. From these features, we obtained satisfactory classification results. Approximation included because some of distinct factor that contained in the clustered MC image.

D. Classification performance

In this part, the data set separated into two parts that are training and testing parts with the data proportion were 74% and 26%, respectively. For training data, we were adding ideal output in the form of ROI from all categories manually to train the classifier and then extracted their features. The below table is classification performance:

TABLE III. CLASSIFICATION RESULT USING DECOMPOSITION IMAGE LEVEL 2 OF DAUBECHIES D4 (IN PERCENT)

Method	**L,H,V,D**	**H,V,D**
Training set	95.31	94.53
Supply test set	80	82.22

*L, V, H, D stand for (approximation image, horizontal, vertical, diagonal details)

TABLE IV. CLASSIFICATION RESULT USING DECOMPOSITION IMAGE LEVEL 3 OF DAUBECHIES D4 (IN PERCENT)

Method	**L,H,V,D**	**H,V,D**
Training set	94.53	94.53

Supply test set	84.44	82.22

*L, V, H, D stand for (approximation image, horizontal, vertical, diagonal details)

Even though MC related with high frequency data. We still found small necessity of approximation image. The MC in the form of cluster could well capture while we considered it. Then, the decomposition level 3 of supply test result was resulting more preferable result in comparison with level 2.

The detected MCs that lied in level 3 of the ROI image nearly similar to the sketch result inside the dataset. We decided, not to apply decomposition level 4 because the decomposition result will be a quarter of previous decomposition images, 16x16 pixels image is too small to be recognized by the system.

IV. Conclusion

In this study, we had improvement in classification result, sensitivity and specificity. In comparison with the previous method, this method was finer. The current classification result was much better indicated by 84.44% compared to 70.8% of result in the previous work. The sensitivity and specificity values were also escalated from 79% and 87% to 90% and 91.43%, respectively. Moreover, this system had the efficiency in time and memory processing in term of detection of breast tissue as well as utilization the SVM classifier.

V. Future Work

The future work that can be developed from this current work is conducting another proper method in single MC detection to obtain superfine clustered MCs. One of the promising methods is Dyadic wavelet transformation with its shift invariant characteristic. That detection is to escalate the sensitivity and specificity performances.

References

[1] Songyang Yu and Ling Guan,“A cad system for the automatic detection of clustered microcalcifications in digitized mammograms films,” IEEE Transactions on Medical Imaging, Vol 19(2), pp. 115-125, 2000.

[2] M. H. Abdallah, A. A. Abubaker, R. S. Qahwaji, M. H. Saleh, “Efficient technique to detect the region of interests in mammogram image,” Journal of Computer Science, Vol 4(8), pp. 652-662, 2008.

[3] Papadopoulos A., Fotiadis D.I., Likas A., “Characterization of clustered microcalcifications in digitized mammograms using neural neutwork and support vector machine”, Artificial Intelligence in Medicine Vol. 34, pp. 141-150, 2005.

[4] A. Tieudeu, C. Daul, A. Kentshop, P. Graebling, D. Wolf, “Texture-based analysis of clustered microcalcifications detected on mammograms,”Digital Signal Processing,Vol 22, pp. 124-132, 2011.

[5] Arai K., Abdullah I.N., Okumura H, “Automated detection method for clustered microcalcification in mammogram image based on statistical textural feature”. International Journal of Advanced Research in Artificial Intelligence (IJARAI) Vol. 1(3), pp. 22-26, 2012.

[6] Lucio F. A. Campos, A. C. Silva, A. K. Barros, “Diagnosis of breast cancer in digital mammograms using independent analysis and neural network,” CIARP, pp. 460-469, 2005.

[7] De Vries Andreas, “Wavelets”, FH Sudwestfalen University of Applied Sciences, Hagen, Germany, 2006.

[8] Ben-Hur Asa, Weston Jason, “A user’s guide to support vector machine”, in Data Mining Techniques for the Life Science, pp 223-239, Humana Press, 2010.

[9] Hsu Chih-Wei, Chang Chih-Chuang, Lin, Chih-Jen, “A practical guide to support vector classification”. Department of Computer Science National Taiwan University, Taiwan, 2010.

Authors Profile

KOHEI ARAI received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a counselor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology from 1998 to 2000. He was a counselor of Saga University for 2002 and 2003. He also was an executive counselor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission A of ICSU/COSPAR since 2008.

INDRA NUGRAHA ABDULLAH was born at Bogor, Indonesia in June 1987. Finished his bachelor degree in Bogor Agricultural University and graduated from Saga University for master degree in the field of Information Science on March 2011. He is currently pursuing to get Ph.D. Degree from the same university with specialization in image processing area. Leaf identification becomes his interest in his latest degree.

HIROSHI OKUMURA was born at Kyoto, Japan in 1964. He received B.E.S.E. and M.E.S.E. degree from Hosei University in 1988 and 1990, respectively, and Ph.D degree on environmental engineering from Chiba University in 1993. He became a research associate at Remote Sensing and Image Research Center, Chiba University first in 1993. Next, he became a research associateand a lecturer at the Department of Electrical Engineering, Nagaoka University of Technology in 1995 and 2000, respectively. He is now an associate professor at the Department of Information Science, Saga University. His research interests are in image and speech processing and remote sensing.