Comparison Classifier: Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) In Digital Mammogram Images

Jeklin Harefa[1], Alexander[2], Mellisa Pratiwi [3]

Abstract— In order to begin the initial check on breast cancer, radiologist can use Computer Aided Diagnosis (CAD) as another option to detect breast cancer. During breast cancer check, human error is often to affecting the result. Several research before have proved that CAD is able to detect breast cancer spot more accurate. The purpose of this research is to find reliable method to classify breast cancer abnormalities. Mammography Image Analysis Society (MIAS) database is used as the sample data to the proposed system in this research. Mammograms are divided into three categorize which are normal, benign and malignant according to MIAS database. Features included in this experiment are extracted by using gray level co-occurrence matrices (GLCM) at 0º, 45º, 90º and 135º with a block size of 128x128. In classification process, this research attempt to compare k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM) classifier in order to achieve the better accuracy. The result shows that SVM outperforms KNN in breast cancer abnormalities classification with 93.88% accuracy.

Keywords: Computer Aided Detection, Mammogram, GLCM, k-Nearest Neighbor, Support Vector Machine

I. INTRODUCTION

One of the leading cancer for women in the world is breast cancer. It has to be the number five killer in term of cancer. The alternative way to reduce the number of death caused by breast cancer is by early detection [1]. One diagnostic tool that usually used in breast cancer detection is mammograms, a digital screening image of tissue [2]. The mammograms are then analyzed by radiologist to detect masses in breast cancer as benign and malignant. There are two types of abnormalities of breast cancer, which are benign and malignant. Malignant is the cancerous cells which are dangerous and potentially re-occur. Otherwise, benign is non-cancerous cell which is easy to remove [3]. Since the process of mammograms interpretation done by radiologist is complicated, the result is high in sensitivity but low in specificity. Thus, in some cases, if the chance of the cancer mass is more than 2%, the patients are required to do a biopsy [4].

There have been different algorithm for detecting and classifying the suspicious cancer cells in mammographic images. The CAD’s output is able to help the radiologist in diagnosis breast cancer whether the breast cancer categorized as benign or malignant [5].

In 2013, Fathima et al. [6] presented the first order and gradient features combined with GLCM, Discrete Wavelet Transform (DWT), run length and higher order gradient features to detect the breast cancer in mammogram images. For classifying the accuracy of the proposed method, a Support Vector Machine (SVM) classifier is used in this experiment. The acceptable results obtained in a rapid and simple manner. The percentage of classification rate was up to 95%. Besides, another improvement obtained from the presented method is the reduction in false positive rate, where the false positive number reduced up to 1 for each 100 images.

Aarthi et al. [7] proposed an application of feature extraction and clustering in mammogram classification. For classifying mammogram images, they used neural networks and Support Vector Machine (SVM) as the classifiers. Initially, for each training and test set they divided before, some preprocessing techniques like noise and background removal are applied on mammogram images to extract the required information. Next, the statistical image features are clustered using k-means algorithm followed by SVM classification to classify the mammogram images as benign or malignant. The result gives promosing accuracy of 86.11%, which is higher than the direct classification approach where the accuracy is 80%.

In 2006, Singh et al. [8] attempt to characterizing the mammogram images using classifier Support Vector Machine (SVM). In this research, the experiment are divided into two sub-problems which the first one are detect and recognize the area of suspicious cancer and the second is categorize the suspicious cancer which already found in the first step into benign or malignant. Based on the proposed method, the region are marked as cancer in mammogram. Then, the marked region will be de-noised and enhanced using a method called morphological enhancement. The last step are finding and extracting the features of microcalcification. The extracted features of microcalcification are classified as benign of malignant by using the SVM classifier.

Another research comes from Oliver et al. [9]. The k-Nearest Neighbor and Decision Tree classifier are used to classify the breast cancer abnormalities. Mammographic images are classified by grouping the pixel that have the

[1] *Teknik Informatika, Bina Nusantara University, Jl. Kebon Jeruk Raya 27, Kebon Jeruk, Jakarta Barat 11530, Indonesia; (e-mail: jeklin_harefa@yahoo.com)*

[2] *Teknik Informatika, Bina Nusantara University, Jl. Kebon Jeruk Raya 27, Kebon Jeruk, Jakarta Barat 11530, Indonesia; (e-mail: alexander.fu@hotmail.com)*

[3] *Teknik Informatika, Bina Nusantara University, Jl. Kebon Jeruk Raya 27, Kebon Jeruk, Jakarta Barat 11530, Indonesia; (e-mail: mel2_pratiwi@yahoo.com)*

similar behavior. This method called as gross segmentation. Afterwards, the extracted features are used to categorize the breast as fatty, glandular or dense breast. In the classification stage, two different classifier are used in order to evaluate the texture features. The experimental results demonstrate the probability of estimating breast density with the proposed algorithm.

In 2015, Jog and Mahadik [10] represented a Grey Level Difference Method (GLDM) and Gabor feature extraction methods along with SVM and k-NN classifiers in order to detect the mammographic images for its malignancy. The result shows that the classification accuracy of 50% with k-NN classifier and 95.83% with SVM classifier in GLDM descriptor. While in Gabor texture feature descriptor, the accuracy of 71.83% is achieved with SVM classifier and 58.33% with k-NN classifier. It can be concluded that the best classification accuracy was achieved in the case of GLDM descriptor.

Ramteke and Yashawant [11] produce an automatic medical images classification in two classes such as Normal and Abnormal based on image features and automatic abnormality detection. The system consists of four stages which are pre-processing, feature extraction, classification and post processing. The k-Nearest Neighbor (kNN) Classifier is used to compare with kernel based Support Vector Machine (SVM) classifier (Linear and RBF) for classifying the image. The result of this experiment, Achieve 80% of classification rate using k-NN classifier which is higher than SVM classifier.

Based on those previous researches, this study attempt to compare the two commonly used classifiers for classifying mammograms into benign and malignant abnormalities in the purpose of finding the better classifier. The first classifier used in this study is k-Nearest Neighbor (k-NN) which is widely used in classification process. The second classifier used is Support Vector Machine (SVM) which is also familiar as the robust method in classification.

As for the features, this study use texture features to be evaluated. Texture feature extraction method will be done by using Gray Level Co-occurrences Matrix (GLCM). GLCM is a common texture feature extraction due to simplicity and efficiency since it has less computational complexity in comparison to other methods like wavelet transform [12].

II. METHODS

The digitized mammogram images have been obtained from Mammography Image Analysis Society (MIAS) database. This MIAS database are organized in UK and can be downloaded in http://peipa.essex.ac.uk/pix/mias/. All the images in this database have a pixel size of 1024 x 1024 and physically in portable gray map (pgm) format. It also includes the ROI of the abnormalities that may be present given by radiologist. The MIAS Database has as many as 330 mammogram images which consists of 208 for normal, 68 for benign, and the rest of 54 for malignant. In this experiment, 70 percent of the MIAS database will be used as training data and the rest of 30 percent will be used as testing data. So there are 231 images for training set and 99 images for testing set.

There are three main processes will be conduct in this experiment for classifying mammogram into benign or malignant. The process will start with pre-processing image, followed by extracting features in each image, and last is classification using two classifier. The two classifier: Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN) will be used to classify the breast cancer of abnormalities [13].

In this experiment, the first step for classifying the mammogram images is pre-processing. Each mammogram image will be pre-processed first for getting better quality of the image. Next, the GLCM features will be implemented to extract the texture features of each mammogram image. The two classifier, which are SVM and k-NN will be used and compared for evaluating and achieve the optimum performance of proposed system.

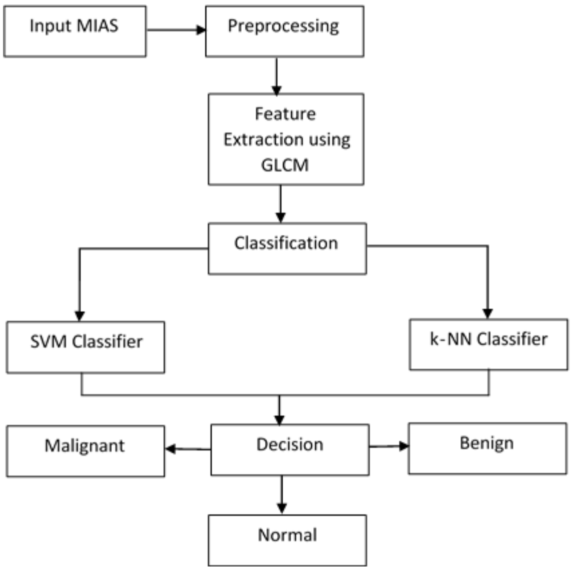

Figure 1. Research Design Flowchart

The pre-processing step focused on improving the image quality by reducing any irrelevant data for better and reliable results. Pre-processing steps are very important in order to search the cancer masses within the background of mammograms. The mammogram image with names of 059.pgm, 212.pgm and 214.pgm are removed from this experiment due to the unknown region of interest. Because the three irrelevant data are benign, so, there will be 327 mammogram images that used in this study. Cropping and resizing each mammogram images to 128 x 128 pixels are another part of pre-processing step in this study.

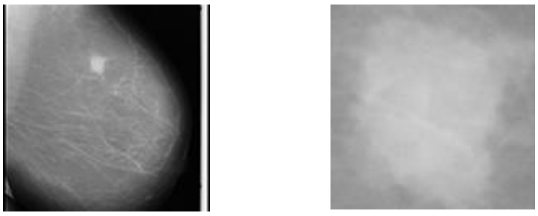

Figure 2. (a) Before image pre-processing (left) (b) After image pre-processing using sample mdb134.pgm (right)

After the pre-processing step is completed, the next step is feature extraction. For mammogram images classification, texture features play important part in differentiating the normal and abnormal breast [14]. One of the known method for textures features extraction is the Gray Level Co-Occurrence Matrix (GLCM) which was proposed by Haralick et. Al [15]. This method has been widely used in many texture analysis applications and it’s already proven that it still better than other texture descriptor. The texture of an image is characterized with GLCM by calculating how often the different combination of pixel which have the gray level value occurred in an image [16].

There are 4 dominant features out of 14 textural features [17] in GLCM. The four features are ASM, Correlation, Sum Entropy and Variance.

- ASM (Angular Second Moment)

ASM is achieved as the amount of the textural uniformity in image (i.e: pixel pairs repetition).

$$ASM = \sum_{i,j} \frac{P(i,j)}{1+|i-j|} \qquad (1)$$

- Correlation

Correlation is used to see the grey level linear dependence to its neighbor.

$$Correlation = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)P(i,j)}{\sigma_i \sigma_j} \qquad (2)$$

- Sum Entropy

Entropy is a measure of the disorder of an image and it gets the largest value when all elements in *P* matrix are equal.

$$SumEntro = \sum_{i=2}^{2Ng} p_{x+y}(i) \log\{p_{x+y}(i)\} \qquad (3)$$

- Variance

Variance of an image is calculated as the average squared derivations from the mean.

$$Var = \sum_i \sum_j (i-j)^2 \, p(i,j) \qquad (4)$$

This study will be used the four dominant features of GLCM with the distance is 1 and the direction θ are 0^0, 45^0, 90^0 and 135^0. The texture of each area in mammogram images will be extracted using these four direction of GLCM.

The third process after extracting feature in mammogram image is classifying the texture features. As mentioned before, this experiment uses Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN). Support Vector Machine (SVM) are supervised learning models that used in computer science for performing classification and regression analysis [18]. This experiment uses a technique called kernel trick to be used in the non-linear data and obtain higher accuracy for classification. This allows the algorithm to find the hyper-plane to differentiate classes in a feature space [19]. RBF kernel is used in this experiment as the most favourite kernel types in SVM [20]. For Gaussian Radial Basis Function:

$$k(x - x') = \exp\left(-\sigma \|x - x'\|\right)^2 \qquad (5)$$

where σ is specified by keyword gamma and must be greater than 0.

This experiment uses two SVM: the first SVM is trained for classifying normal and abnormal training data, and the second SVM is trained for classifying benign and malignant that will be used in the testing phase. If there is a novel input, then the first SVM will classify if the input is normal or abnormal. If the input is abnormal, then the second SVM will automatically categorized the inputted data as benign breast or malignant breast.

The k-Nearest Neighbor (k-NN) is the modest algorithm from among the entire machine learning algorithms [21]. k-NN classifier is used for classifying an object by a majority vote of its neighbor based on *k* most similar vectors presents in feature space, where *k* is a positive integer and typically small [22]. The most similar vector was found using Euclidian Distance between two points using this formula:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (6)$$

where X = ($x_1, x_2, \ldots, x_n$) and Y = ($y_1, y_2, \ldots, y_n$). Here is the simple picture to demonstrate k-nearest neighbor as shown in Fig. 3:

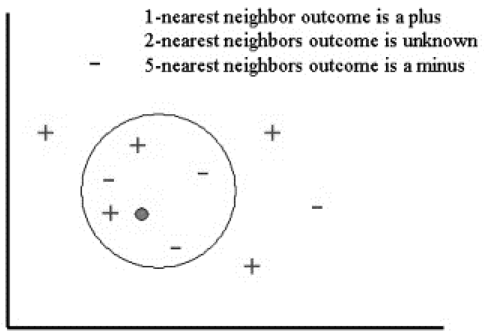
1-nearest neighbor outcome is a plus
2-nearest neighbors outcome is unknown
5-nearest neighbors outcome is a minus

Figure 3 k-Nearest Neighbor demonstration

Based on the Fig. 3., we can see that if k-NN is based on 1 Nearest Neighbor, it is clear that the circle (unknown object) will categorized with a plus (based on the closest point). If number of nearest neighbor is 2, the k-NN will not be able to classify the outcome of circle because the second closest point is a minus. If the number of nearest neighbor is increased to 5, then the k-NN can classify the circle is a minus (3 minus and 2 plus).

For evaluating the accuracy rate of SVM and k-NN classifier, this study uses Confusion Matrix. Sensitivity and specificity are also used as the statistical measure of the cancer detection performance. The accuracy result is used to describe the closeness of a measurement to the true value. The sensitivity is used to identify correctly those who have the cancer if it is present in the breast, while specificity is used to identify correctly those who do not have the cancer masses. Equations (7), (8) and (9) will be used as the evaluation of the performance, respectively.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{7}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{8}$$

$$Specificity = \frac{TN}{TN+FP} \tag{9}$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative. For SVM classifier, this study also gives an overall accuracy with combination of first SVM (normal-abnormal) and the second SVM (benign-malignant) which can be seen in Equation (10).

$$Overall\ Accuracy = \frac{The\ number\ of\ correct\ data}{Total\ data} \tag{10}$$

III. RESULT AND DISCUSSION

The accuracy, sensitivity and specificity result is shown on the Table 1, 2 and 3. Table 1 and Table 2 show the comparison between sigma, accuracy, sensitivity, and specificity for each degree using Support Vector Machine (SVM), while Table 3 shows overall accuracies using SVM. Here is the result of accuracy based on degree:

Table 1 Accuracy, Sensitivity and Specificity of 1st SVM Structures based on Degree

Degree	Sigma	Accuracy	Sensitivity	Specificity
0^o	0.144	91.57%	80.56%	100.00%
45^o	0.144	93.98%	88.89%	97.87%
90^o	0.144	92.77%	88.89%	95.74%
135^o	0.127	91.57%	83.33%	97.87%

Table 1 presents the data with the best results accuracy of classifying normal and abnormal data for any degree. The best accuracy of first structure SVM is obtained from sigma = 0.144 is 93.98% in 45^0.

Table 2 Accuracy, Sensitivity and Specificity of 2nd SVM Structures based on Degree

Degree	Sigma	Accuracy	Sensitivity	Specificity
0^o	0.144	89.66%	92.31%	87.50%
45^o	0.144	100.00%	100.00%	100.00%
90^o	0.144	90.63%	78.57%	100.00%
135^o	0.127	100.00%	100.00%	100.00%

Table 2 presents the data with the best results accuracy of classifying benign and malignant data for any degree. The best accuracy of second structure SVM is 100% where the accuracy is obtained from sigma = 0.144 in 450 and sigma = 0.127 in 1350.

Table 3 Overall Accuracies of SVM Structures based on Degree

Degree	Sigma	Overall Accuracy
0^o	0.144	87.95%
45^o	0.144	93.98%
90^o	0.144	89.16%
135^o	0.127	91.57%

Based on the Table 3, we can conclude that the highest accuracy is achieved at 450 with the accuracy is 93.98% and sigma = 0.144.

For classifying mammogram images using k-Nearest Neighbor (k-NN), this experiment was performing using 83 testing set and 244 training set. Experiment was conducted by using k = 3, k = 5, k = 7, and k = 9 in 0^0, 45^0, 90^0, 135^0 as described in figure 4.

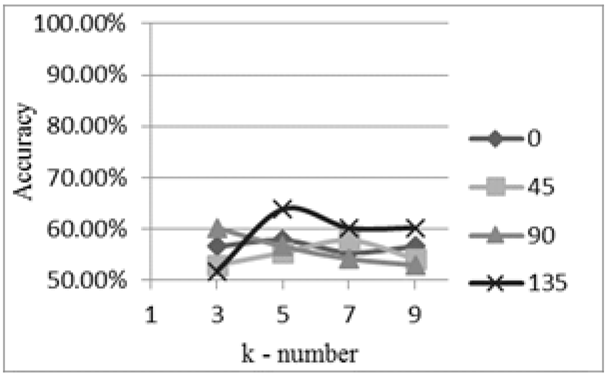

Figure 4 Accuracy Percentage between degree and k number using k-Nearest Neighbor

From the experiments of k-Nearest Neighbor classifier, it can be shown in figure 8 that the highest accuracy is 63.86% in 135^0 and k = 5. Here is the confusion matrix of k-Nearest Neighbor using k = 5 in 135^0:

Table 4 Confusion Matrices of k-NN Classifier

		Automatic Classification		
		Normal	Benign	Malignant
Truth	Normal	45	2	0
	Benign	13	4	3
	Malignant	10	2	4

Confusion matrices as shown in Table 4 should be read as follows: rows indicate the object to recognize (the true class) and columns indicate the label the classifiers associates at this object.

IV. CONCLUSION

This study proposed a reliable for classifying the breast cancer abnormalities. First, each mammogram images will be pre-processed first. Next, the GLCM features will be implemented to obtain the texture images. For classification stage, this study uses and compares two classifiers, which are the k-NN and SVM in order to achieve the best accuracy result. From the result shown, it might be concluded that with the use of GLCM feature, combined with Support Vector Machine (SVM) classifier gives better accuracy than k-Nearest Neighbor (k-NN).

In overall, the accuracy of the GLCM and SVM method is 93.88% with 45^0 and sigma = 0.144. So, the result shows that the GLCM and SVM is reliable method to be used as the bases in developing a CAD system for breast cancer abnormalities classification.

For future work, several other texture feature extraction methods and classifiers will also be combined in the purpose of getting the reliable method to assist the radiologist in interpreting mammograms.

V. REFERENCES

[1] NehaTripathi and S. P. Panda, "A Review on Textural Features Based Computer Aided Diagnostic System for Mammogram Mass Classification Using GLCM & RBFNN," *International Journal of Engineering Trends and Technology (IJETT),* vol. 17, no. 9, pp. 462-464, 2014.

[2] H. Sheshadri and A. Kandaswamy, "Detection of breast cancer by mammogram image segmentation," *Journal of Cancer Research and Therapeutics,* vol. 1, no. 4, pp. 232-234, 2005.

[3] S. Dudea, C. Botar-Jid, D. Dumitriu, D. Vasilescu, S. Manole and M. Lenghel, "Differentiating benign from malignant superficial lymph nodes with sonoelastography," *Medical Ultrasonography,* vol. 15, no. 2, pp. 132-139, 2013.

[4] L. Hadjiiski, B. Sahiner and H.-P. Chan, "Advances in CAD for Diagnosis of Breast Cancer," *Curr Opin Obstet Gynecol,* vol. 18, no. 1, pp. 64-70, February 2006.

[5] K. Bashir and A. Sharma, "Review Paper on Classification on Mammography," *International Journal of Engineering Trends and Technology (IJETT),* vol. 14, no. 4, pp. 169-171, August 2014.

[6] M. M. Fathima, D. Manimegalai and S. Thaiyalnayaki, "Automatic detection of tumor subtype in mammograms based On GLCM and DWT features using SVM," in *International Conference on Information Communication and Embedded Systems (ICICES),* Chennai, 2013.

[7] R. Aarthi, N. K. K. Divya and S. Kavitha, "Application of Feature Extraction and clustering in mammogram classification using Support Vector Machine," in *International Conference on Advanced Computing (ICoAC),* Chennai, 2011.

[8] S. Singh, V. Kumar, H. K. Verma and D. Singh, "SVM Based System for classification of Microcalcifications in Digital Mammograms," in *International Conference of the IEEE Engineering in Medicine and Biology Society,* New York, NY, 2006.

[9] A. Oliver., J. Freixenet and A. Bosch, "Automatic Classification of Breast Tissue," *Pattern Recognition and Image Analysis,* vol. 3523, pp. 431-438, 2005.

[10] N. V. Jog and S. R. Mahadik, "Implementation of Segmentation and Classification Techniques for Mammogram Images," *International Journal of Innovative Research in Science,* vol. 4, no. 2, pp. 422-426, 2015.

[11] D. R. J. Ramteke and K. M. Yashawant, "Automatic Medical Image Classification and Abnormality Detection Using K Nearest Neighbor," *International Journal of Advanced Computer Research,* vol. 2, no. 6, pp. 190-196, December 2012.

[12] P. Mohanaiah, P. Sathyanarayana and L. GuruKumar, "Image Texture Feature Extraction Using GLCM Approach," *International Journal of Scientific and Research Publications,* vol. 3, no. 5, pp. 1-5, 2013.

[13] Y.-D. Zhang, S.-H. Wang, G. Liu and J. Yang, "Computer-aided diagnosis of abnormal breasts in mammogram images by weighted-type fractional Fourier transform," *Advances in Mechanical Engineering,* vol. 8, no. 2, pp. 1-11, 2016.

[14] M. Sharma, R. B. Dubey, Sujata and S. K. Gupta, "Feature Extraction of Mammograms," *International Journal of Advanced Computer Research,* vol. 2, no. 5, pp. 201-209, 2012.

[15] R. Haralick, K. Shanmugam and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 3, no. 6, pp. 610-621, 1973.

[16] N. Zulpe and V. Pawar, "GLCM Textural Features for Brain Tumor Classification," *International Journal of Computer Science,* vol. 9, no. 3, pp. 354-359, 2012.

[17] C.-H. Wei, C.-T. Li and R. Wilson, "General Framework for Content-Based Medical Image Retrieval with its Application to Mammograms," in *Proceedings of the SPIE,* San Diego, CA, 2005.

[18] P. Sharma and M. Kaur, "Classification in Pattern Recognition: A

Review," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, no. 4, pp. 298-306, 2013.

[19] B. E. Boser, I. M. Guyon and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *COLT '92 Proceedings of the fifth annual workshop on Computational learning theory*, Pittsburgh, PA, 1992.

[20] S. Demyanov, J. Bailey, K. Ramamohanarao and C. Leckie, "AIC and BIC based approaches for SVM parameter value estimation with RBF kernels," in *JMLR: Workshop and Conference Proceedings*, 2012.

[21] T. Cover and P. Hart, "Nearest neighbor pattern classification," in *IEEE Transactions on Information Theory*, Menlo Park, 1967.

[22] S. M. Kumar. and G. Balakrishnan, "Classification of Microcalcification in Digital Mammogram using Stochastic Neighbor Embedding and KNN Classifier," in *International Conference on Emerging Technology Trends on Advanced Engineering Research (ICETT'12)*, 2013.