Staging of breast cancer based on the area of the primary tumour

R Embong1,a, M H Sanuddin1 and M S Md Ali1
1Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

aE-mail: rohana@fskm.uitm.edu.my

Abstract. Breast cancer is one of the most common and alarming diseases for women. With early detection and diagnosis, the chances of successful treatment and survival would improve. The diagnosis includes classification or staging of the breast cancer, which plays an important role in the prognosis of the disease hence determining the best treatment for the disease. Breast cancer staging is a way of describing the size of a cancer and how far it has grown. One of the staging numbering systems is the Tumour-Node-Metastasis (TNM) system, which categories cancer into several stages. Most types of cancer have four stages, numbered from I to IV. Normally, staging is determined by doctors by examining the pathology information which describes the spreading area of the cancer. In this study, a series of computer algorithms is applied to produce the area information of the cancerous region. The methodology consists of four phases which are image gathering; image pre-processing using Median filter and PCA; image segmentation using FCM; and staging using the area of the primary tumour as a representation of size in the TNM system. From thirty-five randomly selected mammography images of Malaysian women with malignant tumour, 14.3% is in stage I, 5.7% is in stage II and 80% in stage III. Experimental results also shows that 96.4% accuracy is obtained for stage III but higher errors occur at the boundaries of the staging scale.

1. Introduction
Cancer incidence and mortality are growing worldwide with breast cancer rapidly becomes the leading cause of mortality among women in 2018. This disease contributed to 15% from the total of 4.2 million female cancer deaths [1]. Awareness and early detection of breast cancer may reduce the risk of death [2], [3]. A common modality for breast cancer screening is digital mammography, which is an efficient method to detect breast cancer, even at the early stage [4]. With an increase in the awareness of screening for breast cancer, huge number of mammography images are produced and manual visual examination by expert radiologists becomes a burden. Furthermore, due to human factor, detection of suspicious cancerous may be misinterpreted as a cancer and vice versa, hence leading to false results. Studies have shown that an error between 10% to 30% may occur during the detection process [5] which lead to much effort ([4], [6]) on the automatic process of detection and classification of breast cancer. With the advance in computer technology, computer aided diagnosis (CAD) systems have been developed to assist or act as the second-eye to radiologists, and the application of CAD has proven to improve the performance of radiologists in cancer detection [7].

Many research ([8], [9]) have agreed that accurate diagnosis and treatments are crucial for the survival of cancer patients. The determination of the stage of the cancer remains the most important factor in evaluating the prognosis and choosing the most appropriate treatment for patients with cancer.
According to [9] the staging procedure is based on combining several findings: whether the cancer has metastasized, whether it exhibits invasive characteristics, the diameter of the primary tumour, and the number of malignant lymph nodes. Meanwhile the procedural examination for staging involves physical examination, pathological examination of the tumour, biopsy examination as well as imaging and laboratory blood tests [10]. An accurate estimation of the tumour size is the most basic significant factor in deciding the treatment planning for the cancer disease. Once a cancer is diagnosed, the tumour stage must be accurately determined before therapy can be chosen and the prognosis can be known. Therefore, optimum imaging technique is essential to get good result for initial staging of cancer [10], [11].

Several cancer staging systems have been developed such as Dukes’ system and Tumour-Node-Metastasis (TNM) system, but the Tumour-Node-Metastasis (TNM) system is the most widely applied. The TNM system classifies the stage of cancer based on three major morphological attributes of the malignant tumours, which are the size of the primary tumour (T), the presence and extent of the regional lymph node involvement (N), and the presence of distant metastases (M). Nevertheless, the size of the primary tumour (T) is the most dominant component of the TNM staging system [12], [13], [14]. Both [13] and [14] used area of the primary tumour in mammography images as the representation of size in the TNM staging system. [13] proposed the pixel area scale for stage I, II, and III using Region Growing method and Sample Fuzzy K-means clustering algorithm. Stage IV was not considered because the tumour cancer would vary in size and had spread to other parts of the body, hence further examination was needed. Meanwhile [14] used a different pixel area scale for staging of breast cancer. They applied Fuzzy clustering and Artificial Neural Networks (ANN) and the largest error, 25% occur in stage II using input image from the Mammography Image Analysis Society (MIAS) database. In this study, we applied an improved Fuzzy C-means (FCM) clustering algorithm with image pre-processing using Principal Component Analysis (PCA) and median filter to produce the area information as the representation of the size of the primary tumour in mammography image. In addition, we acquire thirty-five mammography images from the National Cancer Society of Malaysia to investigate staging information of Malaysian women.

2. Methodology

2.1 Mathematical Representation of PCA and FCM

The PCA algorithm is described as follows [15], let \( x = [x_1, x_2, \ldots, x_m]^T \) be an \( m \)-component vector variable and the sample matrix of \( x \) is denoted by

\[
X = \begin{bmatrix}
x_1^1 & x_1^2 & \cdots & x_1^n \\
x_2^1 & x_2^2 & \cdots & x_2^n \\
\vdots & \vdots & \ddots & \vdots \\
x_m^1 & x_m^2 & \cdots & x_m^n
\end{bmatrix}
\]  

(1)

where the discrete samples of \( x_j, i = 1, 2, \ldots, m \) are \( x_j^i, j = 1, 2, \ldots, n \). The \( i \)th row of the sample matrix \( X \) is called the sample vector of \( x_i \) and is written as

\[
X_i = \begin{bmatrix}
x_i^1 \\
x_i^2 \\
\vdots \\
x_i^n
\end{bmatrix}
\]

(2)

The mean value of \( X_i \) is given as

\[
\mu_i = \frac{1}{n} \sum_{j=1}^{n} X_i(j)
\]

(3)
Also, $X_i$ is centralised as follows

$$\overline{X_i} = X_i - \mu_i$$

$$= \left[(\overline{x_i})^1 - \mu_i \right] \left[(\overline{x_i})^2 - \mu_i \right] \cdots \left[(\overline{x_i})^n - \mu_i \right]$$

(4)

Hence, the centralised matrix of $X$ is

$$\overline{X} = \begin{bmatrix} \overline{X_1}^T & \overline{X_2}^T & \cdots & \overline{X_n}^T \end{bmatrix}^T$$

(5)

Finally, the co-variance matrix, $\Omega$ of the centralised data, $\overline{X}$ is calculated as

$$\Omega = \frac{1}{n} \overline{X} \overline{X}^T$$

(6)

Fuzzy C-means clustering (FCM) is an algorithm that optimizes the objective function (7) [16]:

$$J_m(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (u_{ij})^m \| x_i - v_j \|^2$$

(7)

Where $X = \{x_1, x_2, x_3, \ldots, x_n\} \subseteq \mathbb{R}^p$ is the sample set.

$x_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip}\}$ each sample includes $p$ features.

$n$ is the number of samples in $X$.

c ($2 \leq c \leq n$) is the number of clusters.

$U = (u_{ij})_{n \times c}$ is a fuzzy $c$ partition matrix of a sample set $X$.

$u_{ij}$ represents the relative degree (membership value) between the $i$th sample and the $j$th cluster.

$V = \{v_1, v_2, \ldots, v_c\}$ is the set of clustering centre (centroid).

$v_j \in \mathbb{R}^p$ represents the clustering vector of the $j$th cluster.

$\cdot \| \cdot$ is the standard Euclidean distance between the $i$th sample $x_i$ and the clustering centre of the $j$th cluster $v_j$. However, $\| \cdot \|$ can be defined by different distance formula according to one actual needs.

$m$ is the fuzzy weight exponent or the smoothing parameter which determines the degree of the fuzzy partition matrix.

Equation (7) yields the equations for cluster centres $v_j$ and elements of the fuzzy $c$ partition matrix $u_{ij}$

$$v_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m}, \text{ for } j = 1, \ldots, c$$

(8)

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{\| x_i - v_j \|^2}{\| x_i - v_k \|^2} \right)^{\frac{1}{m-1}}}$$

(9)
2.2 Implementation
The implementation of this study consists of four main parts, namely input image, image pre-processing, image segmentation, and staging based on the area of the primary tumour as the representation of size in the TNM system. The process is summarised by the flow chart in figure 1.

![Flow chart of the methodology.](image)

The input are thirty-five randomly selected mammography images with malignant tumour acquired from the National Cancer Society of Malaysia. Images are cropped at the region of interest (ROI) at the size of 200 by 200 pixels, in order to trim down the volume of the dark background since it was not important in the segmentation process [17], inevitably reducing the computational time. The ROI images are pre-processed using Median filter followed by PCA. An important property of PCA is that it can distinguish between the input data and noise in the PCA domain [15], hence might be a suitable noise removal method. The next phase is the segmentation process using Fuzzy C-means (FCM) clustering algorithm as we have proposed in [18]. In the segmentation process, images are clustered into two clusters, namely the foreground (the primary tumour) and the background (the surrounding tissues). Finally, area of the primary tumour is calculated and staging is determined based on the scale proposed by [14] as shown in table 1.

| Stage | Lower limit (pixel) | Upper limit (pixel) |
|-------|---------------------|---------------------|
| I     | 0001                | 2800                |
| II    | 2800                | 3100                |
| III   | >3100               |                     |

The procedures of this study are summarised in figure 2.

![Procedures of this study.](image)

Area = 9097; Stage III
3. Result and Discussion
From the random selection of thirty-five sample of mammography images of Malaysian women with malignant tumour showed that five patients or 14.3% are in stage I, two patients or 5.7% are in stage II and 28 patients or 80% are in stage III. The pie chart in figure 3 summaries the results. These results are consistent with the findings by [19] who stated that the incidence of breast cancer in Asia is rising and is associated with increased mortality. Thus, high percentage of stage III patients contributes to increase in death rate. On the other hand, although the incidence is also increasing in the western countries, the mortality rate is decreasing [19].

![Figure 3. Staging of thirty-five patients.](image)

Experimental results shown in table 2 indicate that higher percentage accuracy, 96.4% is obtained for classification of stage III breast cancer, and this may be due to bigger pixel area in the third category. Also, since the number of mammography images in stage I and II is small, the staging errors in both categories are bigger. Error occurs for images with primary tumour area near the boundaries of the staging scale, which are 2800 and 3100. A more suitable staging classification using fuzzy method may be able to overcome the problem of staging for tumour areas near the boundaries.

| Stage | Test image | Correct staging | Accuracy  |
|-------|------------|-----------------|-----------|
| I     | 5          | 3               | 60%       |
| II    | 2          | 1               | 50%       |
| III   | 28         | 27              | 96.4%     |

4. Conclusion
In conclusion, staging information from a randomly selected mammography images of Malaysian women shows similar trend of higher incidence and mortality rates of breast cancer among Asian women. Also, the proposed series of algorithms using median filter, PCA and FCM was able to correctly stage images with high accuracy at stage III. Suitable method(s) to evaluate tumour areas near the boundaries such as Fuzzy classification using Z-numbers may be able to improve the staging result.

Acknowledgments
The authors would like to acknowledge Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia for the supports and contributions.
References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel R L, Torre L A and Jemal A 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), 394-424.

[2] Salleh S, Mahmud R, Rahman H and Yasiran S S 2017. Speed up Robust Features (SURF) with Principal Component Analysis-Support Vector Machine (PCA-SVM) for benign and malignant classifications. Journal of Fundamental and Applied Sciences, 9(5S), 624-643.

[3] Malek A A, Rahman W E Z W A, Haris M H M and Jalil U M A 2017. Segmenting masses in ultrasound images by using seed based region growing and mathematical morphology. Advanced Science Letters, 23(11), 11512-11516.

[4] Yasiran S S, Salleh S and Mahmud R 2016, June. Haralick texture and invariant moments features for breast cancer classification. In AIP Conference Proceedings, 1750(1), AIP Publishing, 020022.

[5] Bird R E, Wallace T W and Yankaskas B C 1992. Analysis of cancers missed at screening mammography. Radiology, 184(3), 613-617.

[6] Zaman N A K, Rahman W E Z W A, Jumaat A K and Yasiran S S 2015, May. Classification of breast abnormalities using artificial neural network. In AIP Conference Proceedings, 1660(1), AIP Publishing, 050038.

[7] Muntaz H, Hall-Craggs M A, Davidson T, Walmsley K, Thurell W, Kissin M W and Taylor I 1997. Staging of symptomatic primary breast cancer with MR imaging. AJR. American journal of roentgenology, 169(2), 417-424.

[8] Embong R, Aziz N N A, Karim, A A and Ibrahim M R 2017, September. Colour application on mammography image segmentation. Journal of Physics: Conference Series, 890(1), IOP Publishing, 012066.

[9] Colella J A, Llinàs M P and Colella C A 2010. Staging rectal cancer. Radiología (English Edition), 52(1), 18-29.

[10] Förmvik D, Zackrison S, Ljungberg O, Svahn T, Timberg P, Tingberg A and Andersson I 2010. Breast tomosynthesis: accuracy of tumor measurement compared with digital mammography and ultrasonography. Acta radiologica, 51(3), 240-247.

[11] Dromain, C., Boyer, B., Ferre, R., Canale, S., Delaloge, S., & Balleyguier, C. (2013). Computed-aided diagnosis (CAD) in the detection of breast cancer. European journal of radiology, 82(3), 417-423.

[12] Singletary S E and Connolly J L 2006. Breast cancer staging: working with the sixth edition of the AJCC Cancer Staging Manual. CA: a cancer journal for clinicians, 56(1), 37-47.

[13] Karmilasari S W, Hermita M, Agustiyani N P, Hanum Y and Lussiana E T P 2014. Sample K-means clustering method for determining the stage of breast cancer malignancy based on breast size on mammogram image basis. IIACSA) Int J Adv Comput Sci Appl, 5(3), 86-90.

[14] Kulikarni S and Shreedhara K S 2015. Stage Determination of Cancer in Mammogram Image using SOFT CLUSTERING and ANN. LICSME, 4(5), May, 127 – 134.

[15] Zhang L, Dong W, Zhang D and Shi G 2010. Two-stage image denoising by principal component analysis with local pixel grouping. Pattern recognition, 43(4), 1531-1549.

[16] Bezdek J C 1980 A convergence theorem for the fuzzy ISODATA clustering algorithms IEEE Trans on Pattern Anal. Machine Intell. PAMI-2, 1-8.

[17] Dehghani S and Dezfooli M A 2011. A Method For Improve Preprocessing Images Mammography. International Journal of Information and Education Technology, 1(1), 90–93.

[18] Embong R, Rahman W E Z W A and Ahmad T 2010, The Relation of Fuzzy Weight Exponent and Cluster Centre in Fuzzy C-means Clustering Algorithm for Mammography Images, In Fuzzy: From theory to applications, Eds. Daud Mohamad et al., University Publication Centre (UPENA) UiTM, ISBN: 978-967-305-490-9, 169 – 174.
[19] Leong S P, Shen Z Z, Liu T J, Agarwal G, Tajima T, Paik N. S., ... and Foulkes W D 2010. Is breast cancer the same disease in Asian and Western countries?. *World journal of surgery*, 34(10), 2308-2324.