# Tubule Formation Segmentation of Histopathological Image of Breast Cancer by Using Clustering Method

Hadiyyatan Waasilah[1], Riries Rulaningtyas[1, a)], Winarno[1], Anny Setijo Rahaju[2]

[1]*Department of Physics, Faculty of Science and Technology, Universitas Airlangga*
[2]*Department of Pathology, Faculty of medical, Universitas Airlangga*

[a)]Corresponding author: riries-r@fst.unair.ac.id

**Abstract.** Histopathological assessment is one of the examinations that allows the classification of breast cancer based on its level. Histopathological assessment factors are based on tubule formation, nuclear pleomorphism, and the mitotic count. This study only focused on tubule formation. The tubule formation was represented by a lumen surrounded a nucleus. The segmentation of tubule histopathology of breast cancer method was using a combination of k-means clustering and graph cut. The image data used in this study were 15 images of breast cancer histopathology preparations using 5 variations in the number of clusters (k) in the k-means clustering method. The best results of tubule formation segmentation using k = 4, with an average value of balanced accuracy was 81.08% and the most optimal balanced accuracy results was 94.34%.

## INTRODUCTION

Breast cancer (Carsinoma Mammae) is an abnormal or uncontrolled growth and development of cells in the breast tissue that can cause damage to surrounding tissues or organs to other parts of the body. Breast cancer (Carsinoma Mammae) attacks all levels of society in both developed and developing countries, including Indonesia.

In Indonesia, breast cancer has a contribution of 30% and is a type of cancer that beats cervical cancer or cervical cancer which contributes to 24% (Depkes RI, 2012). Based on patient data at dr. Soetomo Surabaya, the number of breast cancer cases in the last two years has increased by 491 cases in 2012 and 574 cases in 2013 (Dewi, 2015).

Histopathological examination allows pathologists to classify cancers based on their grade (Maqlin, 2013). The most common type of breast cancer histopathology is Invasive Ductal Carcinoma (IDC). About 80% of all breast cancers are Invasive Ductal Carcinoma (IDC) (Breastcancer, 2018).

Early diagnosis of patients with Invasive Ductal Carcinoma (IDC) has been using histopathological examination, by providing an assessment of the degree of malignancy of Invasive Ductal Carcinoma (IDC) using the Bloom and Richardson system (Susilo, 2012).

The Bloom and Richardson system used extensive histopathological features of tubule formation and level of core atifia in histopathological assessment. Elston and Ellis used the mitotic activity approach and modified the system developed by Bloom and Richardson. The system is based on tubule formation, nuclear pleomorphism, and the number of mitosis (Hyperastuty, 2017). In this study, researchers only focused on tubule formation. The tubule formation is represented by a lumen surrounded by a cell nucleus (Nguyen, et al).

Histopathology manual analysis is still the main way to identify cancer tissue based on the interpretation of anatomical pathologists. Each anatomical pathologist has a varied assessment based on calculations using the visual ability and expertise of each expert (Hyperastuty, 2017). Therefore we need a special method that can help anatomical pathologists to analyze histopathology semi-automatically to solve the problem of subjective interpretation among anatomical pathologists and improve the accuracy of disease diagnosis.

In a previous study conducted by Agoes Santika Hyperastuty in 2017 entitled "Artificial Neural Network in Determining the Histopathological Grading of Breast Cancer". This study determines the histopathological grading of breast cancer using the thresholding method, edge detection, and k-means clustering for the

segmentation stage and the results of feature extraction still have close to the same value due to the less than optimal segmentation results.

In this study the researchers discussed the histopathological segmentation of tubule formations of breast cancer using a combination of the k-means clustering method and graph cut. This study is expected to provide a second opinion on the recognition of tubule formation patterns to assist pathologists and anatomists in providing prognosis from histopathological images of breast cancer.

## RESEARCH METHODOLOGY

A. Data Preparation

The data used in this study were images of breast cancer histopathology obtained from the Laboratory of Anatomical Pathology, Dr. Soetomo Surabaya. Image data taken as many as 15 images of histopathological preparations of breast cancer in the form of preparations on which there is tissue that has been processed by the anatomical pathology laboratory. The tissue comes from a breast cancer biopsy that has been colored using Hematoxylin and Eosin (H&E) substances. The coloring aims to make it easier for observers to recognize the cell structure. Then the slide analysis was carried out using a microscope equipped with a digital camera with a 40x magnification and taking images of breast cancer histopathology preparations with a size of 1920x2560 pixels stored in a file with the extension (* .Bmp).

B. Pre-processing

Image pre-processing is an improvement in image quality. In this study, in the pre-processing process, the RGB (Red, Green, Blue) color space image conversion was carried out into CIE L * a * b * and contrast stretching.

**Convert RGB to CIE L*a*b***

Image data of breast cancer histopathology preparations in the form of RGB color space images (Red, Green, Blue). RGB image is converted to CIE L * a * b * because the CIE L * a * b * color space can give better results than RGB in measuring the similarity value of color features in the image.

The RGB to CIE L * a * b * transformation begins by performing the following calculations:

$$X = 0{,}412453R + 0{,}357580G + 0{,}180423B \quad (1)$$

$$Y = 0{,}212671R + 0{,}715160G + 0{,}072169B \quad (2)$$

$$Z = 0{,}019334R + 0{,}119193G + 0{,}950227B \quad (3)$$

Next, L*a*b* are defined as

$$L^* = 116f\left(\frac{Y}{Y_n}\right) - 16 \quad (4)$$

$$a^* = 500\left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right] \quad (5)$$

$$b^* = 200\left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right] \quad (6)$$

In this case, f (q) is calculated as follows:

$$f(q) = \begin{cases} q^{\frac{1}{3}}, & jika\ q > 0{,}008856 \\ 7{,}787q + 16/116, & for\ others \end{cases} \quad (7)$$

$X_n$ =0,9642, $Y_n$ = 1, dan $Z_n$ = 0,8249 are the reference of white. (Anonim, 2006).

**Contrast Stretching**

Contrast stretching is a point operation on the original image, which means that the contrast stretching process only depends on the intensity value (gray level) of one pixel and does not depend on other pixels around it (Wakhidah in Asana, 2017). Contrast stretching is used to increase the contrast of the image so that it gets a clearer image.

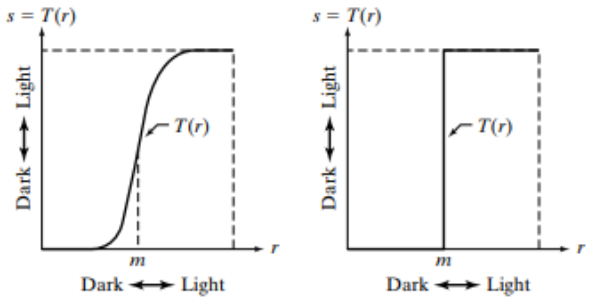



**FIGURE 1**. Graph of Contrast Stretching Transformation and Threshold Transformation. (a) Contrast Stretching Transformation, (b) Threshold Transformation (Source: Gonzalez)

Figure 1 (a) is a contrast stretching transformation function because it compresses the input level lower than m to the dark level limit in the image output. The initial equation is as follows (Gonzalez):

$$s = T(r) = \frac{1}{1+(m/r)^E} \quad (8)$$

where r represents the intensity of the image input, s as the corresponding intensity value in the image output and E as the slope control function. The equation implemented in Matlab to make contrast stretching transformations is as follows:

$$g = 1./(1+(m./(double(f)+eps)).\text{^}E) \quad (9)$$

where E is the slope control function and m is the boundary line to change from dark value to light value. eps is a matlab constant which is the distance between 1.0 and the next largest number that can be represented in double precision floating point.

C. Segmentation

Image segmentation is a process used to separate objects from one another in an image. The image segmentation process carried out in this study uses the k-means clustering method, the opening-closing morphology operation, and graph cut.

**K-means *Clustering***

K-means clustering is a non-hierarchical data clustering method that partitions data into one or more clusters / groups, so that data with the same characteristics and data with different characteristics are grouped into other groups (Agusta, 2007).

K-means clustering algorithm (Noor, 2009):

1. Specify k as the number of clusters desired.
2. Generating a random initial k centroid (cluster center point).
3. Calculate the distance of each data to each centroid.
4. Each data selects the nearest centroids.
5. Determine the position of the new centroids by calculating the average value of the data selected on same centroid.
6. Return to step 3 if the position of the new centroids with the old centroids is not the same.

The calculation of the distance between the centroid and the data can be done using the euclidean distance equation, the equation is as follows (Ong, 2013):

$$D(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2} \quad (10)$$

where D (i, j) is the distance of the data to i to the center of cluster j, $x_{ki}$ is data to I in the data attribute to k, and $x_{kj}$ is the center point to j in the attribute to k.

**Morphological Operation (Opening-Closing)**

From the results of k-means clustering, morphological operations (opening-closing) were performed. The morphology (opening-closing) operation uses a combination of sequential dilation and erosion operations. The opening-closing process begins with an erosion process by smoothing the lines of the object's shape and eliminating objects that are smaller than the specified structural element (SE). Then the process of dilation is to smooth the lines of shape and fill or eliminate small holes between two objects that are close together. And then the image dilation process that has been eroded. After that the results of the morphological

operation (opening-closing) are given a line / contour based on the structure of the cell nucleus and the lumen so that you can see the pattern of the tubule formation.

**Graph Cut**

Graph is a set G = (V, E), where V is a non-empty set of vertices or nodes (v1, v2, v3, ..., vn) and E is a set of edges connecting a pair node {e1, e2, e, 3, ..., en} (Syaifudin, 2017).
The algorithm for graph separation is as follows:
1. The intensity value between object nodes and not objects is made as a node.
2. Set the threshold value. The threshold value is taken from the average intensity value of the experimental object.
3. Each node is then compared to the threshold value.
4. If the node value is greater than the threshold, the node connection is terminated. If it's smaller, then is maintained. So that only remaining pixels with intensity are objects of the tubule formation.

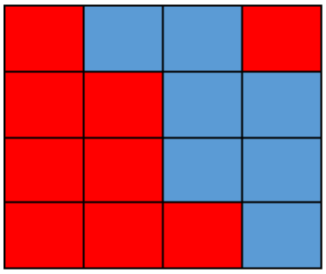


**FIGURE 2**. Cutout of a digital image, to clarify pixels (Source: Perkasa, 2016)

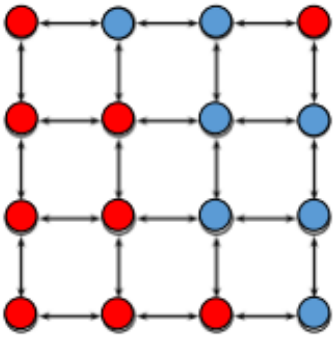


**FIGURE 3.** The shape of the pixel model (Source: Mighty, 2016)

After obtaining the tubule formation pattern, the object that is declared as tubule formation is cut using the image segmented graph cut tool provided by MATLAB. To get the results of the tubule formation, the first step is drawing objects on the foreground and background to initialize the node and giving lines / contour to the results of morphological operations based on the results of cutting the object using graph cut. From the results of object segmentation using graph cut, a tubule formation pattern was obtained in the histopathological image of breast cancer which was then stored on a disk.

D. Segmentation Result Testing

After segmenting using the program, the results of the program segmentation are tested using Matlab software and the results of manual image segmentation marked by doctors using Adobe Photoshop CS6 software using the qualitative balanced accuracy test. Balanced accuracy is a method used to determine the level of conformity of the results of program segmentation with the results of segmentation conducted by experts (ground truth) on the same image. The equation of this method is as follows:

$$Balanced\ accuracy = \frac{\left(\left(\frac{TP}{TP+FN}\right)+\left(\frac{TN}{TN+FN}\right)\right)}{2} \quad (11)$$

Information:
• TP value is obtained, if the test image pixel value = 1 and ground truth image = 1.

• FP value is obtained, if the test image pixel value = 1 and ground truth image = 0.
• The FN value is obtained, if the pixel value of the test image = 0 and the ground truth image = 1.
• The TN value is obtained, if the pixel value of the test image = 0 and the ground truth image = 0.

## RESULT AND DISCUSSION

The research data used in this study has a different image quality, depending on the quality of hematoxylin and eosin (H&E) staining, cell structure, and differences in contrast of each image. This can affect the results of segmentation of the histopathological tubule formation of breast cancer. For that we need an image pre-processing process to improve the quality of each image

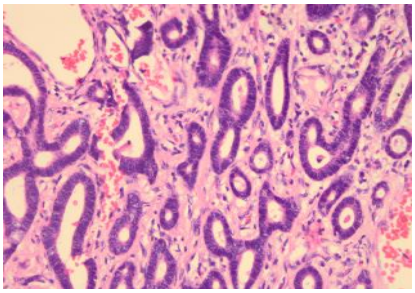


**FIGURE 5.** Original Image of Breast Cancer Histopathology

Image pre-processing is an improvement in image quality. In this study, the first step taken in the preprocessing process is converting the RGB (Red, Green, Blue) color space image to CIE L * a * b * because CIE L * a * b * can give better results when RGB is used in measuring similarity value of color features in the image.

**FIGURE 6**.Regular (Red, Green, Blue) color space image conversion results to CIE L * a * b *

The image that has been converted from the RGB color space to CIE L * a * b * is then repaired using contrast stretching. Contrast stretching is used to increase the contrast of the processed image so that it gets a clearer image.

**FIGURE 7.** contrast stretching image result

Image segmentation is a process used to separate objects from one another in an image. The first step in the image segmentation process carried out in this study is to use the k-means clustering method. In this k-means clustering method, the CIE_a * and CIE_b * color components are used as input in the k-means clustering algorithm. Selection of the CIE_a * and CIE_b * color components is used to get a clear image and no noise. The segmentation process is carried out based on the euclidean distance which classifies pixels

to form color clusters according to the number of clusters (k) specified. The objects obtained from the k-means clustering method are the cell nucleus and the lumen.

(a) (b)

**FIGURE 8**. K-means clustering results (a) Nucleus Image and (b) Lumen Image

From the results of the k-means clustering segmentation, morphological operations (opening-closing) were performed to remove unnecessary structures in the cell nucleus and lumen and make the object area smoother. In the morphological process (opening-closing) the color component used is the CIE_b * color component, because the CIE_b * color component is the best component.

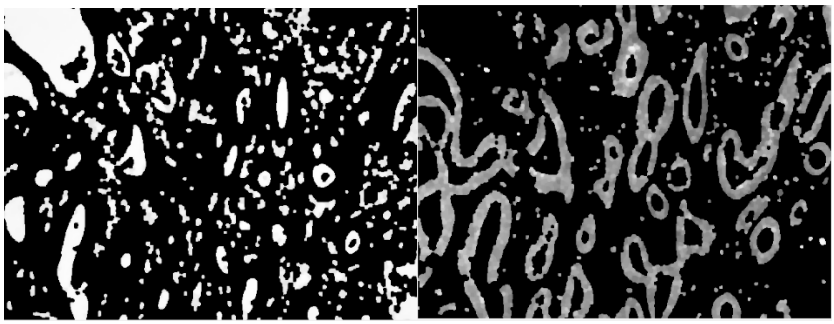


**FIGURE 9.** Morphological operation result (*Opening-Closing*)

In this study, the image of the cell nucleus and lumen was applied to morphological operations (opening-closing) to remove structures that did not include the cell nucleus and lumen in order to obtain better cell nucleus and lumen results. The results of the morphological operation (opening-closing) of the cell nucleus and the lumen are then combined and given a line / contour based on the structure of the cell nucleus and the lumen in order to obtain a pattern of tubule formation, namely the lumen surrounded by the cell nucleus.

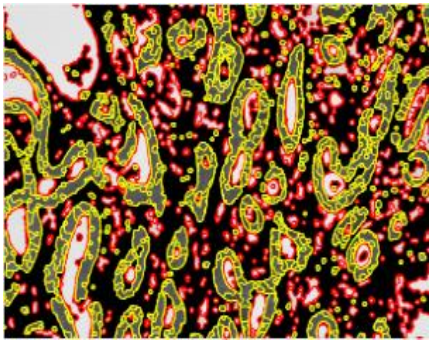


**FIGURE 10**. Result of Cell Nucleus and Lumen Patterns, yellow Contour (Cell Nucleus) and red (Lumen)

After obtaining the tubule formation pattern, the object is cut into the tubule formation using the tools provided by MATLAB, namely the graph cut image segmented. The graph cut in this study can segment tubule formations based on the results of the morphological operation (opening-closing).

**FIGURE 11.** Graph Cut result

From the results of cutting the object using graph cut, to determine the tubule formation pattern, a line / countour is applied to the results of the morphological operation based on the results of cutting the object using graph cut.

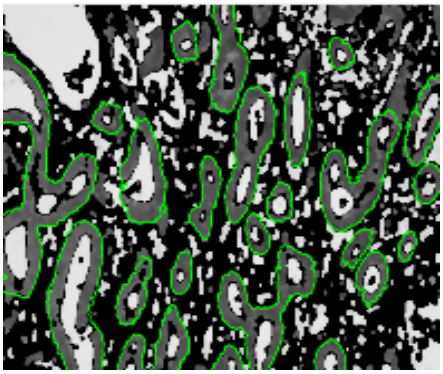
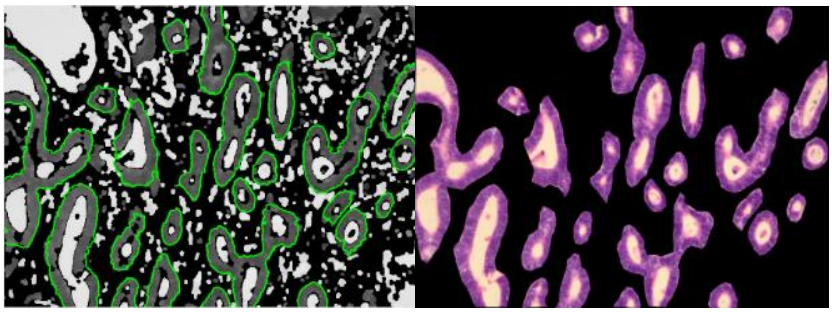


**FIGURE 12**. The result of Tubule Formation Pattern

The results of segmentation of tubule formations using the program are then compared with the results of segmentation of tubule formations manually based on the image marked by the doctor using balanced accuracy. The image of the tubule formation segmentation using the program is used as a test image and the image of the tubule formation segmentation manually is used as a ground truth image. The results of balanced accuracy can be seen in Table 4.1.



(a) (b)

**FIGURE 8**. Tubule Formation Segmentation Results (a) Program and (b) Manual

**Table 4**.1 Balanced Accuracy Results

| Image name | Balanced Accuracy Results | | | | |
|---|---|---|---|---|---|
| | k=2 | k=3 | k=4 | k=5 | k=6 |
| G1 | 63,65% | 72,40% | 83,69% | 83,00% | 69,88% |
| G2 | 76,37% | 78,45% | 79,07% | 76,91% | 59,28% |
| G3 | 68,80% | 81,91% | 72,96% | 80,33% | 59,37% |
| G4 | 90,40% | 93,53% | 94,05% | 91,49% | 92,37% |
| G5 | 63,20% | 64,42% | 73,02% | 62,58% | 52,32% |
| G6 | 73,78% | 80,87% | 89,65% | 89,03% | 89,94% |
| G7 | 60,99% | 93,59% | 90,88% | 93,03% | 88,69% |
| G8 | 50,50% | 71,58% | 74,12% | 70,56% | 71,96% |

| | | | | | |
|---|---|---|---|---|---|
| G9 | 67,10% | 77,93% | 88,99% | 77,81% | 76,20% |
| G10 | 49,75% | 60,21% | 62,13% | 76,65% | 49,07% |
| G11 | 58,20% | 77,89% | 89,22% | 61,95% | 70,57% |
| G12 | 63,16% | 70,60% | 70,84% | 80,20% | 48,86% |
| G13 | 80,73% | 83,52% | 77,79% | 76,15% | 77,77% |
| G14 | 49,53% | 55,62% | 88,80% | 59,60% | 84,55% |
| G15 | 49,47% | 77,50% | 81,01% | 86,75% | 48,88% |
| **averages** | 64,38% | 76,00% | 81,08% | 77,74% | 69,31% |

(Note: Yellow (Highest average result) and Green (Most optimal result))

From the table above it can be seen that the average result of balanced accuracy k = 2 is 64.38%, the average result of balanced accuracy k = 3 is 76%, the average result of balanced accuracy k = 4 is 81.08%, the average result of balanced accuracy k = 5 is 77.74%, and the average result of balanced accuracy k = 6 is 69.31%. From these results it can be seen that the highest average value is at k = 4. Based on the results of data analysis, it can be seen that the percentage of conformity of segmentation results using a program with manual segmentation is based on the image marked by the doctor. The segmentation program is said to be good if the resulting balanced accuracy value is high or has a small error rate.

## CONCLUSION (FIRST LEVEL HEADING)

The highest average balanced accuracy result in the histopathological tubule formation segmentation of breast cancer using the clustering method was 81.08%, and the optimal balanced accuracy result was 94.05%.

## ACKNOWLEDGMENTS

Thank you for the Computational Physics Laboratory, Faculty of Science and Technology, Airlangga University and the Pathology and Anatomy Laboratory of Dr. Soetomo Surabaya which has provided facilities and support to the author in conducting this research.

## REFERENCE

1. Agusta, Yudi. 2007. *K-Means-Penerapan, Permasalahan dan Metode Terkait*. Denpasar: Jurnal Sistem dan Informatika Vol. 3 Pebruari 2007
2. Anonim. 2006. *Image Technology Colour Management-Architecture, Profile Format, and Data Structure*. International Color Consortium ICC.1:2004-10
3. Asana, I Made Dwi Putra, dkk. 2017. *Metode Contrast Stretching untuk Perbaikan Kualitas Citra pada Proses Segmentasi Video***.** Teknologi Elektro Vol. 16, No. 02, Mei – Agustus 2017
4. Breastcancer. 2018. *Invasive Ductal Carcinoma (IDC)*. (http://www.breastcancer.org/symptoms/types/idc, diakses tanggal 1 Februari 2018)
5. Depkes RI. 2012. *Penderita Kanker Diperkirakan Menjadi Penyebab Utama Beban Ekonomi Terus Meningkat*. (http://www.depkes.go.id/article/view/ 1937/penderita-kanker diperkirakan-menjadi-penyebab-utama-beban-ekonomi-terus-meningkat.html, diakses tanggal 1 Februari 2018)
6. Dewi, Gusti Triara dan Lucia Yovita Hendrati. 2015. *Analisis Risiko Kanker Payudara Berdasar Riwayat Pemakaian Kontrasepsi Hormonal dan Usia Menarce*. Surabaya: Jurnal Berkala Epidemiologi Vol. 3 No. 1 Januari 2015
7. Gonzalez, Rafael C, dkk. *Digital Image Processing, Using MATLAB.* Prentice Hall
8. Hyperastuty, Agoes Santika. 2017. *Penentuan Grading Pemeriksaan Histopatologi Kanker Payudara Menggunakan Artificial Neural Network*. Surabaya: Universitas Airlangga Tesis
9. Maqlin P., dkk.2013.*Automatic Detection in Breast Histopatological Images*.India:Journal Advances in Intelligent System and Computing Vol. 202
10. Nguyen, Kien, dkk.*Automatic Glandular and Tubule Region Segmentation in Histological Grading of Breast Cancer*.USA:Ventana Medical System
11. Noor, M. Helmy dan Moch. Hariadi.2009.*Image Clustering Berdasarkan Warna Untuk Identifikasi Kematangan Buah Tomat dengan Metode Valley Tracing*.Yogyakarta:Seminar Nasional Informatika 2009 (semnasIF 2009) UPN "Veteran" Yogyakarta 23 Mei 2009

12. Ong, Johan Oscar.2013.*Implementasi Algoritma K-means Clustering untuk Menentukan Strategi Marketing President University*.Bekasi:President University
13. Perkasa, Rio Dwi Putra. 2016. *Penerapan Teori Graf dan Graf Cut pada Teori Pemisahan Objek Citra Digital*.Makalah IF2120 Matematika Diskrit-Sem. I Tahun 2016/2017
14. Susilo, Imam. 2012. *Ekspresi Protein ER (Estrogen Receptor) pada Kanker Payudara Derajat Keganasan Baik, Sedang, dan Buruk*. Surabaya: Jurnal Ners Vol.7 No.1 April 2012
15. Syaifudin, Yan Watequlis, dkk. *Matematika Diskrit*. Polinema Press