Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 48 (2) 265-270 (2016)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Automated Detection of Malignant Cells Based on Structural Analysis and Naive Bayes Classifier

Z. JAN[++], S. U. KHAN, N. ISLAM, M. A. ANSARI*, B. BALOCH**

Department of Computer Science, Islamia College, Peshawar, Khyber Pakhtunkhwa Pakistan

Received 4th March 2015 and Revised 16rdDecember 2015

Abstract: Breast cancer is the second most common cancer in all over the world. The treatment of breast cancer is possible if the problem is properly identified. To solve this problem it needs such type of automated system that detects it in early stage. This paper presents a new method for detection of malignant cells and their classification in breast cytology images. The proposed method is divided into five phases. In the first phase, image is pre-processed for contrast enhancement followed by noise removal. During the second phase, the image is segmented into foreground and background regions. In third phase total numbers of cells are counted. Using naïve Bayes classifier all cells are classifieds into malignant and benign on the basis of size and shape features of the nucleus. The experiments were performed on local data set provided by the pathology department, Lady Reading Hospital (LRH) Peshawar, Pakistan. The results produced by the proposed technique were validated by a team of senior pathologist of the pathology department LRH. The proposed technique was evaluated individually on the dataset for malignant cells detection and was found to be effective. Experimental results show that scheme has accuracy up to 98.49% by naïve byes classifier with multiple cross validations. Results show that the scheme has improved the performance of malignant cells detection and classification. It is also capable to classify images of Fine Needle Aspiration Biopsy (FNAB) slides with high accuracy.

Keywords: Breast carcinoma, Cytopathologist, Shape features, classifier, Naïve Bayes

1. INTRODUCTION

Breast cancer is one of the most dangerous cancers detected among women in the age of 40-50 years. Precise detection and prediction are crucial to reduce this high death rate. Breast carcinoma is a main health disease not only in Pakistan but all over the world. It is a primary source of death among the female. According to the recent report of World Health Organization (WHO), 7.6 million deaths occur due to cancer every year, in which approximately 0.5 million are caused by Breast tumor only (Boyle, *et al*., 2008). While in Pakistan this ratio goes up to 0.04 million (Pakistan affairs, 2013). There are three known methods for detection of breast cancer i.e. Breast self-examination (BSE), mammography and Fine Needle Aspiration biopsy (FNAB). BSE is basic and primary method used for breast cancer detection. In this method no hardware tools are required; it can also be conducted by women itself at their homes. It might be done regularly once a month after menstrual cycle. Some time breast tissues are affected but not felt, therefore it can be identified by mammography (Breast X-Ray). By mammography it is very difficult to distinguish between benign (not cancerous. It may be increase is in size but not spreading into surrounding area of the body) and malignant (Affect those tissues laying in the nearest region and also spreading into other parts of the body).

FNAB is recommended as the best approach (Martin, *et al.,* 1934). FNAB was first introduced in 1930. It is a progressive method, in which a small sample is extracted from doubtful breast tissue through which pathologists try to detect the cancer and its types. The abnormal breast cytology has a nucleus size larger and an irregular shape than that for normal breast cytology, which is a symptom of abnormality. The fine needle aspiration cytology (FNAC) has been considered as a good method for the pre-operative diagnosis as compared to self examined method and mammography. In this paper, we are going to present a novel approach for segmentation, counting and classification of cells into malignant and benign classes.

(AbderrahimSebri, *et al*., 2009) selected active contour used for segmentation and textural features were extracted by Wavelet transforms. Finally for classification MLP was used. (IssacNiwas, *et al*., 2010) using complex wavelets for features extraction and for classification of malignant and benign cell the k-nearest neighbor classifier was used. Which gives up to 93.33% classified successes (on average).(Amir Fallahi and Shahram Jafari., 2011) Developed an automated system for breast cancer using Bayesian network s a classifier. Its gives 98.1% accuracy results with cross validation fold 3.(MarekKowal, *et al*.,2011) selected adaptive threshold used for segmentation and Gaussian mixture clustering for nuclei features extraction. Classification is done by, k-nearest neighbors, decision trees and naive Bayes. Combined average accuracy result was recorded approximately 98%.(Pawel *et al*.,

[++]Corresponding author, email: Zahoor.jan@icp.edu.pk
* University of Sindh, Jamshoro
** Sindh Agriculture, University, Tando Jam

2013)Proposed system for diagnosing breast cancer based on microscopic cytology images. Circular Hough transform was used for segmentation to detect the circular nuclei shape.SVM was used for classification purpose. Total 737 images were tested by algorithm, which gives 98.51% effectiveness.(Lukasz, *et al*.,2008) developed a system for detection of malignant cell and their grading using microscopic images. The classification was done by Support Vector Machines (SVM) and gives accuracy up to 94.24%.

2. PROPOSED METHOD

The proposed method consists of five phases. The first phase is pre-processing, canting sub five phases. Second is the segmentation phases, canting sub seven phases. The third phase is the counting of total numbers of cells in breast cytology images. In fourth phase the required features are extracted which are then used to classify cells as benign or malignant. **(Fig.1)** shows the phases and all of its parts.

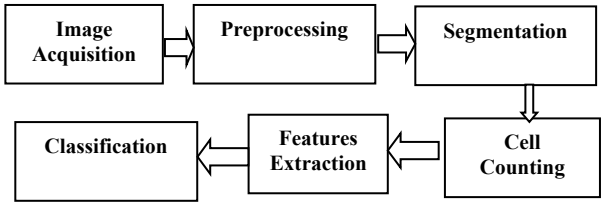


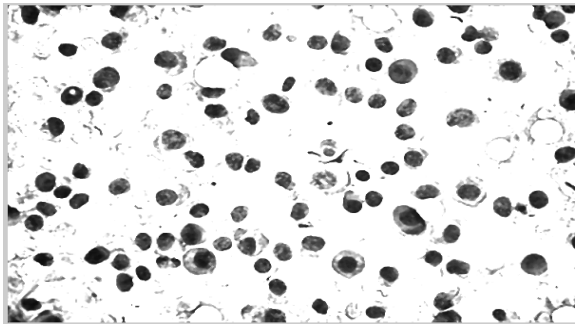
Fig 1: block diagram of Proposed Method

2.1. Image Acquisition

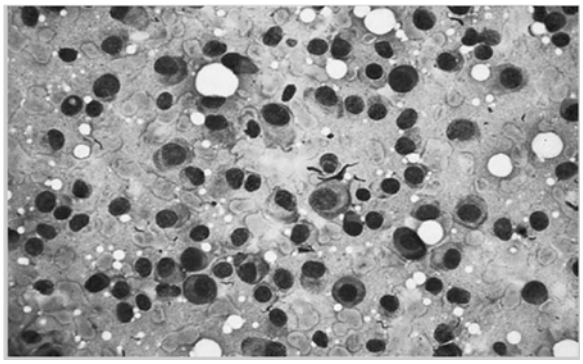
A total 40 numbers of Images were provided by the pathology department LRH, Peshawar for experimental purpose after testing by pathologist and classified with benign and malignant. Captured by a special type of microscope called Olympus BX-51 from breast cytology slides.

2.2. Preprocessing

This is very important step in image processing. In this phase the RGB colored image captured by microscope have been processed by different image processing techniques for enhancement image quality, noise removing and smoothness. Through this step image become clearer for further processing. In the proposed method image adjustment, color space separation, CLAHE and median filter are used.



(a) Original Image



(b) After Preprocessin

Fig.2: (a) Original Image (b) After Preprocessin

2.3. Segmentation

For the determination of benign and malignant cells it needed to separate cell nuclei from background objects and from all others unnecessary objects (blood lesion). Morphological operations were used with the help of 'structuring elements' for smoothing the new area of the image. Otsu's Binarization was used for Conversion of gray scale image into bi-tonal image i-e black and white.

$$\sigma_w^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t) \quad (1)$$

ωi weights demonstrate the two classes probabilities divided by a threshold value t and σ_i^2 variances of these classes. With the help of Complement pixels were converted from white into black and vice versa. Mathematically it can be written as:

$$\left[T_{[t_i,t_j]}(f)\right](x) = \begin{cases} 1\ if t_i \le f(x) \le t_j \\ 0 \quad otherwise \end{cases} \quad (2)$$

To remove extra pixels and shrink cell objects with the help of erosion. The mathematical representation of erosion:

$$I_{errod} = A \Theta B = \{B_w \subseteq A\} \quad (3)$$

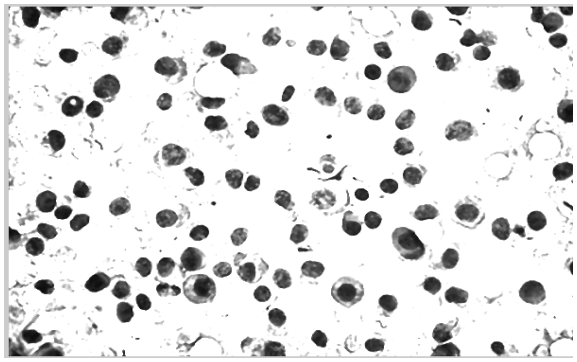
In proposed technique closing used after erosion to reconnect the necessary pixels which are part of cell objects.

$$I_{Closed} = A \bullet B = (A \oplus B) \Theta B \quad (4)$$

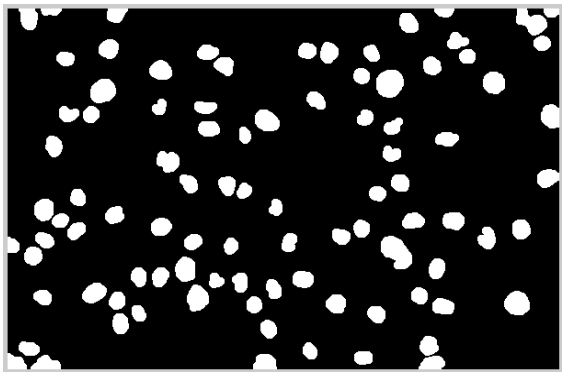
Now to remove all those objects which were not part of cells, which become remain from erosion. In the proposed technique we set the threshold value p=30 pixels, to remove all those objects, whose values is less than 30.

$$I_{Area}\ \text{Filter} = AF = AF\left(\sigma_w^2(t), P\right) \quad (5)$$

After using an area filter if some cells pixels were removed of connected objects, then it will be recovered by dilation process. Finally using of filling holes techniques if there are some spots of holes remain in the cells nuclei then it will be filled.



(a) Before Segmentation

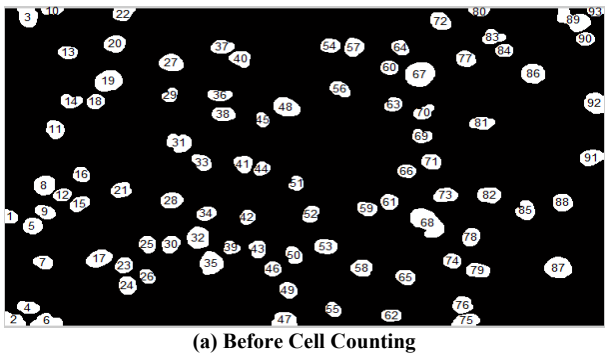


(b) After Segmentation

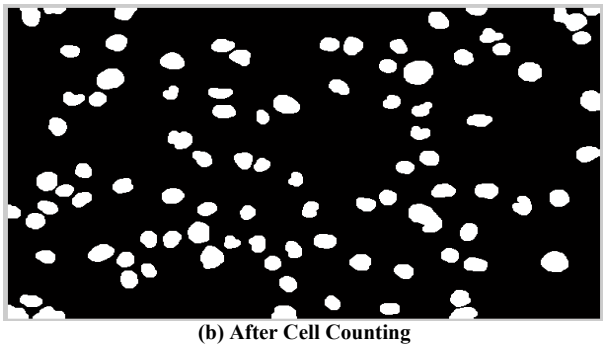
Fig 3: (a) Before Segmentation (b) After Segmentation

2.4. Cell Counting

After segmentation process it is necessary to count the total numbers of cell nuclei in the image, because counting is also a very tedious and difficult task for pathologist to count it manually. The proposed technique counting the total numbers of cells nuclei in breast cytology image, and labeled it with numerical values. By using labeling function, in this technique first is taken in binary image and then converting it into label matrix. The values of matrix elements are in the form of numbers, where 0 represent background and 1 represent object one, 2 represent a second objects and so on using four connected pixels mask in binary image. Scanning the binary image in column wise and finds the object in the lower-left corner and label it.



(a) Before Cell Counting



(b) After Cell Counting

Fig 4: (a) Before Cell Counting (b) AfterCell Counting

2.5. Features Extraction

To efficiently classify cells in cytology image require extraction of strong features. Typically, the image is segmented to isolate different objects from one another as well as from the background and then label .During this step, features like size (area) and shape (compactness) were extracted from each individual cells. According to medical experts four features are used for malignant cell detection (Size, shape, color and ratio between nucleus and cytoplasm).But we were interested in early stage detection. Therefore we have extracted the size and shape of nucleus because initially size and shape of nucleus are affected.

2.2.1. Size

In the proposed technique the size of nucleus in each individual cell is determined and then on the basis of this size the benign and malignant cells are distinguished. Two features regarding size are:

Area Sum of all pixels inside the nucleus region is called an area. It can be calculated by the following formula:

$$\text{Area} = \sum_{i=1}^{n} xi \qquad (6)$$

Then take the average area of all the cells nuclei. The average area is calculated by the following formula:

$$Average = \frac{1}{n}\sum_{i=1}^{n} xi \qquad (7)$$

Where '*I*' mean the set of cells from 1 to n. The sum of the nucleus area are calculated and then divided on the total numbers of cells which produce the average value of cell nuclei area.For reducing the error, we multiply an arbitrary value with this average area as an error factor, and then compare the area of individual nucleus with this value. When the value of single value is greater than this value area than it will be considered as a malignant cell.

Perimeter: Perimeter is the total numbers of the nucleus boundary (*Bi*) pixels. Mathematically it can be written as:

$$P_i = \sum_{(m,n)\epsilon Bi} Bi(m,n). \quad (8)$$

2.2.2. Shape

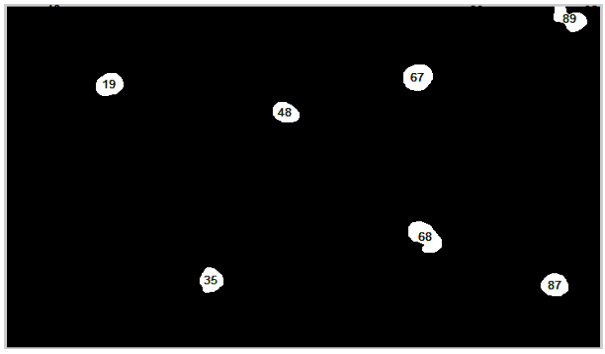
by Shape features, proposed method can find out the shape of nucleus of both normal and abnormal. Therefore, in this work author used the most basic and important shape features.

Compactness

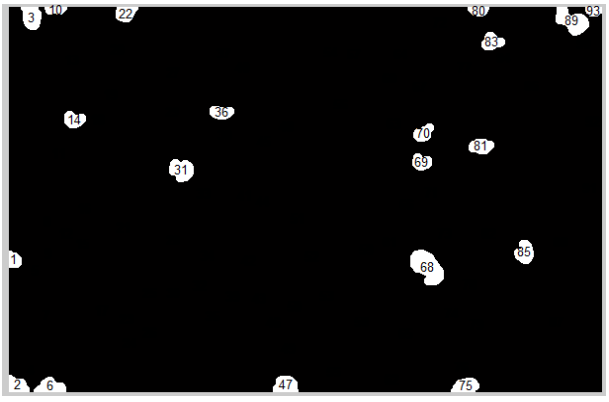
The compactness calculates the roundness of the cell nuclei. The normal cells nuclei are mostly circular in shape. When the circularity of the nucleus changes then, it means that this cell is affected, referred to as malignant cells. Compactness is very important feature for roundness measurement. In this work the compactness feature is used. Mathematically it can be written as:

$$C = P^2/A \quad (9)$$

Where 'P' is the perimeter of cell nuclei and 'A' shows the total area of the cell nuclei. The proposed system computes the compactness of each individual cells nucleus and then takes their mean multiplying with error factor F. For malignant cells detection the compactness result of each individual cells nuclei is compared with this value. If the compactness of single nuclei is greater than the calculated value than these cells should be considered as a malignant cells.



(a). Effect Cells on the basis of Area =7



(b)Effected cells on the basis of compactness=20

Fig.5: (a). Effect Cells on the basis of Area =7

3. CLASSIFICATION

In this phase, the author is describing the classification of the breast cells into two separate classes malignant and benign Using the known and specific classifier and describing their result Pattern classification is used for the prejudice between classes of patterns. The process of prejudice is not sure to implement for all patterns. It is also possible that may be some of classification may misclassify the input patterns and some classify the patterns hundred percent. This difference leads us toward the errors which are measured by classifier. The performance of the classifier is considered to be better, when it produce less numbers of errors. When the accuracy is closer to the 100% then it means that the performance of the classifier is better. Classification of the FNAC images collecting the features vector as input for classifier and its produce two-output vector for benign and malignant. The Naïve Byes classifier is used for classification.

4. RESULT AND DISCUSSION

In this section the detail explanation of images data sets, result of proposed technique compared with pathologist are discussed step by step.

4.1. Dataset

The main and important contribution of the proposed method is using of local dataset, provided by the department of pathology Lady Reading Hospital Peshawar. All the images have been tested by pathologist manually and compared with proposed technique. All the images have been taken by a special type of microscope Olympus BX-51 and stored in computer in jpeg format.

4.2. Experimental Result

The proposed algorithm were tested on a dataset consist of 40 images of breast cytology, which were already tested by pathologist manually. The Compression results of both proposed technique and pathologist are shows in table1. The results of extracted features were stored with ***".csv"***file detected by algorithm. The proposed algorithm totally based on nucleus not on cytoplasm.

Table1 shows the comparison result of images tested by pathologist as well as proposed system. Total 40 images were tested of datasets. Colum first show numbers of tested images, Colum second show total numbers of cells in tested images. After detection of malignant cells the tables also explained the gradingof the breast cancer, according to criteria discussed above.

4.3. Classification Result

After extraction of required features to check the performance of the proposed technique by the help of known classifier. In this Paper the naïve bays classifier

is used for testing the performance of proposed method with cross validation folds (CV) (2, 5, 10, and 50). The most known and binary machine learning algorithm Naive Bayes classifier was used for performance evaluation of the proposed method.

Cross Validation This means that CV produces a reasonable assessment of test performance (i.e. when the training form with 100% is tested next to another hidden test set). including say 5 folds only means the folds are 80% trained which can be exposed to have great effect on the heftiness (train to test) of the produced copy.

Table 1.Detection of malignant cells and their comparison with pathologist result

Image No	Total Numbers of Cells	By Pathologist		By System	Proposed
1	93	65	28	68	25
2	176	141	35	137	39
3	94	66	28	64	30
4	87	64	23	58	29
5	89	59	30	61	28
6	118	82	36	86	32
7	104	76	28	74	30
8	126	90	36	85	41
9	84	56	28	58	26
10	171	119	52	116	55
11	175	130	45	133	42
12	185	145	40	143	42
13	167	134	33	132	35
14	170	125	45	127	43
15	190	151	39	142	48
16	152	127	25	126	26
17	25	22	3	24	1
18	92	65	27	62	30
19	93	78	15	73	20
20	110	81	29	74	36
21	105	78	27	80	25
22	113	72	41	70	43
23	135	82	53	84	51
24	110	80	30	78	32
25	95	74	21	69	26
26	116	80	36	84	32
27	106	73	33	76	30
28	121	76	45	80	41
29	84	62	22	58	26
30	174	117	57	119	55
31	148	121	27	122	26
32	120	77	43	70	50
33	96	66	30	66	30
34	102	79	23	82	20
35	112	82	30	76	36
36	173	134	39	131	42
37	185	150	35	143	42
38	169	126	43	134	35
39	173	143	30	130	43
40	191	136	55	143	48
Total	5129	3784	1345	3902	1391

4.4.1. Confusion Matrix

Also called as a contingency matrix is a precise table design that allows visualization of the performance of an algorithm. Each and every field of the matrix shows the instance in a calculated class, while each row represents the instances in a real class. it formulate easy to perceive if the method is confusing in two classes (i.e. benign and malignant).

True Positive (TP) For a single instance test positive for a certain condition and become positive (i.e., have to fulfill the condition).

False Positive (FP) For a single instance test positive for a certain condition and also become negative (i.e., do not have fulfill the condition).

$$\Pr ecision = \frac{TP}{(TP + FP)}$$

$$\text{Re } call = \frac{TP}{(TP + FN)}$$

On the basis of these facts we can easily evaluated a confusion matrix.

Table 2 Confusion Matrix

	Nucleus classified as positive	Nucleus classified as Negative
Truly Positive nucleus	TP	FN
False Positive nucleus	FP	TN

5. CONCLUSSION

In this paper, we proposed a method for detection of malignant cells and classification in microscopic images of humanbreast. In pre-processing Phase different image processing techniques are used to enhance the image for segmentation phase with high quality. The image is segmented into foreground and background objects. To extract and differentiate between malignant and nonmalignant cells correctly, it is important to extract the relevant features of malignancy. Therefore we are using the simplest and important features size and shape, which are mainly used by medical experts. After extracting the features of malignancy cells can be classified into two classes with the help of naïve bayes classifier and produce 98.49% accuracy. The results encourage further researchers into automatic cell image segmentation especially in biomedical applications. We have tried to solve the problem up to some extent, but there is still a large room for improvement.

In future it will be trained with multi classifier for comparison purpose. we will focusing on classifying the segmented cells into both benign or malignant using multiple features and use multi lens resolution like 100x, 110x, 140x, 160x, 200x for image capturing. The problems faced in this regard are lack of devices for capturing cytology images.

REFERENCES:

Boyle, P., and B. Levin, (2008). World cancer report 2008. IARC Press, International Agency for Research On Cancer. Lyon, France, Available:http://www.cabdirect.org.

Fallahi, A., and S. Jafari, (2011). "An expert system for detection of breast cancer using data preprocessing and bayesian network." international Journal of Advanced Science and Technology, 34: 65-70.

Filipczuk, P., T. Fevens, A. Krzyzak, and R. Monczak, (2013). "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies." Medical Imaging, IEEE Transactions 32 (12): 2169-2178.

George, Y. M., B. M. Bagoury, H. H. Zayed, and M. I. Roushdy, (2013). "Automated cell nuclei segmentation for breast fine needle aspiration cytology." Signal Processing, 93(10): 2804- 2816.

Gurcan, M. N., T. Pan, H. Shimada, and J. Saltz, (2006). Image analysis for neuroblastoma classification: segmentation of cell nuclei. In Engineering in Medicine and Biology Society, 2006. EMBS 06. 28th, Annual International Conference on.4844-4847. New York, IEEE.

Hrebień, M., P. Steć, T. Nieczkowski, and A. Obuchowicz, (2008). "Segmentation of breast cancer fine needle biopsy cytological images." International Journal of Applied Mathematics and Computer Science, 18(2): 159-170

Jeleń, L., T. Fevens, and A. Krzyżak, (2008). "Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies." International Journal of Applied Mathematics and Computer Science, 18(1): 75-83.

Kowal, M., P. Filipczuk, A. Obuchowicz, and J. Korbicz, (2011). "Computer aided diagnosis of breast cancer using Gaussian mixture cytological image segmentation." Journal of Medical Informatics & Technologies, 17: 257-262

Malek, J., A. Sebri, S. Mabrouk, K. Torki, R. Tourki, (2009). Automated breast cancer diagnosis based on GVF-Snake segmentation, wavelet features extraction and fuzzy classification. Journal of Signal Processing Systems, 55(1-3):49-66.

Martin, H. E, and E. B. Ellis, (1934). Aspiration Biopsy. Surg Gynecol Obstet, 59:578-589.

Niwas, I., Palanisamy, P. and Sujathan, K. (2010 October). Wavelet based feature extraction method for breast cancer cytology images. In Industrial Electronics and Applications (ISIEA), 2010 IEEE Symposium on. 686-690. Penang, IEEE.

Pakistan affairs.[Online].http://www.pakistanaffairs.phreads 6116- Breast- cancer- kills- 40-000 women-annually- in Pakistan. [Accessd Dec,2013].

Primkhajee, P. C., P. Phukpattaranont, S. Limsiroratana, P. Boonyaphiphat, and K. Kayasut, (2010). Performance Evaluation of Automated Algorithm for Breast Cancer Cell Counting. International Journal of Computer and Electrical Engineering, 2(4):637.

Yang, X., H. Li, and X. Zhou, (2006). Nuclei Segmentation using marker-controlled watershed, tracking using mean shift, and Kalman filter in time-lapse microscopy. Circuits and Systems I: Regular Papers, IEEE Transactions on, 53(11):2405-2414.

Zhou, X., H. Li, J. Yan, and S. T. Wong, (2009). "A novel cell segmentation method and cell phase identification using Markov model." Information Technology in Biomedicine, IEEE Transactions, 13(2):152-157.