Initiative for Thyroid Cancer Diagnosis: Decision Support System for Anaplastic Thyroid Cancer

Jamil Ahmed Chandio[1], M. Abdul Rehman Soomrani[1], Attaullah Sehito[1], Shafaq Siddiqui[1]

Abstract

Due to the high level exposure of biomedical image analysis, Medical image mining has become one of the well-established research area(s) of machine learning. AI (Artificial Intelligence) techniques have been vastly used to solve the complex classification problems of thyroid cancer. Since the persistence of copycat chromatin properties and unavailability of nuclei measurement techniques, it is really problem for doctors to determine the initial phases of nuclei enlargement and to assess the early changes of chromatin distribution. For example involvement of multiple transparent overlapping of nuclei may become the cause of confusion to infer the growth pattern of nuclei variations. Un-decidable nuclei eccentric properties may become one of the leading causes for misdiagnosis in Anaplast cancers. In-order to mitigate all above stated problems this paper proposes a novel methodology so called "Decision Support System for Anaplast Thyroid Cancer" and it proposes a medical data preparation algorithm AD (Analpast_Cancers) which helps to select the appropriate features of Anaplast cancers such as (1) enlargement of nuclei, (2) persistence of irregularity in nuclei and existence of hyper chromatin. Proposed methodology comprises over four major layers, the first layer deals with the noise reduction, detection of nuclei edges and object clusters. The Second layer selects the features of object of interest such as nuclei enlargement, irregularity and hyper chromatin. The Third layer constructs the decision model to extract the hidden patterns of disease associated variables and the final layer evaluates the performance evaluation by using confusion matrix, precision and recall measures. The overall classification accuracy is measured about 97.2% with 10-k fold cross validation.

1. Introduction

Recently biomedical image inference of DICOM (Digital communication in medicine) images have been witnessed one of the active research area(s) of machine learning and AI (Artificial Intelligence) base techniques have shown significant impact upon the diagnostic process. Various CAD (computer added diagnosis) systems have been proposed to solve the classification problems of malignant diseases such as lung, breast, head & neck, lymphatic system, thyroid and other cancers. Some of the very nice approaches [1],[2],[3],[4] were proposed to solve the classification problem of cancer disease.

Infact, it is really one of the challenging field to identify the object of interest for different organs of human body as stated above because the classification of poorly differentiated, well differentiated and undifferentiated cancers have been found with lots of variations and divergent properties. For example, thyroid Anaplast cancer is one of the aggressive malignancy and its growth rate is higher than the other type of cancers.

Since the DICOM images of thyroid disease provides sufficient information to diagnose the Anaplast cancers but due to the use of poor staining material such as H and E, it may deceive doctors while examining the

---

[1] *Department of Computer Science Sukkur IBA University Sukkur, Pakistan*
Corresponding author: Jameelahmed.phdcs@iba-suk.edu.pk

mimic chromatin properties either it is hyper chromatin or not. Secondly, due to the unavailability of nuclei measurement techniques, at very early stage, irregular set of nuclei may produce point of confusion for doctors to determine the initial stage of enlargement from concerned group of nuclei. Involvement of multiple transparent overlapping nuclei may also be prone to one of the causes confusion to infer the growth pattern of nuclei grooves and early eccentric properties of nuclei are considered one of the difficult decisions because detachment of cell wall may be initial phase. In-order to solve all the above stated problems this paper proposes a system "Decision Support System for Anaplast Thyroid Cancer" to assist the doctors to diagnose the Anaplast cancer at early stage. Methodology of proposed system comprises over four layers and it proposes an algorithm for data preparation for Anaplast cancer [Algorithm 1]. In first layer noise reduction techniques have been used to reduce the noise of DICOM image by using adaptive threshold method and edges of objects have been detected by using canny edge detection algorithm whereas regional clusters have been tagged by using watershed algorithm. In second layer our proposed data preparation algorithm detects the properties and behaviors of Anaplast cancer such as enlargement of nuclei, irregularity of nuclei and hyper chromatin related features. In the third layer, random forest AI base technique has been used to construct the decision model to decide about the existence of cancerous material. In final layer performance evaluation have been conducted by observing confusion matrix, precision and recall measures. The overall performance has been shown by using AUC (area under curve). The measured accuracy of proposed system is about 97.20%. A real world dataset has been used for Anaplast cancers received from SMBBMU (Shaheed Muhtarma Benazir Bhutto Medical University) Pakistan as such datasets are unavailable in literature.

Rest of paper is organized in five sections. Introduction is presented in section one. Related works in section two. Proposed methodology is shown in section three. Results in section four and conclusion are described in section five.

2. Related Works

Basically this paper falls into the category of productive mining and deals with the classification problem of DICOM (Digital communication in medicine) images of biopsy. Specially; this paper proposes a preprocessing algorithm for Thyroid related Anaplast cancers which are most aggressive type of malignancy comprises over dissimilar growth patterns in terms of shape, size and other morphological properties, which are significantly important to diagnose at proper stage. Following related works have been seen in the recent past.

SVM (support vector machine) and AdaBoost machine learning techniques were compared for breast cancer classification problem [1]. The best classification accuracy was approximated as 87.42%. This paper presents a novel algorithm in data preparation phase and uses random forest AI base technique to construct the classification model for cancerous and non-cancerous thyroid Anaplast malignancies.

A system [2] using Convolutional neural network based machine learning technique was proposed for thyroid disease classification. The DICOM images need significant pre-processing techniques for every individual class of disease such as well-differentiated, poorly differentiated and others. All the cancer types may not be pre-processed with the same procedure because the key building blocks nuclei are considered most important micro architectural components, which usually have been found with different properties for every histo-pathological class. In data preparation phase proposed algorithm selects enlargement, irregularity and hyper chromatin related features. A comparative [3] study was conducted by using various segmentation algorithms based on clustering i.e. K-means and watershed than a supervised learning approach was used to construct the decision model obtained accuracies for template matching strategy were measured as

72% and 87%. Since the histopathology focuses on dissimilar micro structures of different human cells and tissues, whereas in a single organ various type of cancer diseases may be appeared because every disease has its own properties which needs significant efforts in data preparation phases.

On the concept of Neural Network [4] different AI base algorithms i.e. Scaled Conjugate Gradient, BFGS Quasi-Newton, Gradient Descent method and Bayesian regularization algorithms were used and best approximated accuracies for thyroid disease datasets were recorded as respectively 90.5%, 86.30% and 83.50%. Since the DICOM images of thyroid disease provide sufficient information to diagnose the Anaplast cancers, it deceives doctors when hyper chromatin may be observed with nearby properties due to the use of poor staining material and at very early stage irregular nuclei could not be seen properly due to the unavailability of nuclei measurement techniques, whereas enlargement of cells could easily be picked by the doctors but due to the involvement of multiple transparent overlapping nuclei in deceiving nature may create confused situations. It is direly needed to propose an efficient system with the support of systematic data preparation algorithm which recommends deepest decisions about the evidences of cancerous material at early stages.

3. Methodlogy

This paper proposes a system which deals with the predictive mining in the field of machine learning. The system comprises over four layers, in the first layer noise reductions have been done and in the second layer several disease related properties have been selected under the umbrella of behavior detection and feature selection. In the third layer classification models have been constructed and in the final layer performance evaluation have been done. Since the DICOM images of Anaplast cancers need significant methodology to pre-process, the nuclei behaviors of Anaplast cancers have dissimilar behaviors that are very difficult to select object of interest in proper way. For example enlargement of nuclei, irregularity in the nuclear patterns and variations in chromatin distribution are very difficult to interpret as digital set of objects. This paper also contributes a data preparation algorithm AD (Analpast_Cancers) which performs all the above stated tasks effectively and every observation is recorded carefully with the assistance of expert medical panel.

Algorithm 1: Anaplast_Diagnosis
Input: DICOM Dataset as D
Output: Enlargement, Irregularities, Hyperchromatin and Class Label Cancerous = Yes/No

$$Datset \leftarrow Anaplast_{Cells} \|\Delta f\| \leftarrow \sqrt{\left(\frac{\partial f}{dx}\right)^2 + \left(\frac{\partial f}{dy}\right)^2}$$

$Goto = every\ unit\ of\ dataset\ DasD_n\ (x_i, y_i)$
$Find\ \sigma^2 w(D) \leftarrow x1(D)\sigma_1^2 + y1(D)\sigma_1^2(D)$
$foreach P_i \in x_i(D) do$

    $Enlargement \leftarrow P \leftarrow g(x,y)^{x1(D)} \leq 0_n(P) \leq 1$
    $if\ g(x,y)^{x1(D)} = 1$
        $Heterogeneity \leftarrow g_h(\tau) = h_b(x) * g(x)$
    $Count \leftarrow \sum_{j=n}^{i=1} Size\ of\ p_i\ 1 + + \|Enlargment\|$
    $if\ g(x,y) \leftarrow Size\ of p_i = 0$
    $Count \leftarrow \sum_{j=n}^{i=1} p_i\ 0 + + \|irragularity\|$
    $if\ g(x,y) \leftarrow Size\ of p_i = p_i$
    $Count \leftarrow \sum_{j=n}^{i=1} p_i = p_i + + \|hyperchromatism\|$
    $endif$
$Return \leftarrow Enlargment, irragularity, hyperchromatism,$
    $Class\ Label \leftarrow Cancerous = Yes|No_{Cells}$
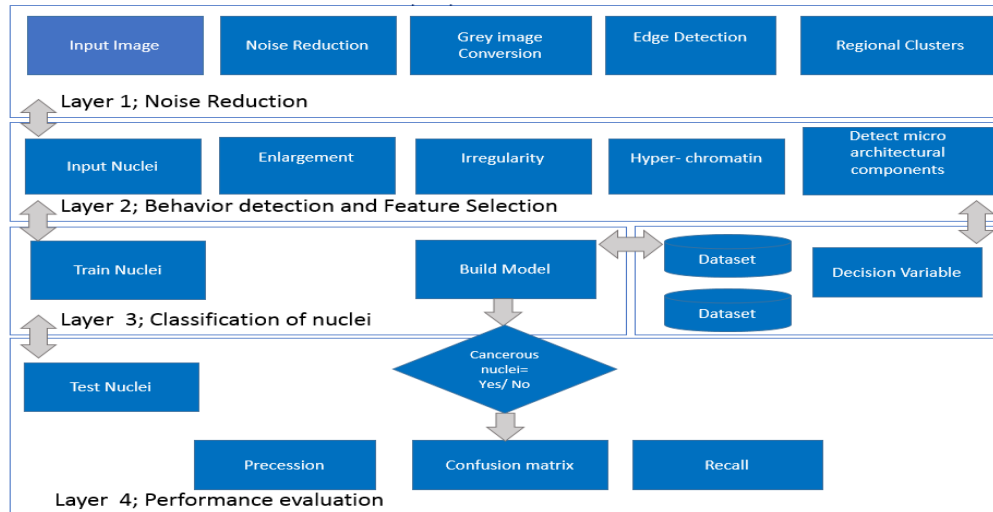
Algorithm1: AD (Anaplast Diagnosis)

Figure 1: Decision Support System for Anaplast Thyroid Cancer workflow

3.1. Layer 1: Noise Reduction

Proposed algorithm has to perform three numbers of tasks. Firstly to detect the enlargement of cells. Secondly to find out the irregularity and third the hyper chromatin. Let us consider a medical image of biopsy consisting of the several number of nuclei and each nuclei occupies size, shape, color and many features. Computationally, the algorithm reduces the noise by applying the image pixel derivation as per eq. (1).

$$\|\Delta f\| \leftarrow \sqrt{\left(\frac{\partial f}{dx}\right)^2 + \left(\frac{\partial f}{dy}\right)^2} \qquad (1)$$

Medical image $M = \{x+h_1, x + h_2, x + h_3 \dots \dots \dots \dots x + h_n\}$ is consisting upon the h spaces in vertical and $N = \{y + h_1, y + h_2, y + h_3 \dots \dots \dots \dots y + h_n\}$ horizontal vectors. If complexity of huge vectors is selected as features of large size medical images, it may require high computational power to compute the correlation of the associated pixels. Proposed algorithm reduces the unnecessary information of medical image and detects the nuclei by eq. (1), where homogenous intensity based pixels regions are formed as foreground of second derivation. Since $g(x, y)$ is position of particular pixel p which represents the each

nucleus considering homogenous threshold eq. (2).

$$g(x, y) \quad = (x_i, y_i) \leq 0 \ \ (P) \leq 1 \ (2)$$

Proposed algorithm visits each pixel and selects appropriate set of weighted pixels p designated as regions and tagged as Enlargement of nuclei where shape of various nuclei behaviors are considered with same visual properties. the task of this subsection of algorithm is to extract the statistical morphological features. Additionally, canny edge detection algorithm is used to record the shape related changes between the set of nuclei in medical images.

3.2. Canny edge detection:

Nuclei edges provide additional information about the selected regions of enlarged objects where shape sizes could be recorded for further analysis and eq.(3) describes that all the edges of medical image objects are the lies between the $(G) = \sqrt{G(x_i, y_i)^2}$, since the G is targeted connected set of lines to be formed around the (x_i, y_i) spatial locations. G_x and G_y Angles could also be used to measure the directions of objects which are involved in expected expansion over multiple regions eq. (4). This

helps to quantify the max and min distances for overlap objects and by $(\frac{G_x}{G_y})$ dividing the spatial positions logical separation could be done.

$$Edge_{Nuclei}(G) = \sqrt{G(x_i, y_i)^2} \qquad (3)$$

$$Angle_{Nuclei\ (\theta)} = tan^{-1}(\frac{G_x}{G_y}) \qquad (4)$$

3.3. Layer 2: Behavior detection and feature selection:

- Enlargement:

$$\boldsymbol{Enlargment} = \sum_{j=n}^{i=1} \boldsymbol{Sizeofp} = 1 \qquad (5)$$

Let's consider (x_i, y_i) are the spatial locations residing in the vector space of $f(G)$ where shape are not equal to the same size eq. (5). It is said to enlarge shapes that are counted with the assistance of auto image separation techniques as per eq (6). The image particles are considered as seed and each seed is cropped with its mean value and surrounded area is mapped with edges that the necessary information of particular object can be obtained.

$$\boldsymbol{seed}(\boldsymbol{\mu x, y}) = \{(\frac{\pi_x}{\pi_y})\boldsymbol{Sum}(\boldsymbol{a(x, y)})\}$$

$$(6)$$

- **Irregularity:**

Mostly the enlargement detection deceives in terms of object detection, since the proposed preprocessing technique measure the center of enlarged sequences of nuclei and rings acquired through canny edge detection are transformed around the large objects $(x_i + h_i, y_i + h_j)$ where (h_i, h_j) are transformed as boundaries of multiple objects.

$$\mathbf{X_{circle\ c}} = \frac{1}{M}\sum_{i=1}^{n} \mathbf{x_i\ m_i}(\frac{\pi_{x+h}}{\pi_y + h}) \qquad (7)$$

The next task of our algorithm is to encircle the closed objects $X_{circle\ c}$ where every point is denoted by M and numbers of associated pixels have been calculated with aggregated values eq. (7). Since the locations

$(x_i + h_i(x_i + h_i, y_i + h_j))$ are encircled. The irregular margins M have been counted and recorded in each observation of training and testing datasets.

- **Hyper chromatin:**

The hyper chromatin is also known as access of chromatin could be found in cells. The cell wall of nuclei has irregular quantities of the chromatin fluid. It is a condition where nuclei could not able to maintain its shape related properties due to the detachment of surrounding nuclei walls and loss of nucleus properties at DNA levels. Computationally proposed subsection of algorithm records a high concentrated mean value of chromatin color movement based features and color deviation is quantified on the basis of mass value and recorded into the observation.

$$Hyperchromatin = \sqrt{\frac{\sum_{i=1}^{n^*}(y_i - \hat{y}_i)^2}{n^*}} \qquad \textbf{(8)}$$

Hyper chromatin is one of the key feature in which each nuclei has to lose the chromatin. Since the loss of chromatin may incurred due to the eccentric nucleolus of nuclei or the expansion in the size of nuclei may be affected to nuclei shape, size and behaviors. Aggregated Color movements were measured with the assistance of distance matrix, where color spectrum is shown in [Figure 2] to represent the behaviors of chromatin in each nuclei with sounding set of nuclei. Let us suppose every DICOM image D is represented as collection of pixels contain the information based upon $X = \{X = x_1, x_2 \ldots \ldots . x_n\}$ and $Y = \{y_1, y_2 \ldots \ldots . y_n\}$ where each set of pixels $P(X, Y) = \{m_1, m_2 \ldots \ldots . m_n\}$ qualifies an object with distinct set of features $F = \{f_1, f_2 \ldots \ldots . f_n\}$ on $H = \{h_1, h_2 \ldots \ldots . h_n\}$.

3.4. Layer 3: Classification of nuclei:

Random forests machine learning algorithm is widely used to predict the different classification problems due to

several advantages to regress the rank relationships of important variables. Deep decision trees have ability to aggregate more than one decision related to every class label attribute.

Let's consider DICOM dataset D consists having several no of important variables $D = \{(X_i, Y_i)\} \, _{i=1}^{n}$. the decision model can be constructed as aggregated decision variable by fitting by measuring the J_{th} feature in training stage, since the decision tree J_{th} features are distressed datasets because of aggregated permutation operations (out of bag error) conducted for overall trees.

$$\hat{y} = \sum_{i=1}^{n} W(x_i, \acute{x}) y_i \qquad (9)$$

Let's suppose $\hat{y}$ is a class of functions in $x^,$set of points which formalized by weighted function known as W. thus following weighted decision tree can be constructed as per eq. (9) and (10).

$$\hat{y} = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} W_j(x_i, \acute{x}) y_i = \sum_{i=1}^{n} \left(\frac{1}{m} \sum_{j=1}^{m} W_j(x_i, \acute{x}) \right) y_i \qquad (10)$$

3.5. Layer 4: performance evaluation:

A total number of 20 biopsy images were used for training and testing purposes and 1829 number of nuclei were detected from those images. The confusion metrics [Table 1] shown for cancerous class which classified 948 observations and non-cancerous classified a number of 830 instances. . The precision eq. (11) and recall eq. (12) was approximated for cancerous classes 98.63% and about 97.73% was measured for non-cancerous classes.

$$Precision = \frac{NumberofTrePositives}{NumberofTruePositives + FalsePositives} \qquad (11)$$

$$Recall = Sensitivity = \frac{TruePositive}{FalsePositive} \qquad (12)$$

$$Sepecify = \frac{TrueNegetive}{TruePositive + FalsePositive} \qquad (13)$$

4. Results

In [Figure 2] there are five columns. The column one is presented as an image input where all the images are belonging to the Anaplast thyroid cancer with different histo-pathological material H and E. In second column the points of individual cells have been shown, where each point denotes number of enlarge nuclei. In column number three the irregular nuclei are placed where each set of overlapping of nuclei are represented. The system generated circles show that there is several numbers of nuclei which are eccentric due to the loss of cell wall. In column number four hyper chromatin, level is shown by distance matrix in a color spectrum which reveals that the all those nuclei are considered as hyper chromatin where red color entropies are found with highest aggregated values. The ratio of chromatin is said to be hyper when the surrounded features are present with low red color ratio using color spectrum properties. Column five is placed to show the summarization of each observation. All these detected features have been recorded in a dataset for training and testing purposes.

Figure 2: Results of preprocessing algorithm Anaplast Diagnosis

Table-I: Confusion matrix

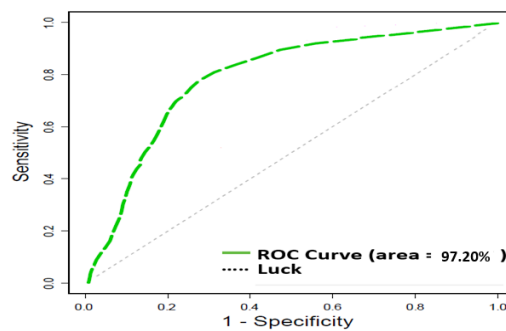|	*Cancerous*	*Non-Cancerous*
Cancerous	948	31
Non-Cancerous	20	830

Table-II: Overall performance of proposed methodology

	Raw images	*No of Extracted nuclei*	*No of classified Nuclei*	*No of miss-classified Nuclei*	*Precession*	*Recall*
Cancerous	10	979	948	31	98.63%	98.83%
Non-Cancerous	10	850	830	20	97.73%	97.64%

Table-III: Comparison of our system with literature

Approaches	*Image Type*	*Cancer Type*	*Technique*	*Accuracy*
1	Ultrasound Image	Follicular	SVM	97.50 %
			AdaBoost	87.42%.
2	FNAC Images	Follicular	NN	91.00 %
3	FNAC Images	Medullary	Templated matching strategy	87.00%
4	FNAB Images	Papillary	Scaled Conjugate Gradient	90.5%,
			BFGS Quasi-Newton,	85.40%
			Gradient Descent method	86.30%
			Bayesian regularization	83.50%
Our Proposed approach	FNAB Images	Anaplast	Random forest	97.20%

Figure 3: Estimated ROC Curve for the proposed

5. Conclusion

This paper contributes a novel data preparation algorithm AD (Anaplast Diagnosis) for Anaplast cancers, since the feature selection is one of the difficult tasks in medical images diagnosis. It needs significant efforts from the perspective of image feature engineering to select the related features from DICOM (Digital Communication in Medicine) images of thyroid biopsy such as enlargement of nuclei, irregular nuclei as individual & also in sequences and different chromatin distribution level features.

The methodology of proposed system comprises over four interconnected layers, where layer one has been assigned the job to reduce the noise and the layer two has been assigned core tasks to select appropriate features by measuring the aggregate values of involved pixels as color movements of selected objects. The layer three has been used to construct the classification model by using the random forest algorithm and the final layer has to perform the performance evaluation of the system.

The numbers of 1829 nuclei were detected from 20 images of cancerous and non-cancerous classes. The confusion matrix show that a total number of 948 instances out of 979 observations were classified as cancerous class and out of 850 instances 830 observations were classified for non-cancerous class.

The overall accuracy of the system was recorded as 97.2% with 10-k fold cross validation. Since such datasets are unavailable in literature, used dataset was received form SMBBMU, Pakistan. By observing the experiment in this research, we conclude that special data preparation algorithms are required to be developed for each histo-pathological medical image classification problem distinctly, because every DICOM image can be identified with its own properties which are mostly different in nature from each other.

REFERENCES

[1] Lashkari, AmirEhsan, and Mohammad Firouzmand. "Early Breast Cancer Detection in Thermogram Images using Supervised and Unsupervised Algorithms." The Journal of Engineering Vol 7, No 3, pp. 114-124, (2016)

[2] Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M. and Chang, E.I. , May. "Deep learning of feature representation with multiple instance learning for medical image analysis". In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference pp. 1626-1630. (2014)

[3] Cheng Chen, Wei Wang, John A. Ozolek, and Gustavo K. Rohde., "A flexible and robust approach for segmenting cell nuclei from 2D microscopy images using supervised learning and template matching". Journal of Cytometry A. 2013 May ; 83(5): p.p 495–507, (2013)

[4] Pourahmad, S., Azad, M., Paydar, S., & Reza, H. "Prediction of malignancy in suspected thyroid tumour patients by three different methods of classification in data mining". In First International Conference on advanced information technologies and applications pp. 1-8, (2012).

[5] Glotsos, D, Tsantis, S. Kybic J, Daskalakis, A. "Pattern recognition based segmentation versus wavelet maxima chain edge representation for nuclei detectionin microscopy images of thyroid nodules". Euromedica medical center,

Department of Medical Imaging, Athens, Greece (2013).

[6] Gopinath, B, Shanthi, N.. "Support Vector Machine Based Diagnostic System for Thyroid Cancer using Statistical Texture Features", Asian Pacific J Cancer Prev, 14 (1), pp. 97-102. (2013).

[7] Gopinath, B, Shanthi, N. "Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images", Australas Phys Eng Sci Med Vol (36), pp. 219–230, (2013).

[8] Stavros T., "Improving diagnostic accuracy in the classification of thyroid cancer by combining quantitative information extracted from both ultrasound and cytological images". 1st IC-SCCE-Athens, pp. 8-10, (2004).

[9] http://www.thyroidmanager.org/chapter/fine-needle-aspiration-biopsy-of-the-thyroid-gland/ accessed on 15-01-2014.

[10] Thyroid Cancer: Web Site, http://www.medicinenet.com/thyroid_cancer/article.htm accessed on 15-01-2014

[11] Han J., Kamber M., and Pei P. "Data Mining: Concepts and Techniques",. 3rd (ed.) The Morgan Kaufmann Series in Data Management Systems (2011).

[12] Rafeal, C. G., Richard, E. W., (2007). "Digital Image Processing" , 3rd (ed). Prentice Hall. (2007).

[13] Hussein, Mohamed, Amitabh Varshney, and Larry Davis. "On implementing graph cuts on cuda." First Workshop on General Purpose Processing on Graphics Processing Units. Vol. (1). (2007).

[14] Boykov, Y., & Funka-Lea, G. "Graph cuts and efficient ND image segmentation". International journal of computer vision, 70(2), pp. 109-131, (2006).

[15] Malik, J., Belongie, S., Leung, T., & Shi, J. "Contour and texture analysis for image segmentation". In Perceptual Organization for artificial vision systems Vol(1) pp. 139-172. (2000).

[16] Al-Kofahi, Y., Lassoued, W., Lee, W., & Roysam, B. "Improved automatic detection and segmentation of cell nuclei in histopathology images".Biomedical Engineering, IEEE Transactions on, 57(4), pp. 841-852. (2010).

[17] Liu, X., Huo, Z., & Zhang, J. "Automated segmentation of breast lesions in ultrasound images", In Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the IEEE pp. 7433-7435, (2006).

[18] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350. (1963).