A Robust Feature Extraction and Selection Method for the Recognition of Lymphocytes versus Acute Lymphoblastic Leukemia

Hayan T. Madhloom[1], Sameem Abdul Kareem[2]
Department of Artificial Intelligence, Faculty of Computer Science and Information Technology
University of Malaya
Kuala Lumpur, Malaysia
Hayan.tariq@siswa.um.edu.my
Sameem@um.edu.my

Hany Ariffin[3]
Department of Pediatric, Faculty of Medicine
University of Malaya
Kuala Lumpur, Malaysia
Hany@um.edu.my

Abstract— An essential part of the diagnosis and treatment of leukemia is the visual examination of the patient's peripheral blood smear under the microscope. Morphological changes in the white blood cells are commonly used to determine the nature of the malignant cells, namely blasts. Manual techniques are labor intensive slow, subjected to error and costly. A computerized system can be used as an aiding tool for the specialist in order to improve and accelerate the morphological analysis process. This paper presents and application of feature extraction, selection and cell classification to the recognition and differentiation of normal lymphocytes versus abnormal lymphoblast cells on the image of peripheral blood smears. This is considered as a very useful procedure in the initial treatment process of leukemia patients. A computerized recognition system has been developed, and the results of its numerical verification are presented and discussed. The methodology demonstrates that the application of pattern recognition is a powerful tool for the differentiation of normal lymphocytes and acute lymphoblastic leukemia, leading to the improvement in the early effective treatment for leukemia.

Keywords- Shape Features, Texture Features, Image Segmentation, Leukemia diagnosis

I. INTRODUCTION

Leukemia is the commonest childhood malignancy and accounts for 30% of childhood cases yearly. Acute lymphoblastic leukemia (***ALL***) is the predominant type, contributing approximately to 80% of total cases of childhood leukemia [1]. In ***ALL***, there is an abnormal, uncontrollable proliferation of lymphoid precursors called lymphoblasts in the bone marrow with arrested maturation [2]. In many cases, these lymphoblasts escape the bone marrow comportment and circulate in the patient's peripheral blood. The French-American-British (FAB) classification categorizes acute lymphoblastic leukemia into three morphological subtypes (L1, L2 and L3). [3]

Oral steroids eg. prednisolone are the keystone in the treatment of children with ***ALL***. The response to this agent has been shown to have significant prognostic value [4,5,6,7,8]. Many ***ALL*** treatment protocols have a "window" of prednisolone monotherapy to risk-stratify patients according to treatment response. To assess steroid response, the peripheral blood film (PBF) of the patients has to be examined on the 8th day of therapy where the numbers of circulating leukemia cells are determined. The patient with more than 1000 lymphoblasts per uL in the peripheral blood after one week of oral prednisolone therapy are regarded as "poor-risk" and will receive more intensive chemotherapy whereas those who are found to have less than 1000 lymphoblasts per ul of circulating blasts are considered "good-risk" and will have less intensive therapy. Thus, it is important to have a precise assessment of the number of circulating lymphoblasts. The process of counting lymphoblast cells versus normal lymphocytes is currently done manually by the hematologist. The accuracy of this counting procedure has many implications to the patient's treatment and ultimate prognosis. Hence, it would be useful if there is computer software that can aid in the identification of these abnormal leukemia cells in the peripheral blood smear at the 8th day of prednisolone therapy to increase the confidence of an accurate count by the hematologist.

II. LITERATURE REVIEW

The most accurate blood cell recognition system that exists nowadays in the hospitals and medical centers is work based on the concepts of a light or laser beam that points on each cell and then a computerized system analyzes the reflected signal in order to determine the type of the cell such as the flow cytometer. Although this medical instrument is very robust and accurate, however, it is very expensive and not all the medical center and hospitals around the world can afford to buy such device. Since the last two decades, researchers start to look for a blood

978-0-7695-4959-0/13 $25.00 © 2013 IEEE
DOI 10.1109/ACSAT.2012.62

recognition system that is based on image processing and artificial intelligence, which can be significantly cheaper than the other earlier mentioned technology. Blood cell recognition systems that are based on artificial intelligence concepts have developed remarkably during the past few years where new blood cell counters become available in the world market [9,10], offering a different count of various blood cells. These normally apply the back propagation neural network as a classification tool. The accuracy of this system compared to human expert is about 15%. Thus, this system is unable to classify abnormal cells with a really accepted accuracy demanded in the hospital procedure for all types of cells in the blood [10]. Besides, the images have to go through a careful and precise procedure in order to test the quality before feeding them to the system since this sort of system needs very high quality images. Therefore, much work need to be done to produce an efficient, robust and accurate blood recognition system that can accept various kinds of blood cell whether it is normal or abnormal and give perfect results comparable to the specialists.

The research and development has continued, and much work have been done in the area of blood cell recognition, either dealing with red blood cells (RBC) like the researches that address the problems of malaria and anemia classification; or differential blood count where the focus is clearly on normal white blood cells (WBC) or some other work on leukemia diagnosis [11].

As this work deals directly with the problem of differentiating between normal lymphocytes and lymphoblast cells, which are leukemia cells, there are very few researches that address the problem of normal WBC versus leukemia cells [12].

All the work done previously either tries to discover the differential blood count which deals with normal WBCs or leukemia diagnosis; such as distinguishing between acute lymphoblast leukemia (***ALL***) and acute myeloid leukemia (***AML***), or distinguishing between subtypes of one leukemia types like (L1-L2-L3) in ***ALL***.

Reta, et al [13] introduced a system that can categorize the two types of leukemia ***ALL*** and ***AML***. Various features and attributes were used such as geometrical, statistical, texture, size ratio, and principle component analysis; however, feature selection was not used; which in turn affects the classification accuracy. Furthermore Farag A [14] discovered a method for ***ALL*** and ***AML*** differentiation, the author depended on the biological feature related to the thickness of the cytoplasm as a feature that can be used for the classification. Some other researchers such as Scotti [15, 16] who has contributed significantly to this field, worked on both normal and leukemic cells. In 2004, Scotti worked on a differential blood count system where 23 features were extracted from the cells, most of the features are geometrical and only the mean gray-level intensity of the cytoplasm was used as a statistical feature. A forward selection method based on the nearest neighbor classifier was used as a feature selection method. The classification was done using various classifiers such as k-nearest neighbor (KNN) and feed forward neural network (FFNN). The results showed that the parallel neural network obtained the lowest error rate. However, the KNN obtained better results than a conventional feed forward neural network (FFNN). In 2005 Scotti [12] repeated almost the same steps of his previous experiment carried out in 2004 [15], however, this time the experiment was done on ***ALL*** cells to classify the three subtypes of ***ALL*** (L1-L2-L3). The experiment showed the FFNN obtained a better result than the KNN. For the studies done by Scotti, texture features were not used; however, texture features can play a significant role in the classification of both normal and abnormal WBCs [17]. Osowski, S et al,[11] presented a way to recognize a myeloid leukemia using bone marrow aspirate; various features were used including texture, shape and statistical. The classification was done using support vector machine, (SVM), but the number of misclassification cells were relatively high due to the image segmentation method used. Osowski, S continued his work by repeating the same experiment in [18], however, this time a linear SVM ranking was used as features selection method and only 30 features were ranked as the best. The classification is done using SVM, but the percentage error is still high at 18.71%. In 2007 and 2009 Osowski improved the classification result [19,20] by using genetic algorithm as a feature selection method. Using the genetic algorithm instead of linear SVM ranking, the result improved by 30%. However, the increase of accuracy comes with calculation cost.

III. MATERIALS AND MATHODS

A. Patients and Peripheral Blood Acquisition

The dataset used in this research consists of (260 cells). These cells were divided into two parts, one part contain 130 cells (90 are lymphocytes and the other 90 are ***ALL)*** this part is used as a training set. 80 cells (40 lymphocytes and 40 **ALL**) were kept as a test set. The PBF have been prepared using the standard staining method of May-Grunwald-Giemsa. The digitization process was done using an Olympus UC30 camera that is mounted on an optical microscope with a magnification of 400x. The UC 30 camera is a 3.2 mega pixels digital color camera with CCD chip. The live frame rate is 7.0 frames per second at 2080×1544, which is the resolution that is used for the images in this research. The dataset of the blood cells used in this experiment was obtained with the cooperation of the department of pediatric oncology at University of Malaya Medical Center (UMMC) in Kuala Lumpur, Malaysia. Fig 1 shows a sample image containing ***ALL*** and lymphocytes.

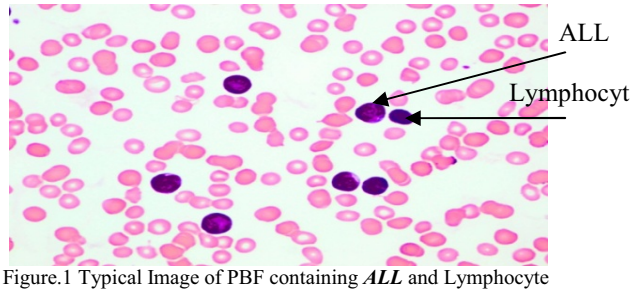

Figure.1 Typical Image of PBF containing ***ALL*** and Lymphocyte

B. Cell Localization and Segmentation

This research is a continuation of the experiments we conducted earlier [21, 22] which was particularly concerned with WBC (normal and abnormal) localization and segmentation. The localization was done using color features with the morphological reconstruction to isolate lymphoblast cells from a microscope image that contains many cells. The experiment was tested on 180 microscope blood images, and it effectively produced 100% accuracy for the localization and 90-95% of cell segmentation as compared to a manual segmentation that was done by a hematologist.

C. Feature Extraction

Visual information plays an important role in clinical diagnosis. It is widely known that there is no common distinct group of features that can be suitable for all kinds of computer vision application, for instance the features that can be used for leukemia classification purposes is not totally the same as the features that can be used for a breast cancer diagnosis system since the two region of interest (ROI), have different appearances, so the feature is usually selected based on the application. All the differences between the lymphocytes and ***ALL*** should be considered with great attention paid for the distinctive features. Normally during the procedure of manually observing a blood slide, the hematologist uses apparent features which are considered as a straightforward process; however, it is a very difficult task to mimic these features and formulate it in order to be used for the computer vision system. It is also very important to know that the efficiency of various features that are extracted from the cells cannot be evaluated without considering the efficiency of the system stages prior to the feature extraction stages namely the single cell localization and segmentation and also the slide preparation since it may affect the quality of the images used.

For the purposes of lymphocyte-lymphoblast classification, two types of features are extracted namely shape and texture features.

The shape features is a set of numbers that are produced to describe a given shape property such as size, axis length, convex area, etc. A descriptor attempts to quantify shape in ways that agree with the human perception. Good and accurate object classification based on shape features requires the classification to group objects with similar shapes from a set of images. Usually, the feature descriptors are arranged in the form of a vector. This group of features plays a significant role in the differentiation between normal lymphocytes and ***ALL***, since the ***ALL*** is an immature cell from a lymphocyte descendant and it may have different shape characteristics than the normal lymphocyte. Fig 2 illustrates the representation of the shapes features that are extracted from each cell, and column two of Table I shows all the shape features and the corresponding ROI.

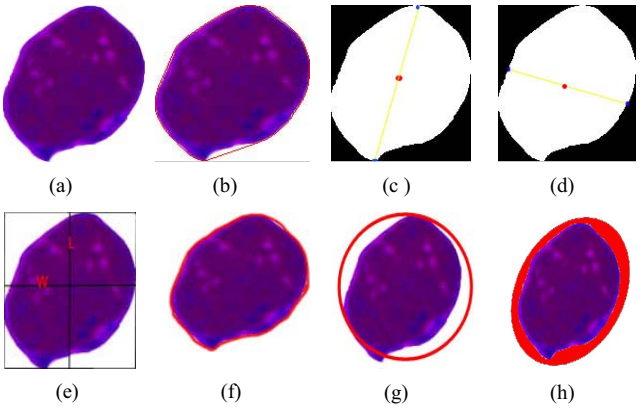

Figure 2 Illustration of Shape features

a) Original Cell b) Convex area c) Major axis length d)Minor axis length e) Bounding Box f) perimeter g) Circularity h) Bounding ellipse.

Texture features is described as a repeating pattern of local variations in image intensity, where the intensity variations are high among different areas of the image. Two types of texture features were used that is the First-order texture (histogram statistics) and the second-order texture (based on gray level co-occurrence matrix). Table I, column three shows a list of all the texture features.

TABLE I. SUMMARIZATION OF ALL THE EXTRACTED FEATURES BASED OF THE REGION OF INTEREST

Region of Cell	**Shape Features**		**Texture Features**	
Nucleus	Area Convex hull Major Axis Length Minor Axis Length Bounding Box Eccentricity Perimeter Circularity		First-order	Mean Standard Deviation Smoothness Skewness Uniformity Entropy
	Elliptical Features	Major and Minor Axis Length Eccentricity Perimeter diameter	Second-order	Homogeneity Contrast Entropy Correlation Energy Variance Cluster Shade Cluster Prominence
Cytoplasm	No shape features		No texture features	
Whole Cell	Area Ratio of nucleus to the whole cell		No texture features	

IV. FEATURE SELECTION AND CLASSIFICATION

A group of 30 features were extracted from each cell. Not all of these features may be useful. Some may show a high mutual correlation, so there is no point in using all the features for classification.

The process of selecting the best features is described below:

1. The data will be normalized to restrict the values of all features within a predetermined range. This step will help to overcome the problem of large values for some features that may affect the classification performance. The data is normalized to have zero mean and standard deviation equal to 1 using the following equation

$$\widehat{X}_i = \frac{X_i - \bar{X}}{\sigma} \quad i = 1,2 \dots \dots N \quad \dots\dots(1)$$

Where $\bar{X}$ and σ are the mean and standard deviation computed from the values of a feature X_i and $\widehat{X}_i$ is the normalized value.

2. The features are ranked using scalar feature selection method with a Fisher's Discrimination Ratio (FDR) taking into consideration the cross-correlation among all the features. The top seven features will be selected.
3. Exhaustive search will be carried out to select the best three features using a scatter measure as a measure for class separation.

For the purpose of data classification, the k-nearest neighbor was used. This classifier is a statistical classifier. ***k*** was chosen equal to 3 and the Euclidean distance was used as a distance metric.

V. RESULTS AND DISCUSSION

A combination of shape and texture descriptors was used. 15 shape features were employed, some of them were extracted from the nucleus and some others obtained from the whole cell. 15 texture features were extracted from the nucleus of both types of cells, the cytoplasm shape and texture was not used as a measure since there was not much difference between the two classes of cells.

The dataset were normalized. The normalization gave a significant contribution to the classification performance. It made the range of the values lie within a predetermined range of [0,1] and this step will positively affect the classifier performance, since some of the features have large values, and without normalization these large values may have an adverse impact on the classifier.

The scalar feature selection was used to rank the features so the number of features can be reduced by selecting only the top features that have the best discrimination measure. The discrimination measure was done using FDR, supported by cross-correlation measure that will also measure the degree of similarity between the features. The top seven most discriminate features were selected as shown in Table II below arranged in descending order according to the FDR

Table II. THE TOP SEVEN FEATURES RESULTED FROM SCALAR FEATURE SELECTION

Sequence	***Feature***	***Discrimination Measure***
1	Nucleus Major Axis	3.118
2	Nucleus Convex Area	2.895
3	Nucleus Diameter	2.812
4	Nucleus Bounding Box Area	2.808
5	Nucleus Cluster shade	2.804
6	Nucleus Cluster Prominence	2.800
7	Minor axis of the bounding ellipse	2.787

Using these seven features would have resulted in getting a high classification error rate, since some of the features have very close discrimination measures to each other; therefore an exhaustive search was applied on these seven features to find the combination of the best three features. Even though the exhaustive search is considered slow in terms of processing, however there were only seven features to be searched, hence the processing speed was not affected. The result after the exhaustive search is shown in Table III and the features are plotted in a hyper plane as shown in Fig.3. The three features reveal a significant distribution for the test set of the two classes.

Table III. THE RESULTED FEATURES AFTER PERFORMING EXHAUSTIVE SEARCH

Selected features	***Feature***
3	Nucleus Diameter
6	Nucleus Cluster Prominence
7	Minor axis of the bounding ellipse

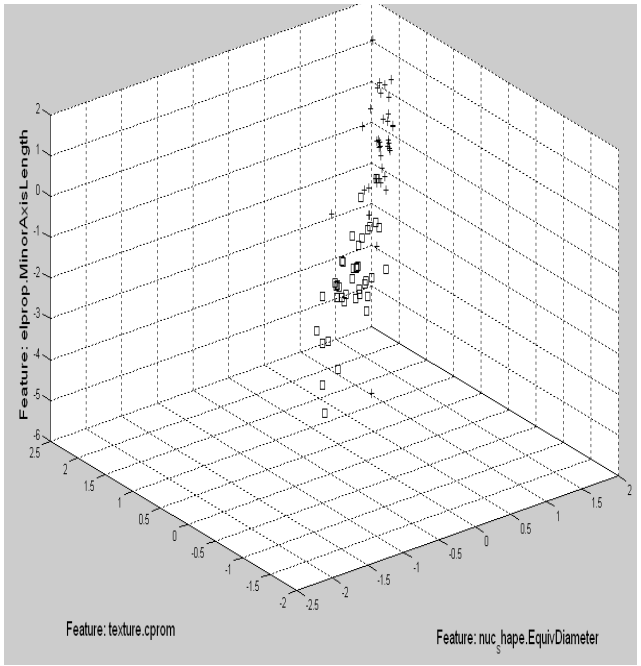

Figure.3 The distribution of the resulted three features in a 3-D hyperplane

A k-nearest neighbor classifier (*k*-NN) with Euclidean distance was used; the test set distribution which consisted of 80 cells (40 normal lymphocytes and 40 ***ALL***) is shown in Fig.4. From Fig.4 it is clear that there are three misclassified cells, these cells were of type ***ALL*** and was wrongly classified as normal Lymphocyte. This misclassification may be due to the similarity between the two types especially in their morphological features. The accuracy that was obtained from this experiment was about 92.5%

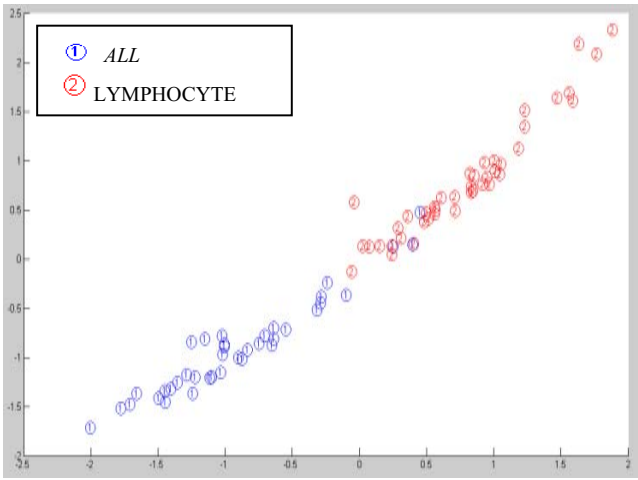

Figure 4. The classification result

The clinical impact of this research is that it will provide the ability to physicians to examine a patient's PBF for counting the number of normal lymphocytes against leukemia ***ALL***. This process is considered as a key step in determining the intensity of treatment that will be given to the patient. The result of this research can be used also for the evaluation of the patient's clinical state upon follow-up.

In Malaysia, a country of 13 states with 27 million people, there are only four tertiary-referral centers for childhood cancer and less than 30 trained pediatric hemato-oncologists. Hence, having a tool to facilitate the counting of normal versus abnormal cells of children having ***ALL*** would be beneficial to clinicians and laboratories in terms of speeding up the process and verifying the results especially if it is done by a junior personnel. This tool can also be used for education purposes where the trainee can use the computer-based tool to validate his findings.

VI. CONCLUSION

This paper presented a method to distinguish between normal lymphocytes and ***ALL***. The focus of the research was to find the useful features that can strongly participate in the classification and recognition of the two classes. The numerical experiment performed on the set of 260 blood cells confirms the promising results of the proposed method. The important advantage of this approach is the feature set normalization followed by feature ranking and subsequently obtaining the optimum set of features by exhaustive search. Applying feature selection without data normalization produces very poor classification accuracy [23]. The focus in this research was on producing the best features that can give the best classification performance. In future other types of leukemia will be included such as acute myeloid leukemia ***AML.*** The method can also be expanded to test the classification accuracy for ***ALL*** versus ***AML.*** Besides that, various classification methods such as neural network and support vector machine can also be used and the results compared.

REFERENCES

[1] LAG Ries, D Melbert, and M, Krapcho, "SEER Cancer Statistics Review, 1975–2005". Bethesda: National Cancer Institute, 2008.

[2] CL. Sawyers, CT. Denny, and ON. Witte, "Leukemia and thedisruption of normal hematopoeisis". Cell vol. 64 issue. 2 pp.337–350, 1991.

[3] B.J Bain., "Leukemia diagnosis, 3rd ed, Blackwell Publishing, 1991. pp.2

[4] ML. Den Boer, DO. Harms, R. Pieters, et al. "Patient stratification based on prednisolone-vincristine-asparaginase resistance profiles in children with acute lymphoblastic leukemia". Journal of clinical oncology; vol. 21 no. 17 pp.*3262*–3268, Sep. 2003

[5] T. Hongo, S. Yajima, M. Sakurai, Y. Horikoshi, R. Hanada. "In vitro drug sensitivity testing can predict induction failure and early relapse of childhood acute lymphoblastic leukemia". Blood; vol. 89 no. 8, pp. *2959*–2965. Apr. 1997

[6] GJ. Kaspers, R. Pieters, CH. Van Zantwijk, ER. Van Wering, A. Van Der Does-Van Den Berg, AJ. Veerman "Prednisolone resistance in childhood acute lymphoblastic leukemia: vitro-vivo correlations and cross-resistance to other drugs". Blood; vol. 92, no. 1, pp.259–*266.* Jul.1998

[7] M. Lauten, M. Stanulla, M. Zimmermann, K. Welte, H. Riehm, M. Schrappe. "Clinical outcome of patients with childhood acute lymphoblastic leukaemia and an initial leukaemic blood blast count of less than 1000 per microliter". Klinische Padiatrie; vol 213 no. 4, pp. 169–174 Jul-Aug. 2001

[8] R. Pieters, ML. den Boer, M. Durian, et al. "Relation between age, immunophenotype and in vitro drug resistance in 395 children with acute lymphoblastic leukemia–implications for treatment of infants". Leukemia; vol. 12 no. 9, pp.1344–1348. Sep. 1998

[9] H. Ceelie, RB. Dinkelaar, W. Van Gelder, "Examination of peripheral blood film using automated microscopy; evaluation of diffmaster Octavia and cellavision" Journal of Clinical Pathology, vol. 60 no. 1, pp. 72-79, Jan 2007.

[10] B. Swolin,P. Simonsson ,S. Backman ,I. Löfqvist, I. Bredin, M. Johnsson, "Differential counting of blood leukocytes using automated microscopy and a decision support system based on artificial neural networks – evaluation of diffmastertm octavia". clinical & laboratory haematology, vol. 25, no. 3, pp. 139-147, Jun. 2003.

[11] F. Sadeghian, Z. Seman, AR. Ramli, BHA. Kahar, M. Saripan "A Framework for White Blood Cell Segmentation in Microscopic Blood Images Using Digital Image Processing". Biological Procedures Online vol. 11, no.1, pp. 196-206, Jun. 2009.

[12] F. Scotti "Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In: IEEE International Conference on Computational Intelligence for Measurement Systems and Applications. 2005, pp. 96 - 101

[13] C. Reta, L. Altamirano, J. Gonzalez, R. Diaz, J..Guichard "Segmentation of bone marrow cell images for morphological classification of acute leukemia. in Twenty-Third International Florida Artificial Intelligence Research Society Conference May, 2010; Florida,USA; pp. 86-91

[14] A. Farag "Computer based acute leukemia classification". in Proc of the 46th IEEE International Midwest Symposium on Circuits and Systems, 2004, pp. 701- 703

[15] V. Piuri. F. Scotti. "Morphological classification of blood leucocytes by microscope images". IEEE International Conference on Computational Intelligence for Measurement Systems and Applications. 2004: pp. 103 - 108.

[16] F. Scotti "Automatic Morphological analysis for acute leukemia identification in peripheral blood microscope images". in IEEE International Conference on Computational Intelligence for Measurement Systems and Applications. 2005, pp. 96 – 101.

[17] D. Sabino, L. Costa, EG. Rizzatti, MA. Zago. "A texture approach to leukocyte recognition". Real-Time Imaging, vol.10 no.4 pp. 205-216, Aug. 2004

[18] T. Markiewicz, S. Osowski,B. Marianska., L. Moszczyn'ski. "Automatic Recognition of the Blood Cells of Myelogenous Leukemia Using SVM". in International Joint Conference on Neural Networks, Montreal, Canada; 2005: pp.2496-2501.

[19] S. Osowski,S. Robert,T. Markiewicz,K. Siwek "Application of Support Vector Machine and Genetic Algorithm for Improved Blood Cell Recognition". IEEE Transactions on Instrumentation and Measurement 2009, vol. 58, no.7, pp. 2159-2168, Jul. 2009.

[20] R. Siroic, S. Osowski, T. Markiewicz, K. Siwek "Support vector machine and genetic algorithm for efricient blood cell recognition". In Proc. IEEE Instrumentation and Measurement Technology. 2007: pp. 1-6.

[21] H. Madhloom,SA. Kareem,H. Ariffin, AA. Zaidan, HO. Alanazi, BB. Zaidan. "An automated white blood cell nucleus localization and segmentation using image arithmetic and automatic threshold". Journal of Applied Sciences, vol. 10, no.11, pp.959-966. 2010.

[22] H. Madhloom, SA. Kareem, H. Ariffin. "An image processing application for the localization and segmentation of lymphoblast cell using peripheral blood images", Journal of Medical Systems, vol. 36, no. 4 pp. 2149-2158. Aug 2011

[23] S. Theodoridis , A. Pikrakis K. Koutroumbas, "Pattern recognition" 4th ed, Academic Press 2008 pp.263