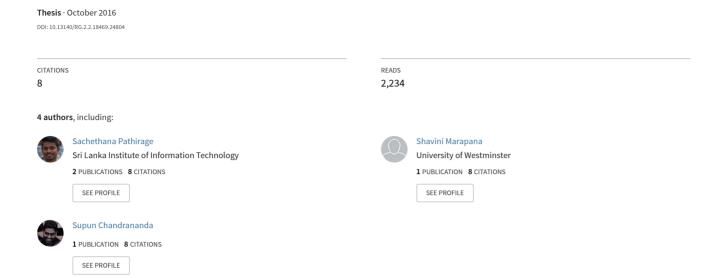
Detection of Leukemia using Image Processing and Machine Learning



Detection of Leukemia using Image Processing and Machine Learning

Supun S. Chandrananda, Sachethana H. Pathirage, Shavini R. Marapana, Nishika M. Amarathunga and Sanjeevi S. Chandrasiri

Abstract—Counting of blood cells plays a very important role in the health sector. The old conventional method used in hospital laboratories involves the manual counting of blood cells using a device called Haemocytometer. But this process is extremely monotonous, time consuming, which leads to inaccurate results. In order to overcome these complications, this research presents a fully automated software solution, enriched with image processing and machine learning techniques to detect and to count the number of RBC, WBC and Platelets cells in the blood sample images and to classify diverse types of Leukaemia. Several problems and missing features in existing white blood cell classifiers were addressed by implementing an effusively automated method using a multi-class classifier. This approach identifies various colour feature statistics with geographical measures for machine learning centred on supervised learning. Also a training model which is capable of being resumed was implemented in the research. The proposed classifier was able to avoid the numerous drawbacks identified by previous researchers and shortcomings on manual identification with a success rate of 84.8%. The system focus on four types of Leukaemia which are Acute Lymphoblastic Leukaemia (ALL), Chronic Lymphocytic Leukaemia (CLL), Acute Myelogenous Leukaemia (AML or Myeloblastic), Chronic Myelogenous Leukaemia (CML).

Keywords— Image Processing; Machine Learning; Multiclass Classifier; Haemocytometer; Circular Hough; Blob Detection; Leukaemia; White Blood Cells;

I. Introduction

Leukemia is a variety of cancer that commences in the bone marrow which results in high numbers of abnormal white blood cells [1]. These white blood cells are not completely developed and are called blasts or leukaemia cells. Leukaemia can be divided into four major categories including Acute Lymphoblastic Leukaemia (ALL), Chronic Lymphocytic Leukaemia (CLL), Acute Myelogenous Leukaemia (AML), and Chronic Myelogenous Leukaemia (CML). The research aims to detect Leukemia cancer types using image processing machine learning techniques. The system use morphological features in microscopic images to examine changes on texture, geometry, colour and statistical analysis. It is believed that this system would increase the diagnostic accuracy and consistency of the haematologist and laboratory practitioner in the daily diagnostic routine.

Problems mislaid and existing in current WBC classifiers were identified as usage of binary classifiers which only states whether a single type of WBC has leukaemia instead of multiclass classifiers since it requires more work, misclassification was between several WBC types and the accuracy levels

stated in the systems are rarely above 70%. According to the framework proposed by F. Sadeghian [14] for segmenting white blood cells using the integration of concepts in digital image processing consists of an incorporation of several digital image processing algorithms. Twenty microscopic blood images were tested, and the framework managed to obtain 92% accuracy for nucleus segmentation and 78% for cytoplasm segmentation [14]. The results indicate that the framework is able to extract the nucleus and cytoplasm region in a WBC image sample. The approach scheme consists two parts: The nucleus segmentation based on morphological analysis, and the cytoplasm segmentation is based on pixel-intensity thresholding [2].

II. METHODS AND PROCEDURES

The research aims to compare diverse image processing techniques to extract and automatically count blood cells from a microscopic blood image. The exact count of cells will benefit to determine the early detection of disease and are accomplished with a sequence of 4 major steps.

A. Pre Processing / Segmentation and Counting Red Blood Cells, White Blood Cells and Platelets

In medical diagnosis, Complete Blood Count (CBC) [7] is considered as a significant aspect. The Table I summaries normal complete blood count results for adults. This can be used to evaluate the health level of the person and to detect the disorders like Leukaemia. There are mainly three classifications of cells. Red blood cells (RBC), White blood cells (WBC), and Platelets. The abnormal count of cells can be used as an indication to detect the presence of disease of a person. When compared to the manual process with the automated, though the automated analysers give fast and reliable results regarding the count, automated process cannot reliably count the abnormal cells and overlapped cells that are present in the blood image.

TABLE I STANDARD CBC OF A HEALTHY PERSON

Blood Cell Type	Women	Men	Unit
WBC	4-5	4.5-6.0	M/µL
RBC	4.5-11	4.5-11	K/µL
Platelets	150-450	150-450	K/µL

1) Image Acquisition: The digital microscope is interfaced to a computer and the microscopic images are obtained as digital images.

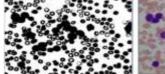
- 2) Image Pre-processing: Acquired images are preprocessed to remove superfluous noise and minor substances that are not measured as blood cells.
- 3) Image Enhancement: For better segmentation results of the blood cells, the obtained image has to be enhanced. This is achieved with few image processing techniques including contrast adjustment, grey-scale, edge detection, and spatial smoothing filtering [6].
- 4) Image Segmentation: This process involves selecting the region of interest in the image. This describes the area which comprises the blood cells. Circular Hough transform [17] is applied and not much of the image segmentation is needed because the applied transform looks only for the circular objects in the image.
- 5) Detection of Blood Cells: The Circular Hough transform searches for the blood cells in the image and then detects them. The function "draw circle" draws circles around the detected cells. Even the overlapped circles are detected.
- 6) Counting Cells: Counting the number of cells drawn gives the total number of blood cells in the image.

The Fig. 1 separate WBC from the background image using grey binary image to canny threshold (Left). And thin red circles had drawn around the detected WBCs (Right).



Fig. 1. Detected WBCs from the original image

The Fig. 2 elucidates the pre-processed image for counting RBCs in the image (Left). By using the Hough circles method thin circles drawn around the detected RBCs in the image (Right). The same method is used to detect and count the platelets.



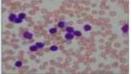


Fig. 2. Detected RBCs from the original image

7) Formula for counting: Actual count of the blood sample is calculated with the image area (Eq. 1). Now such samples are usually diluted with an anticoagulant liquid to avoid clotting [17]. In such cases we have to multiply the count by the dilution factor.

C – Cell count of the image

DF - Dilution Factor

D - Depth of the height of the counting chamber

A - Input image area

Total Count = $C \times DF / (A \times D)$ (1)

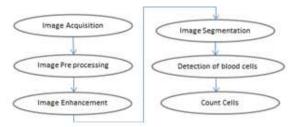


Fig. 3. Overall steps to Count Blood Cells

B. White Blood Cell Classification

It is essential to segment the WBCs from each image. Among the three RGB color components, the nucleus region in violet color has the least value of the green channel when compared to other regions such as cytoplasm or background [3]. Therefore, the nucleus region enhances in the input images by averaging the pixel values in the red and blue channels together and then dividing the sum by the intensity value of the green channel.

Erosion and dilation morphological operators with flat disk structuring element are applied to the nucleus enhanced image to remove any noisy pixels. Some types of WBCs have nucleus with multiple lobes, such as the eosinophil and the neutrophil [3]. The minimum area for each lobe can be calculated [3]. Morphology dilation is applied with a flat disk structuring merge them into one cell.

The WBC segmentation is carried forward according to the size and the shape of the cells. AForge's [4] built-in functions are used to calculate the area, width, height, number of occurrences per cell and other variables for each of these cells in order to obtain something meaningful to the classifier.

The classifier itself starts up and repeats the above process for each cell in each WBC cell folder and learns. Then it stores the data into a CSV file for re-use on the next classification of image to speed up learning time. The classifier will input the current WBC, which is from the patient's image to the SVM and then outputs the WBC category as an integer value. Since the classifier class is static, it doesn't repeat learning process for each image in a single blood sample. The WBC classifier simply needs the path to:

- Blood sample image
- WBC mask of the blood sample image

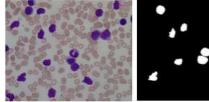


Fig. 4. Sample inputs to the WBC Classifier

The Fig. 4 explicates more how it looks like. It contains two examples of raw blood a sample (Left) and WBC masks (Right). The blob detector function [16] extracts the WBC's using the mask as the cell areas. At this point, if the count is zero, the classifier stops immediately. The extracted WBCs looks exactly like the WBCs in the input image and now ready to be processed through the WBC analyzing modules

0,144,160,16823,0.730164930555556, 78.6702728271484,69.159423828125, 88.6363677978516,273.442626953125,81

Fig. 5. Example output double array for a single WBC



Fig. 6. Separated WBCs from input image

Now these cells will be used as input to the next function which is to calculate the cell metrics and to obtain a double array. The input to this function is a single extracted WBC image presented in Fig. 6. This function is used for all the cells in the training data set in order to build up a 2D array for the SVM as well. As discussed in the methodology, colour based and geometry based metrics are collected per each cell. Colour based feature extraction runs through several functions in order to extract cell features (i.e.: nucleus, cytoplasm, holes) as necessary.

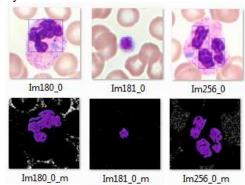
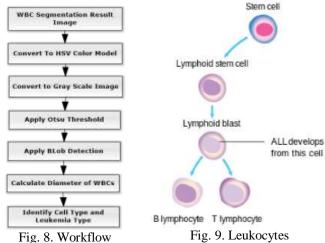


Fig. 7. Nuclei segmentation

An examination of the images in the middle of processing was necessary in order to check whether the colour channels are correct, function parameters are correctly filtering the required regions. The Fig. 7 shows some examples which were obtained while analyzing nucleus and the number of nucleus. As the third cell has three nucleuses, that count is also taken as an input to the metrics as well. For geographical, geometry based feature extraction, several variables are collected. As stated in the background context, the dark blue, violet areas were measured for diameter, outline width & height, area, fullness, blood sample image, center coordinates colour standard deviation, colour mean, count of occurrences. In cases where there are more than 1 area present, the average area was calculated. After collecting the metrics, it was presented as a double array as in Fig. 5. A double array of this format as listed in Fig. 5, is used for both learn and input for classification. Technically the machine is using a second degree polynomial kernel. Before classifying an image, then classifier automatically learns itself. For that it scans through the learning folder which has a specified folder structure, as this is dynamic, the user can contribute to the learning materials by copying new cell images to the folder. Each cell image inside these folders are scanned for metrics, and at the end of it a cache file is built up to speed up the next learning (not the next cell image). The next time the classifier is used, it will check for the cache file and uses that instead of reading each image file reducing disk I/Os. The classifier class is static, it only learns when it is first requested by an object. For each and every input, it will output an integer which will be mapped to the WBC class later. At the end of the classification, the classified WBCs will be stored in folders for their class.

C. Acute Lymphoblastic Leukemia (ALL) and Chronic lymphocytic Leukemia (CLL)

ALL and CLL affect a group of leukocytes called lymphocytes which is one of WBC type. Basically we can identify three types of lymphocytes. Those are normal lymphocytes shown in Fig. 10, mature lymphocytes shown in Fig. 11 and blast lymphocytes also called as lymphoblast shown in Fig. 12. These two diseases are identified by the count of each lymphocyte types [9]. Mature lymphocytes cell count is more than 25% which indicates human is CLL addicted [9]. Lymphoblast count is also more than 25% which directs as CLL addicted [9]. To identify ALL and CLL first we need to identify the type of lymphocytes. To do that we use a morphological feature of cell, it is nucleus diameter of lymphocytes. And then should get the count of each lymphocyte type. As shown in Fig. 8, the process starts by collecting the lymphocytes, which are separated in WBC segmentation. Then that image is converted to HSV colour model, since leukocyte nucleus are more contrasted in the S component of the HSV colour model as shown Fig. 14 [10]. The images are converted to grayscale images and then to binary images. To obtain the binary images used Otsu Thresholding [10] technique. Blob detection is applied and the diameter of the leukocyte nucleus is measured.



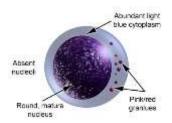




Fig. 10. Normal Leukocyte

Scant cytoses with High N:C without granules

Less prominent, variable, immanus promatine or control of the con



Fig. 13. Original input image

Fig. 11. Lymphoblast cell

The intermediate result images obtained are shown below. The original image is shown in Fig. 14.





Fig. 14. Saturation component image

Fig. 15. Grayscale image

Saturation component image which is selected for processing is shown in Fig. 14. In Fig. 15 illustrates gray scale image and Fig. 16 indicates a binary image gained by thresholding algorithm. In the binary image, final segmented nucleus is shown. After smearing Blob Detection output image is shown in Fig. 17. These obtained nuclei are further used to calculate diameter and to lymphocyte types.





Fig. 16. Binary image

Fig. 17. Blob detected image

Then we should classify the lymphocyte type using the nucleus diameter. In fact, we have analyzed 200 Normal Leukocyte cells from 72 blood samples and found the diameter of each. Comparing those values, the diameter range of normal Leukocytes is calculated. Using that deliberate ranges we presumed the diameter ranges of mature leukocytes and Lymphoblast cells.

1) Used formulas to calculate diameter of cell nucleus:

DMM = digital microscope magnification

1 mm = 3.77 pixels 1 micron = 0.001 mm Radius = Diameter / 2Perimeter = $2 \times \pi \times Radius$

Area =
$$\pi \times (Radius)^2$$

Diameter = $\underline{(diameter\ in\ Pixels) \times 1000}$
 $\underline{(3.77 \times DMM)}$ (2)

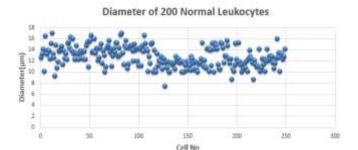


Fig. 18. Analysis result

- Normal lymphocyte nucleus diameter between 7.5 and 17.0 micro meters.
- Lymphoblast cell nucleus diameter is greater than 17.0 micro meters.
- Mature lymphocyte cell nucleus diameter less than 7.5 micro meters

According to the above diameter ranges, the number of lymphocyte cells in a blood sample image can be calculated to classify the disease.

D. AML and CML Detection

- 1) Chronic Myeloid Leukaemia: It's a disease in which the bone marrow produces too many white blood cells [18]. Chronic myelogenous leukaemia (also called CML or chronic granulocytic leukaemia) is a slowly progressing blood and bone marrow disease that usually occurs during or after middle age, and rarely occurs in children [18].
- 2) Acute Myeloid Leukaemia (AML): does not ordinarily form tumours. It is in all of the bone marrow in the body and, in some cases, spreads to other organs, such as the liver and spleen [18]. Therefore AML is not operated like most other cancers. The outlook for a person with AML depends on other information, such as the subtype of AML (determined by lab tests), the patient's age, and other lab test results. The first trace to a diagnosis of AML is typically an abnormal result on a complete blood count. While an excess of abnormal white blood cells (leucocytosis) is a common finding.
- *3) Feature Extraction:* It's a difficult step in image processing which outlines a large set of redundant data into a set of features of reduced dimension. Features extracted are Multi-resolution Texture, Area, Colour, and Overall Similarity Measure [19].

Shape Feature: Haematologists believed [20] that the important feature for classification is the shape of the nucleus. The region and boundary based shape features are extracted in

order to analyze the shape of the nucleus. The shape features are area and compactness. Shape based features is used to detect different shapes like circle, rectangle, ellipse, squares etc. Our blood cells also have different size and shapes. So to detect geometrical shape of cells, this shape based features are very useful method to detect different type of cells and their shapes.

Colour Feature: An important feature that human perceive while visualizing [12] is the colour. This feature is considered for extraction from nucleus region. While in AML, [13] the blasts are larger and irregular form and usually multiple nucleoli with the presence of Auer rode [14].

- 4) Nucleus and cytoplasm selection: The WBC identified in above WBC classifier can now be used to extract the nucleus and cytoplasm. To carry out this step, Blob detection [13] methods have been used. In order to overcome problems discussed previous level we need to measure the size of the cells we input to the unit. Separate nucleus and cytoplasm of all cells. Blob detection [16] and channel filtering algorithms has been used [21].
- 5) Blob Detection: The circle detection algorithm is based on the finding, that all circle's edge points have the same distance to its centre, which matches to circle's radius [16].
- 6) Channel Filtering: The filter organizes colour channels' filtering by filling with definite values, which are inside/outside of the quantified value's range. The filter allows filling certain ranges of RGB colour channels with specified value. The filter accepts twenty two and thirty two BPP (Bits Per Pixel) colour images for processing [21]. For classification of the abnormal cells with their particular type and subtypes, observation is carried out to detect the abnormalities in the nucleus or the cytoplasm of the cells. This identification is very significant and the diagnosis to determine the risk level of the patient is provided via generating a standard report. Blob detected cell and separated cytoplasm and nucleus are shown in Fig. 19.





Fig. 19. Blob detected cell and separated cytoplasm and nucleus

7) Image Classification: The primary intention is in this section is to detect AML and CML using geographical feature extraction and colour based feature extraction. In this final phase, the features extracted are used to provide the final reaction. All the features extracted are listed into the different columns with their values. Once the classified WBC images are given as an input to the proposed system, then the feature values are calculated. The values of the test image features are checked with the parameters implemented. Three parameters are being used to detect abnormalities. Whole cell size (area), Nucleus size, a rage of Nucleus and cytoplasm.

III. RESULTS

TABLE 2
ACCURACY TEST RESULTS FOR CELL COUNTING

Cell Type	No of Samples	Correct	Incorrect	Accuracy (%)
WBC	15	15	0	100
RBC	15	13	2	86.7
PLT	15	11	4	73.3
Overall	45	39	6	86.7

The WBC classifier was tested against two aspects, speed and accuracy using 15 sample images. The following two tests were run to classify 11 WBCs, with 130 training images. Both tests consumed up to 28MBs of RAM.

TABLE 3
CLASSIFIER PERFORMANCE TEST RESULTS

Task	Without cache	With cache
Segment WBCs from mask and input	0.240	0.254
Time to Learn	29.942	0.866
Time to classify and finish	0.321	0.375
Total Time	31.998	1.495

A test was carried out for the accuracy of the classifier with WBC cells per each category. The WBCs were obtained by extracting them from the ALL-IDB [1] blood sample images set rather than individual cells.

TABLE 4
CLASSIFIER ACCURACY TEST RESULTS

WBC Type	Count	Correct	Incorrect	Accuracy (%)
Basophil	40	32	8	77.5
Eosinophil	25	22	3	84.0
Neutrophil	60	48	12	85.0
Monocyte	25	17	8	68.0
Lymphocyte	100	93	7	79.0
Overall	250	212	38	84.8

 ${\bf TABLE~5}$ ${\bf Measured~Cell~Diameters~for~ALL/CLL~Detection}$

WBC Type	Size	
Eosinophil	12 to 15 μm in diameter	
Basophil	9 to 10 μm in diameter	
Lymphocyte	8 to 10 μm in diameter	

Detection is applied to 85 images of blood smear under ALL and CLL. Image data set details of [1]

- Magnification of the microscope: 300 to 500
- Image format: JPG,24 bit,1712 x 1368 pixels

 $\label{eq:table 6} Table \ 6$ Test Performance to identify Leukocyte type

Task	Time (s)
Time to identify leukocyte type of one blood sample image	11.44

TABLE 7 ACCRACY TEST RESULT FOR ALL AND CLL DETECTION

Sample	No. of	No. of Identifications		A 001140011
Type	Samples	Correct	Incorrect	Accuracy
ALL (+)	21	20	1	95%
CLL (+)	14	13	1	93%
Normal	50	45	12	90%
Overall	85	78	7	92%

TABLE 8
INFECTED CELLS IDENTIFIED BY THE SYSTEM

	Basophile	Eosinophile	Lymphocyte
Healthy		67	
Infected		13	

IV. CONCLUSIONS

For the cell count, testing was done by taking five images for each patient. The system resulted in 86.7% accuracy of the tested images and it only took 8 seconds to give the result. For the classifier, since the training data is 130 images and test data is 250 images the trained sample to test sample ratio is higher, for more accuracy, the training sample has to be very high in count. It is proved that the use of cache is a very good approach for the classifier hence it saves 95.32% of your time (30.5 seconds). We detect ALL and CLL from digital microscope images using image processing techniques followed by Morphological feature extractions. The proposed method, mainly considered shape features like diameter, radius, area and perimeter of the cell nucleus for the identify type of leukocyte. The system contributed 91.76% accuracy to tested image data set. The next stage is anticipated to improve the accuracy level up to 100% with the help from the medical division in Sri Lanka to introduce a fully automated system to Cancer hospitals.

ACKNOWLEDGMENT

We are very grateful to our supervisor Ms. Sanjeevi Chandrasiri who is a Lecturer in Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka, for the immense support and guidance throughout this research. Furthermore we are thankful to Mr. Fabio Scotti of University of Milan, Italy for providing an image database [1] within a very short time.

REFERENCES

[1] S. Fabio, L. D. Ruggero and P. Vincenzo, "ALL-IDB Acute Lymphoblastic Leukemia Image Database for Image Processing," Department of Information Technology - Università degli Studi di

- Milano, [Online]. Available: http://homes.di.unimi.it/scotti/all/. [Accessed 30 08 2016].
- [2] N. Theera-Umpon and S. Dhompongsa, Morphological granulometric features of nucleus in automatic bone marrow white blood cell classification, IEEE Trans Inf Technol Biomed, 2007.
- [3] ADNAN KHASHMAN and ESAM AL-ZGOUL, Image Segmentation of Blood Cells in Leukemia Patients, vol. I.
- [4] "AForge.NET :: Computer Vision, Artificial Intelligence, Robotics," [Online]. Available: http://www.aforgenet.com/. [Accessed 30 08 2016].
- [5] "Platelet Count Using Image Processing" being submitted by Rakshita Tandon, Rohan Sharma and Sahil Goel of the Department of Computer Engineering, Netaji Subhas Institute of Technology, New Delhi, 2nd June. 2014.
- [6] "Red Blood Cells Classification using Image Processing" Navin D. Jambhekar , Department of Computer Science. ISSN: 2249-7846 (Online) pp 151-154, Nov. 2011.
- [7] "Automated Red Blood Cells Counting in Peripheral Blood Smear Image Using Circular Hough Transform" Mohd Azhar Abdul Razak, Department of Electronic and Computer Engineering Faculty of Electrical Engineering. 2014
- [8] "Counting of Wbcs and Rbcs from Blood Images Using Gray Thresholding", Pooja R. Patil, G.S. Sable, Gauri Anandgaonkar International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308
- [9] Hayan Tareq Abdul Wahhab, "Classification Of Acute Leukemia Using Image Processing And Machine Learning Techniques" Phd Thesis, Faculty Of Computer Science & Information Technology University Of Malaya, Kuala Lumpur, 2015. [Accessed Feb.29 2016]
- [10] Trupti A. Kulkarni-Joshi, Dilip S. Bhosale, "A Fast Segmentation Scheme for Acute Lymphoblastic Leukemia Detection," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.3,no.2,Feb 2320 – 3765,2014
- [11] Esti Suryani, Wiharto, Nizomjon Polvonov, "Identification and Counting White Blood Cells and Red Blood Cells using Image Processing Case Study of Leukemia," International journal of Computer Science & Network Solutions, Vol. 2, no. 6 June 2014
- [12] T. Markiewicz, S. Osowski, B., Marianska, L., Moszczynski, "Automatic Recognition of the Blood Cells of Myelogenous Leukemia Using SVM", Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 – August 4, 2005, pp. 2496-2501.
- [13] N., H., A., Halim, M., Y., Mashor, R., Hassan, "Automatic Blasts Counting for Acute Leukemia Based on Blood Samples", International Journal of Research and Reviews in Computer Science, Vol. 2, No. 4, August 2011, pp. 971-976.
- [14] F., Sadeghian, Z., Seman, A., R., Ramli, B., H., A., 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, 2009, pp. 196-206.
- [15] V., Piurri, F., Scotti, "Morphological Classification of Blood Leukocytes by Microscope Images", IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Boston, MA, USA, 14 – 16 July 2004, pp. 103-108
- [16] "AForge.NET :: Framework Features Blobs Processing", Aforgenet.com, 2016. [Online]. Available: http://www.aforgenet.com/framework/features/blobs_processing.html. [Accessed: 30- Sep- 2016].
- [17] S. Chandrasiri and P. Samarasinghe, "Morphology Based Automatic Disease Analysis Through Evaluation of Red Blood Cells," 2014.
- [18] "Myeloid leukemia", *Wikipedia*, 2016. [Online]. Available: https://en.wikipedia.org/wiki/Myeloid_leukemia. [Accessed: 30- Sep-2016]
- [19] Chary, R., Lakshmi, D.R. and Sunitha, K.V.N., 2012. Feature extraction methods for color image similarity. arXiv preprint arXiv:1204.2336.
- [20] S. R. N. J. Monika Jain, "Detection of Cervical Cancer Using Image Processing Tools," Greater Noida.
- [21] "ChannelFiltering Class", Aforgenet.com, 2016. [Online]. Available: http://www.aforgenet.com/framework/docs/html/e053df99-306d-1bbe-52ec-44f327fe789b.htm. [Accessed: 30- Sep- 2016].