10.14445/23488549/IJECE-ICCREST-P102

SSRG International Journal of Electronics and Communication Engineering (SSRG – IJECE) – Special Issue ICCREST Mar 2019

Detection of Acute Lymphocytic Leukemia using Statistical Features

S. Hariprasath [#1], T. Dharani [*2], Dr. M. Santhi [#3]
[#1] Assistant Professor, PG Scholar[*2], Professor: HOD of ECE [#3] & Dept. of ECE
Saranathan College of Engineering, Trichy, India

Abstract

Accurate diagnosis of Leukemia is an important issue in the medical field in order to provide effective treatment to the patient. Leukemia is caused due to the abnormalities in the lymphatic (immune) system in our body. The chance of getting affected by leukemia is more common in children than adults. Cell types that involved in leukemia are white blood cells which are potent infection fighters. When leukemia causes abnormal production of WBCs which do not function properly is known as Acute Lymphocytic Leukemia (ALL) whereas, in other type of leukemia called Chronic Lymphocytic Leukemia (CLL), immature WBCs are capable of performing their functions normally. When compared to Chronic Leukemias, Acute Leukemias are more hazardous. In this paper Acute Lymphocytic Leukemia (ALL) is focused. In this work, from the given dataset that consisting of both benign (healthy) & malignant cells, Leukemic cells are detected & classified based on blast cells' morphology. There are various image processing techniques to detect leukemia and its types. Linear, SVM and classifiers are analyzed. SVM-R produced accuracy, sensitivity and specificity of 86.67%, 85%, and 90% respectively for noisy data.

Keywords - *Leukemia, Pattern Classification system, linear classifier, SVM.*

I. INTRODUCTION

White Blood Cells has the major contribution to the immune system of our body. It also helps in detecting diseases such as leukemia. The death rate in our country has been increasing in every year due to many hazardous diseases. Leukemia (Blood Cancer) is considered as the most threating disease since it may cause immediate death. It causes the production of abnormal white blood cells that are responsible for fighting diseases. Leukemic cells continue to grow, multiply & divide and results in decreasing the capability to fight against infections, control bleeding and transport oxygen [1].

Different types of leukemia depend on the type of blood cell that results in cancer. This classification is based on two systems namely: French-American-British (FAB) classification and World Health Organization (WHO) [2].When leukemia occurs in lymphocytes, it is known as lymphocytic leukemia. Hence further subtypes of lymphocytic leukemia are Acute Lymphocytic Leukemia (ALL) and Chronic Lymphocytic Leukemia (CLL). The other types of leukemia are known as Acute Myeloid Leukemia (AML) and Chronic Myeloid Leukemia (CML) that occurs due to abnormal myelocytes [3]. According to FAB classification, the subtypes of ALL & AML are L1, L2 and L3, and M1, M2, M3, M4, M5, M6 and M7 respectively.

Since Acute Leukemias are more hazardous, it is necessary to diagnose them at the early stage for medication [10]. By observing the blood cells from microscopic images, many diseases can be diagnosed and treated. In this work, only ALL is considered which affects cells called lymphocytes. ALL primarily affects children and adults over 50 years of age. ALL can affect children who are younger than 5 years of age and declines and begins to rise again after 50 years of age. Hence ALL can even leads to death when it is untreated. Early diagnosis of ALL is very crucial for a person's recovery. This diagnosis and detection of ALL is based on identification of lymphoblasts by microscopy. The best way to distinguish malignant cells from benign cells is to utilize the morphological features of a blast cell nucleus where immature white blood cells are known as 'blast cells' [4]. The discrimination between healthy and malignant cells can be made by using their cell structure. The differences in the cells structure of healthy and malignant cell is shown in figure 1.

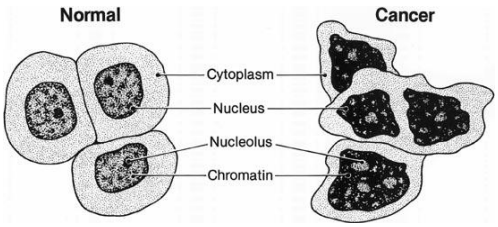

Figure 1. Difference between healthy and malignant cell structure

II. EXISTING SYSTEMS

Various morphological, shape and textural features have been taken into account for the detection and classification of a disease.

Commonly used cancer database for detection and classification is ALL-IDB. In 'ALL-IDB: the acute lymphoblastic leukemia image database for image processing' by Ruggero donida

labati, vincenzo piuri, fabio scotti, an automated method for detection of lymphoblast is done in the sequence of steps such as segmentation, identification of white cells, identification of lymphocytes [6]. The database used is ALL-IDB that has two distinct versions such as (ALL-IDB1 & ALL-IDB2).

Jyoti Rawat Annapurna Singh, H.S. Bhadauria, Jitendra Virmani, Jagtar Singh Devgun, has proposed a system that utilizes 331 features on 240 microscopic blood smear images (images from American Society of Hematology) to classify them into healthy and subtypes of Acute Leukemia by using a genetic algorithm based support vector machine classifier [7]. They also analyzed various kernel functions of SVM and achieved an accuracy of above 95%. However the system may get affected when larger dataset is used and irregular illumination of microscopic images occurs.

Subclasses of ALL (L1, L2) and subclasses of AML (M2, M3, and M5) are classified by Shaikh Mohammed Bilal and Sachin Deshpande. This method utilizes LAB color space by CIE for the dataset image. After the cell separation process, the texture specifications of the image is determined. It is done by using 'Wold's decomposition procedure'. It consists of three fields namely: harmonic field, stochastic field, evanescent field. Harmonic and evanescent fields define structural component of the image. To initialize the harmonic field, sine transformed by DFT or FFT is done [8].

To initialize evanescent field, Hough transform along with DFT is used. Texture specifications such as regularity, direction and random variations are considered. A 2-D Wold decomposition of homogeneous random fields is applied to retrieval of database image and appears to offer a perceptually more satisfying measure of similarity in patterns.

These phases differ in the amount of blast cells present. Chronic phase has blast count <10%, accelerated phase has blast count <20% and blast crisis phase is the most affected phase that has >20% of blast count. But in Acute leukemia there are no such phases can be detected.

Preeti Jagadev and Virani compared various image segmentation algorithms such as K means clustering algorithm, Marker controlled Watershed algorithm and HSV color based segmentation algorithm. In K means clustering algorithm, the value of k plays a major role. Generally, the value of k is taken as 3. The clusters taken are nucleus, background and other cells. To detect the distance between two clusters, 'Euclidean metric' is used. In addition to k means clustering, feature extraction techniques such as GLDM and GLCM that is Grey Level Difference and Grey Level Co-occurrence matrices can also be used [9].

Biji G and Hariharan have proposed a solution for an optimization problem for the WBC detection using Electromagnetism like Optimization (EMO) based circle detector [10]. EMO technique considers each solution as a charged particle and its charge is determined by objective function. The main phases of EMO algorithm are: initialization, local search, calculation of total force vector and movement. The detection of WBC is implemented as follows: Segmenting the WBCs using Diffused Expectation Maximization (DEM) algorithm. By morphological edge detection method, the edge map of segmented image is obtained. To identify WBCs, the parameter values for each circle is defined.

Madhloom et al, proposed a methodology in which the image segmentation addressed several key issues in blast cells segmentation including, the blast cell localization sub-imaging, color variation and segregation of touching cells [11].

Each blast cell is segmented into the nucleus and the cytoplasm. This stage produces two outputs: (i) a sub image of the blast cell extracted and placed on a white background, (ii) a nucleus sub image extracted from the blast cell sub-image. The determined blast cell and its nucleus are the regions of interest (ROI) to be analyzed in the succeeding stages of the research.

III. METHODLOGY

The steps involved in the pattern classification system is given in figure 2. Each step is explained in following sections.

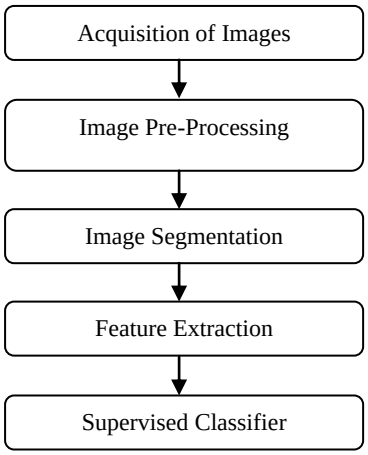

Figure 2. General Processing steps in Pattern classification System

A. Image Acquisition

Initial step in the process is to collect the dataset of leucocytes. Dataset used in this methodology is ALL-IDB which contains images of the blood smears of leukemic persons and images of the blood smears of non-leukemic persons. The database ALL IDB consists of ALL_IDB1 and ALL_IDB2 datasets consisting of pictures that were captured by optical laboratory magnifier as well as a Canon Power Shot G5 camera. The resolution of the JPEG format 24 bit pictures is 2592 x 194. ALL_IDB

consists of 100 pictures (35 healthy images and 65 malignant images) having 39000 blood components in which the lymphocytes are labeled by skilled oncologists [6]. The sample images for benign and malignant cell is given in figure 3.

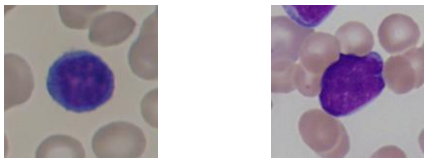

Figure 3. Samples from ALL-IDB for healthy and malignant cells

B. Image Pre-processing

The principle objective of image enhancement technique is to process an image so that the resultant image is more suitable than the original for a particular application. The steps involved in image pre-processing before feature extraction is shown in figure 5. WBC identification was made possible by conversion to the CMYK color model. In fact, leucocytes are more contrasted in the Y component of CMYK color model because the yellow color is present in all elements of the image, except leucocytes. Redistribution of image grey levels is necessary to make the subsequent segmentation process easier.

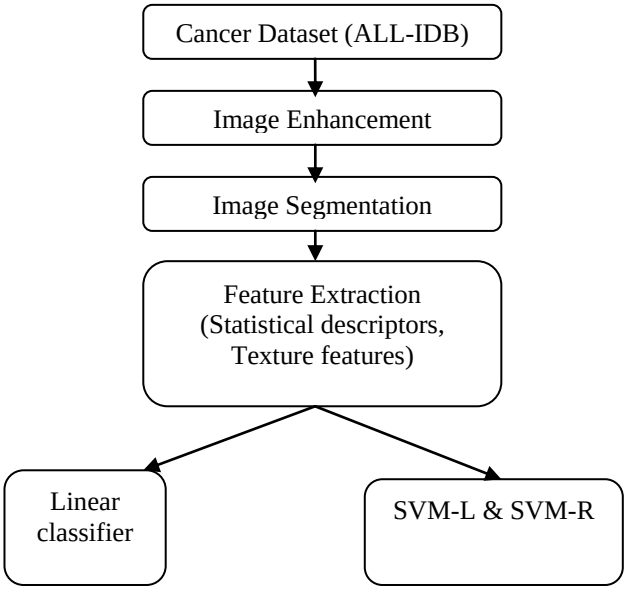

Figure 4. Detection and classification of ALL

C. Image Segmentation

Then, to enhance the contrast of the image histogram equalization or contrast stretching can be used by adjusting the image intensity. Segmentation is achieved using an automatically calculated threshold. Here, the threshold value based on the triangle method or Zack algorithm. The triangle method is applied to the image histogram, resulting in a straight line that connects the highest histogram value (h[b max]) and the lowest histogram value (h[b min]), where b max and b min indicate the values of the grey levels where the histogram h[x] reaches its maximum and minimum, respectively.

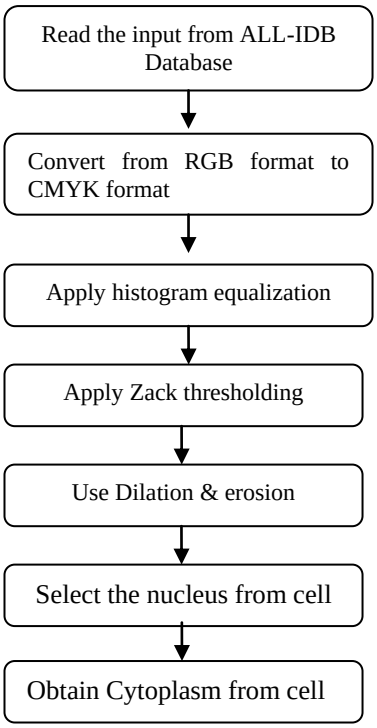

Figure 4. Image Pre-processing & Image Segmentation

Then, the distance (d) between the marked line and the histogram values between b min and b max is calculated. The threshold value is fixed where the distance (d) reaches its maximum at a particular intensity value. This algorithm is particularly effective when the histogram displays clear valleys between high and weak peaks present in the Y component histograms generated from leucocytes and red blood cells. Area opening is done in order to clean up the image by deleting all of the objects with as size lesser than the structuring element. The two morphological operations involved are erosion and dilation. In dilation, the value of pixel in output is maximum for all pixels in the input pixel's neighborhood whereas in erosion, the value of pixel in output is minimum for pixels in the input pixel's neighborhood whereas

Nucleus selection:

The nucleus selection approach takes advantage in which WBC nuclei are more in contrast on the green component of the RGB color space. However, in this color space, the threshold operation described by Otsu does not produce clean results, especially in the presence of granulocytes, because granules are selected erroneously as part of the nucleus. To avoid this issue, the binary image obtained from the green component is combined with

the binary image obtained from the a* component of the CIE Lab color space via threshold operation [9].

Finally, to obtain the cytoplasm, a subtraction operation is performed between the binary image containing the whole leucocyte and the image containing only the nucleus.

D. Feature Extraction:

Since discrimination of benign and malignant cells purely depends upon the cell morphology, features are extracted from the nucleus of the leucocytes. Features and their equations are given in table 1.

1. Shape descriptors

The following nine shape descriptors were extracted from each blast cell and its nucleus [12]:

Area: It is represented by the number of pixels in the ROI. This feature was used to quantify the size of the ROI.

Eccentricity: The Eccentricity represents the ratio of the length ***L*** and width ***W*** of the minimal bounding box of the ROI, that is, it measures the degree in which the blast cell and its nucleus resemble an ellipse.

Elongation: Another measure which can be calculated from the bounding box is elongation. It describes the extent of elongation of the blast cell and its nucleus.

Solidity: A solidity value of 1 signifies a solid object, and a value less than 1 signifies an object with an irregular boundary. The solidity value used for the threshold is calculated directly from the image containing only the individual leucocytes, and when this image is empty, a default value of 0.90 is used.

Circularity*:* Circularity as the ratio of the perimeter squared to the area. This measure describes how much the ROI is similar to a circle and it reflects the complexity of the object boundary.

Perimeter*:* Number of boundary pixels that belong to an object. This measure quantifies the distance of the outside boundary of the blast cell or its nucleus.

Rectangularity*:* Rectangularity is represented as the ratio of the ROI's area to the area of its minimum bounding box. This measure describes how much the ROI is similar to a rectangle.

Roundness: It shows the deviation in the shape of the cells.

Orientation: It is used to identify the symmetry of image intensity.

2. Statistical descriptors

Shape based features may be susceptible to errors in segmentation. Thus, these descriptors are used together with regional descriptors, which are less susceptible to errors. Among these are statistical descriptors, which are the most discriminatory features of blood cells [13]. The statistical descriptors are also known as color descriptors such as mean, standard deviation, skew ness, kurtosis and entropy, which are calculated from sub images in shades of grey.

Mean- Mean is the average value of pixels within the region of interest that represents the brightness of the image.

Entropy- It is used to measure the randomness or disorder of an image.

Standard Deviation- By using the Standard Deviation we can determine a way of analyzing what is normal, extra-large or extra-small.

Skewness- It measures lack of symmetry. The zero value indicates that the distribution of the intensity values is relatively equal on both sides of the mean.

Kurtosis- Measures the peak of the distribution of the intensity values around the mean.

3. Texture descriptors

To examine the texture of an object, gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. It calculates how often pixel pairs with specific values and in a specified spatial relationship occurs. 20 GLCM features such as auto correlation, contrast, correlation, entropy, energy, homogeneity etc. are extracted from the cell's nucleus [14]. The results of features are shown in the table 3.

TABLE I Features Extracted

S.No.	Feature extracted	Mathematical formula
1	Eccentricity	$1-\frac{minor\ axis}{major\ axis}$
2	Elongation	$\frac{\sqrt{(majoraxis^2 - minoraxis^2)}}{major\ axis}$
3	Solidity	$Convex\_Area/Area$
4	Circularity	$\frac{1}{(4*\pi*Area)/(perimeter)^2}$
5	Rectangularity	$\frac{area}{majoraxis*minoraxis}$

6	Roundness	$(4 * \pi * Area)/(perimeter)^2$
7	Skewness	$\frac{E(X-\mu)^3}{\sigma^3}$
8	Kurtosis	$\frac{E(X-\mu)^4}{\sigma^4}$
9	Correlation	$\frac{\sum_i \sum_j (ij)\, p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
10	Entropy	$-\sum_i \sum_j p(i,j) \log(p(i,j))$
11	Energy	$\sum_{n=0}^{Ng-1} n^2 \{\sum_{i=1}^{Ng-1} p(i,j)\}$
12	Sum of variance	$\sum_{i=2}^{2Ng} (i - fs)^2 p_{x+y\,(i)}$

E. Classification

The given set of samples can be classified into healthy and malignant classes by using a classifier at the end. This classifier can be a linear classifier, SVM classifier [15].

1. Linear classifier

Initially only 15 features (shape and color descriptors) are used for the detection and classification by a linear classifier. Since the initial task is to classify the image into two classes namely: healthy and malignant cells, a 'decision threshold' is used. From 15 extracted features, 5 features such as solidity, entropy, circularity, standard deviation and Skewness are selected in order to detect the malignant cells from healthy cells.

2. Support Vector Machine Classifier

A Support Vector Machine is a supervised classifier and it is defined by a hyperplane. For a given set of training inputs, the classifier provides an optimal hyperplane that categorizes the new samples [16]. In 2D plane, a hyperplane is considered as a line that divides a plane into two parts. Mathematically hyperplane can be expressed as,

$$f(x) = \beta_0 + \beta(x)$$

Where β is a weight vector, β0 is a bias.

During implementation, bias is taken as 0.3344.Training samples that are closest to the hyperplane are known as support vectors. A set of mathematical functions that are used in SVM are called kernel. A kernel functions provides the required form of the given input. There are different types of kernels such as linear, nonlinear, Radial Basis Functions (RBF), Sigmoid. Widely used kernel function is RBF which has finite response along the x-axis. Formulas are given in table 2.

TABLE II Kernel Functions Of Svm

Kernel function	Formula
Linear	$G(x1,x2) = x1'x2$
RBF/ Gaussian	$G(x1,x2) = \exp(-\lvert\lvert x1 - x2\rvert\rvert)^2$

IV. IMPLEMENTATION DETAILS

From figure 2, in image pre-processing and segmentation steps after converting the input image from RGB format to CMYK format, histogram equalization is applied. After analyzing various values, threshold is fixed as 0.3608. By zack / triangle method, threshold is fixed as 0.5664 for thresholding the images.

To obtain optimal results using linear classifier, threshold values calculated from the selected features are adjusted and various performance metrics are calculated.

To fix the threshold values:

- No. of healthy images taken =20
- No. of malignant images taken=40

Testing (for various threshold values):

- No. of healthy images taken =15
- No. of malignant images taken=25

If the features of a given input image satisfies the decision threshold, it is classified as a malignant one. The results of linear classifier is given in table 4.In order to validate a SVM classifier, Cross fold validation is used.

TABLE III Texture Features For Benign And Malignant Samples

Type of cell	Auto Correlation	Contrast	Energy	Entropy	Homogeneity
Benign	20.24	0.0298	0.4417	1.196	0.9851
Malignant	24.12	0.0483	0.2595	1.623	0.9701

K-fold cross validation:

It is a resampling procedure which is used to estimate the skill of machine learning models for new data.

Where k is the number of groups that a given sample is to be split into. Generally k is chosen as 5 or 10.

V. EXPERIMENTAL RESULTS

To calculate the performance metrics such as accuracy, sensitivity and specificity, TP, FP, TN and FN are needed to be calculated.

True positive (TP) = correct identification of class as malignant.
False positive (FP) = incorrect identification of class as malignant.
True negative (TN) = correct identification of class as healthy.
False negative (FN) = incorrect identification of class as healthy.

Accuracy: The accuracy of a test is its ability to differentiate the leukemic and healthy cases correctly. Mathematically, this can be stated as

$$Accuracy = TP + TN / (TP + TN + FP + FN). \quad (1)$$

Sensitivity: The sensitivity of a test is its ability to determine the leukemic cases correctly. Mathematically, this can be stated as:

$$Sensitivity = TP / (TP + FN). \quad (2)$$

Specificity: The specificity is used to determine the healthy cases correctly. This can be stated as:

$$Specificity = TN / (TN + FP) \quad (3)$$

TABLE IV Performance Metrics Of A Linear Classifier For Various Threshold Values.

Threshold fixed	Sensitivity	Specificity	Accuracy
Solidity<0.8, Skewness (mean)< -0.3, Circularity >1, Roundness < 0.8, Entropy >5	8%	86%	33%
Solidity<0.88, Skewness (mean)< -0.3, Circularity >1, Roundness < 0.8, Entropy >5	64%	60%	56%
Solidity<0.88, Skewness (mean)< -0.3, Circularity >1.3, Roundness < 0.8, Entropy >5	64%	86%	58%
Solidity<0.88, Standard deviation >35, Circularity >1.3, Roundness < 0.8, Entropy >5	72%	86%	68%
Solidity<0.88, Standard deviation >40, Circularity >1.3, Roundness < 0.8, Entropy >5	76%	86%	71.1%

TABLE V Performance Metrics For Svm-L, Svm-R.

SVM kernel function	K value	Sensitivity	Specificity	Accuracy	K-Fold loss
RBF	4	85%	90%	86.67%	0.0125
	5	85%	90%	86.67%	0.0225
	10	90%	92.4%	90%	0.01435
Linear	4	80%	75.65%	76.67%	0.0143
	5	10%	55%	65%	0.0286
	10	10%	55%	65%	0.0286

TABLE VI Performance Metrics For Various Classifiers (Noise Free Data)

Classifier	Sensitivity	Specificity	Accuracy	K-Fold Loss
Linear	76%	86%	71.1%	--------
SVM-L	10%	55%	65%	0.0286
SVM-R	85%	90%	86.67%	0.0125

Table 5 & 6 shows the results of SVM for linear and RBF kernel functions. Since the sample points are permutated before classification for each iteration, the results may get varied. . SVM-R produced accuracy, sensitivity and specificity of 93.5%, 90%, and 92% respectively for noisy data.

VI. CONCLUSION

In the proposed work, morphological and statistical features of blast cell is analyzed to discriminate benign and leukemic cells. For classification, SVM-R, SVM-L and KNN classifiers are used and compared. Obtained results shows that SVM-R produced better results only when noise free data is used. The future scope of this work is to develop a classification system that classifies the leukemic cells (ALL) into its subtypes-L1, L2 and L3. For this subtype classification, both nucleus and cytoplasmic features are needed to be extracted.

ACKNOWLEDGMENT

We would like to thank **N. Shaikh Mohammed Bilal, M.E., Assistant Professor,** Department of Computer Science Engineering, KJ Somaiya College of Engineering and Mumbai for providing the data sets and guidance during the project work.

REFERENCES

[1] https://www.cancercenter.com
[2] http://www.hematology.org/Patients/Cancers/Leukemia.aspx.
[3] https://www.verywellhealth.com/understanding-white-blood-cells-and-counts-2249217
[4] https://www.mayoclinic.org/diseases-conditions/acute-lymphocytic-leukemia/symptoms-causes/syc-20369077
[5] https://www.cancer.gov/types/leukemia
[6] Ruggero Donida Labiti, Vincenzo Piuri, Fabio Scotti "ALL-IDB Acute Lymphoblastic Leukemia Image Database for image processing". IEEE Transactions, Pg. No. 2045-2048, (2011).
[7] Jyoti Rawat, Annapurna Singh, H.S. Bhadauria, Jitendra Virmani, Jagtar Singh Devgun, "Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia", Bio cybernetics and Biomedical Engineering Journal, Volume 37, Issue 4, July 2017, Pages 637-654.
[8] Shaikh Mohammed Bilal N, Sachin Deshpande "Computer aided leukemia detection using image processing techniques". International Conference on Recent Trends in Electronics Information & communication technology (RTEICT), May 19-20, (2017).
[9] Preeti Jagadev, H. G. Virani (2017). "Detection of Leukemia and its types using image processing and Machine Learning". International conference on Trends in Electronics and Informatics (ICEI).
[10] Biji G, Hariharan S (2017), "An Efficient Peripheral Blood Smear Image Analysis Technique for Leukemia Detection", International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud).
[11] Madhloom, H Kareem, S. Ariffin, H. (2012). An Image Processing Application for the Localization and Segmentation of Lymphoblast Cell Using Peripheral Blood Images. Journal of Medical Systems, 36(4).
[12] Himali Vaghela, Hardik Modi, Manoj Pandiya, Potdar M.B, "A Novel Approach to detect chronic leukemia using shape based feature extraction and identification with digital image processing. International Journal Applied Information Systems (IJAIS). Volume 11, No.5 (2016).
[13] Himali Vaghela, Hardik Modi, Manoj Pandiya, Potdar M.B "Leukemia Detection using Digital Image Processing Techniques". International Journal of Applied Information Systems (IJAIS), Volume 10-No.1 (2015).
[14] Preetham Kumar and Shazad Maneck Udwadia, "Automatic Detection of Acute Myeloid Leukemia from Microscopic Blood Smear Image", International Conference on Advances in Computing, Communications and Informatics, ICACCI (2017)
[15] Jyoti Rawat, Annapurna Singh, H.S. Bhadauria, Jitendra Virmani, Jagtar Singh Devgun, "Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia", Bio cybernetics and Biomedical Engineering Journal, Volume 37, Issue 4, July 2017, Pages 637-654.
[16] Kumar PS and Vasuki S, "Automated diagnosis of acute lymphocytic leukemia and acute myeloid leukemia using multi-SV", J Biomedical Image Bioengineering, (2017). Volume 1 Issue 1