Applications of Machine Learning and Image Processing Techniques in the Detection of Leukemia

Siddhika Arunachalam

Dept. of Computer Engineering, Sardar Patel Institute of Technology, University of Mumbai, Mumbai, India

Corresponding Author: siddhi97am@gmail.com,   Tel.: +919969286067

Abstract— The cancer of blood and bone marrow is called Leukemia. It is a result of the uncontrolled reproduction of immature white blood cells. It hampers the ability of the body to fight infection. In Leukemia, the white blood cells (WBC) are generally affected. There are different types of Leukemia namely: Acute myelogenous Leukemia (AML), Acute lymphocytic Leukemia (ALL), Chronic myelogenous Leukemia (CML), Chronic lymphocytic Leukemia (CLL). This paper proposes to automate the Leukemia detection process using machine learning and different techniques of image processing. The dataset consists of images of blood smear which is of both Leukemia and non-Leukemia patients. K-means Clustering, Marker-Controlled Watershed segmentation, and HSV colour-based segmentation are the image segmentation algorithms that have been used. Various features from the segmented lymphocyte images are extracted since the structural components of normal and Leukemic lymphocytes differ remarkably. The SVM classifier, which is a machine learning technique, is used to further classify Leukemia into its different types. This paper aims at identifying Leukemia and determine its types whether it is AML, ALL, CML or CLL which takes the classification process one step further as a majority of the previous works have been restricted to just detection of Leukemia or classifying into few of the main subtypes. The proposed system is successfully implemented using MATLAB.

Keywords— Leukemia; Image Processing; Machine Learning; SVM Classifier; MATLAB

I. INTRODUCTION

There are mainly three types of cells in the blood: red blood cells that carry oxygen, white blood cells that fight infection and platelets that help blood clot. The bone marrow makes the greatest number of red blood cells out of the billions of cells which it makes every day. The ratio of red blood cells to white blood cells in the body is 1000:1. This means that there are 1000 red blood cells for every 1 white blood cell. When there is Leukemia, the body makes a greater number of white cells than it is required. As the normal white blood cells do, the Leukemia cells are unable to fight the infection. Because of their large numbers, they start to affect the normal functioning of the organs. The cell grows and divides more rapidly and continues to live when normal cells would die, due to certain abnormalities, which leads to a fewer number of healthy red blood cells, indicating the symptoms of Leukemia. Here, Fig. 1 (a) shows the bone marrow and different types of blood cells it produces. In the bone marrow, the blood cells are made. The spongy material that is present in the center of the bones and generates about 95 percent of the body's blood cells is the bone marrow [1]. There are different types of white blood cells such as lymphocyte, monocyte, eosinophil, basophil, and neurophil. Fig. 1 (b) shows how the blood cells that are formed in the bone marrow first start as a stem cell. The preliminary stage of all blood cells is a "stem cell". Many diverse blood cells evolve such as the white blood cells, red blood cells and the platelets as the stem cell grows.
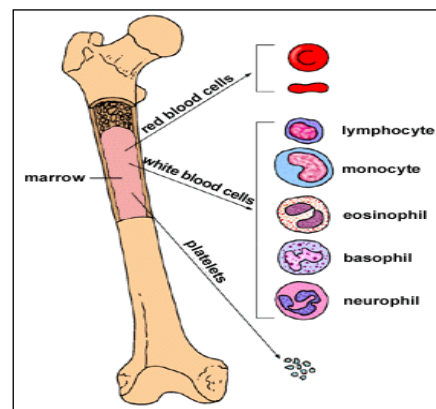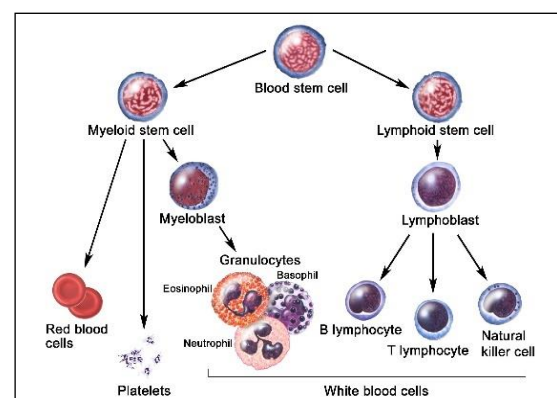

Fig.1. (a) Bone marrow and blood cells

Fig.1. (b) Process of formation of Myeloid and Lymphoid series of cell

Based on how fast Leukemia progresses, it is classified as:

- **Acute Leukemia**- The abnormal blood cells are generally the immature blood cells which are also referred to as blasts. The disease worsens quickly as these abnormal blood cells multiply rapidly and they are unable to carry out their normal functions. Timely and aggressive treatment is required for Acute Leukemia.

- **Chronic Leukemia**- More mature blood cells are involved in chronic Leukemia. These blood cells can function in a normal manner for a while and replicate more slowly and. Sometimes chronic Leukemia can go unnoticed or undiagnosed for years as it produces no early symptoms.

The major four types of Leukemia which will be classified using SVM classifier are as follows:

- **Acute lymphocytic Leukemia (ALL)**- It grows quickly and affects the lymphoid line of blood cells. It is more usual in juveniles.

- **Acute myelogenous Leukemia (AML)-** It affects the myeloid line of blood cells and grows quickly. It is more common in adults.

- **Chronic lymphocytic Leukemia (CLL)-** A large number of lymphocytes are made by bone marrow and it grows slowly. The patient may feel healthy for years without needing any treatment.

- **Chronic myelogenous Leukemia (CML)-** It generally grows slowly at the beginning and affects the myeloid line of the blood cells. In this type of Leukemia, a person may have little or nil symptoms for months or years. It mostly affects adults.

Fig. 2 shows how the patient with Leukemia show increased number of lymphocyte and monocytes in their blood smear samples as compared to the non-leukemic patients.
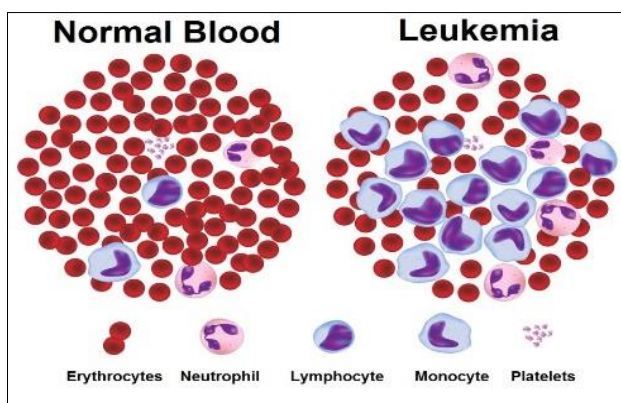

Fig. 2 The difference between normal blood cells and Leukemia blood cells

Microscopic study of human blood is used by the haematologist. This leads to the requirement of methods that include segmentation, microscopic colour imaging, clustering and classification. This allows the recognition of people afflicted from Leukemia. The process of inspecting the microscopic images visually by haematologist can be time-consuming, tiring and also prone to human error. Hence, the image processing and machine learning system is instantly needed that can conquer the related limitations in visual inspection. The prompt and quick detection of Leukemia will profoundly aid in providing the relevant treatment. The main objective of this paper is to implement machine learning and image processing techniques to detect Leukemia and its types. The organization of the paper is as follows: Section II contains the previous work, Section III contains the proposed work, Section IV provide experimental results and the conclusion and future work discussed in Section V.

II. PREVIOUS WORK

In [3], an automated Leukemia detection system is proposed where some filtering techniques and k-means clustering approach is used. Some features such as perimeter, mean, area, centroid, entropy, skewness, smoothness and standard deviation are extracted and calculated. To detect whether the cell has Leukemia or not, the neural network is used.

In [4] proposes a technique to automatically detect and classify Acute myeloid Leukemia (AML) blood smear images. For segmentation, the K-means clustering algorithm is used and for classification KNN, NN, and SVM algorithms are used. For texture description, the local binary pattern is used.

Different image processing techniques are applied to detect Leukemia in [5]. The number of methods used to detect immature cells is linear contrast stretching, histogram equalization, morphological techniques like area opening, area closing, dilation, erosion. Histogram equalization and Watershed transformation algorithms are also used.

In [6], a review of various computer-aided diagnosis techniques is discussed. The parameters that are mainly used for the review are various enhancement techniques, segmentation algorithms, various features extraction methods, classifiers and accuracy.

In [7], the main focus is on the geometry of cells like perimeter, area and statistical measures like a standard deviation that helps to separate the leukemic cells from other blood components using various processing tools. Then, after the statistical properties are identified, Leukemia is detected.

For the detection of Leukemia various geometrical attributes such as perimeter, area of the white blood cell nuclei is investigated in [8]. Statistical parameters such as standard deviation and mean are used for initial segmentation. Then, various image processing algorithms like image enhancement techniques, morphological operations, Thresholding and Labelling are implemented with the help of tools such as LabVIEW and MATLAB.

Geographical measures with different colour feature statistics are identified in [9], which is focused on supervised learning. A training model is also implemented in this research. A multi-class classifier was implemented to address missing features in the present white blood cell classifiers.

III.   PROPOSED WORK

Fig. 3 gives a stepwise process description for the effective detection and classification of Leukemia.
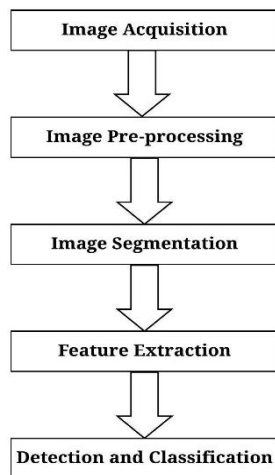

Fig. 3 Proposed process of Leukemia detection and classification

3.1 *Image Acquisition*

The images of the blood smears of Leukemic patients having all the main types of Leukemia and non-leukemic patients have been obtained from various online databases.

3.2 *Image Pre-processing*

The images that are obtained during the image acquisition phase may not be suitable for detection and classification purposes. Hence image pre-processing is performed to improve the image quality and make the image more appropriate for further processing steps. There may be some noise and blurriness in the image acquired. Using median filtering noise is removed from the image and using Wiener filter blurriness is eliminated from the image. The images procured were in the CMYK form. Therefore, these images are pre-processed and transformed into RGB form.

3.3 *Image Segmentation Algorithm*

Based on the features of the pixels in the image, an image is partitioned to several parts or regions. This process is identified as image segmentation. It is typically used to locate boundaries and objects in images. Some of the image segmentation algorithms used here are K-means clustering, Marker-controlled-watershed and HSV colour-based segmentation algorithm.

a)   *K-means Clustering Algorithm*

The interest area is segmented from the background in an image using K-means Clustering algorithm.

This clustering algorithm works in the proposed system as follows:

Step 1: The pre-processed image that is in RGB form is taken as an input in this step.
Step 2: The image is converted from RGB to grayscale as the RGB colour space is difficult to segment.
Step 3: Then the image is converted from grayscale to L*a*b* colour space. The 'L*' in lab space represents luminosity layer, 'a*' indicates the place where the colour falls along red-green axis and 'b*' represents where the colour falls along the blue-yellow axis. It helps to quantify the visual differences by distinguishing the colours from one another. 'a*' and 'b*' layers contain all the colour information.
Step 4: The value of k is considered as 3. These 3 clusters correspond to background (low saturation and high luminance), nucleus (high saturation) and other cells (e.g., cytoplasm, leukocyte and erythrocyte).
Step 5: Euclidean distance is used to measure the distance between the two clusters.
Step 6: Every pixel in the image is labeled with the cluster index.
Step 7: The objects can be separated by colour using pixel labels which will have an outcome in the form of three images.
Step 8: The dark blue cluster can be separated from the light blue cluster by using the 'L*' layer from the L*a*b* colour space. The cell nuclei have dark blue colour [11].
Step 9: Using the cluster center value, the index of the cluster which contains blue objects can be determined programmatically. By determining experimentally, it is found that the smallest cluster center value is of blue cluster.
Step 10: A mask is used to label the pixels belonging to the nuclei and the blue nuclei is displayed as a separate image.

b)   *Marker-Controlled   Watershed   Segmentation Algorithm*

It is mainly suitable and robust to segment objects that have closed contours, where the regions have been expressed by the boundaries. This segmentation is primarily used to separate the objects that are touching in the image. The binary image is used as a marker image for Watershed Segmentation. It typically works by identifying or "marking" the foreground objects and the background locations. This eventually helps to identify the boundary lines among the connected cells and effectively isolate all the connected nuclei from the image into its distinct nuclei.

The basic procedure for Marker-controlled watershed segmentation algorithm is as follows:

Step 1: The colour image is read and converted into grayscale.
Step 2: The Gradient Magnitude is used as the Segmentation Function.
Step 3: The Foreground objects are marked
Step 4: The Background Markers are computed.

Step 5: The Watershed Transform of the Segmentation function is computed and the result is visualized.

c) *HSV Colour-based Segmentation Algorithm*

HSV stands for Hue, Saturation and Value respectively. According to the model, Hue dimension denotes the "colour", Saturation dimension denotes the supremacy of that colour and the Value dimension denotes the brightness. Most of the WBC information is contained in the H component. The structure information of the WBC nucleus is covered in the S component. The image intensity is separated from the colour information in HSV. To select the colour tones in an image and apply a filter on the particular region, HSV colour space is used. This helps in eliminating dust, scratches and noise [12].

d) *Feature Extraction*

From the initial set of data, a number of features are extracted which helps to yield informative and non-redundant values. These features thus help in analyzing the results. The features that are mainly extracted are as follows:

- **Cell size**- It is the total number of pixels in the cytoplasm and nucleus area.
- **Mean-** It is the average value of pixels inside the region of interest (ROI) that characterizes the brightness of the image.
- **Variance-** The average of the squared differences from the Mean is called the variance.
- **Standard Deviation**- A way to analyze which cell is normal, extra small or extra-large is denoted by the standard deviation. The probability of the maximum likelihood of a pixel can also be analysed.
- **Entropy-** It is used for measuring the disorder or randomness of an image.
- **Correlation-** It represents the correlation between the neighborhood and its pixel values. How strongly pairs of variables are correlated are also indicated.
- **Skewness-** It is a measure of the deficiency of symmetry. On both sides of the mean, the distribution of the intensity values is relatively equal is indicated by the zero value.
- **Contrast-** The amount of local variations existing in an image is measured by contrast.
- **Kurtosis**- The peak of the intensity values about the mean is measured using Kurtosis.
- **Smoothness-** It facilitates the capturing of significant patterns in the data while excluding noise in the image.
- **Homogeneity:** It measures the degree of variation for the texture in an image. The greater the changes in gray values, the lesser the homogeneity. The homogeneity range is [0, 1].

e) *Classification*

The Support Vector Machine (SVM) classifier is used to classify the cells as leukemic and non-leukemic. The classification is accomplished by detecting the hyper-plane that separates the two classes. The kernel trick is used for nonlinear classification in SVM. The weights are updated using the training data while the classifier performance is validated using the testing data [10].

IV. EXPERIMENTAL RESULTS

The dataset contains images of blood smear. Fig. 4 (a) shows the output of segmentation obtained after the implementation of the K-means Clustering algorithm. Here, the value of K is taken as 3.

The output of the Marker-Controlled Watershed segmentation algorithm is shown in Fig. 4 (b) and the segmentation output achieved from the execution of the HSV colour-based segmentation algorithm is shown in Fig 4 (c).

After the image segmentation process using various methods mentioned above the feature extraction process is performed which is shown below in Table 1, 2, 3 respectively. The detection and classification process performed by the SVM classifier is shown in Fig. 5 (a), (b). If the Leukemia is detected, then the type of Leukemia is identified i.e. whether it is AML, ALL, CLL or CML.
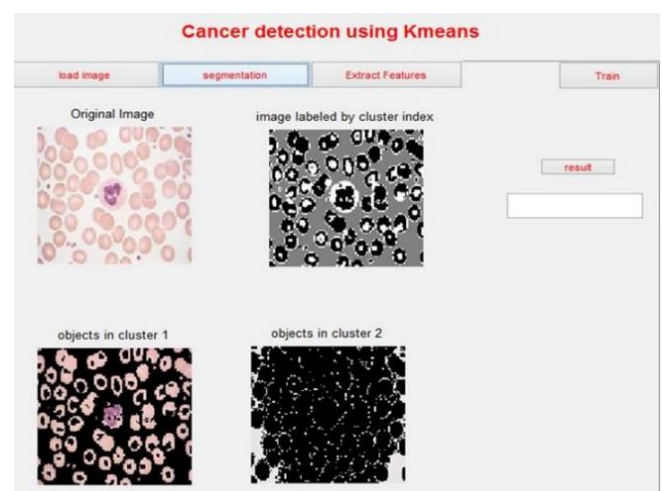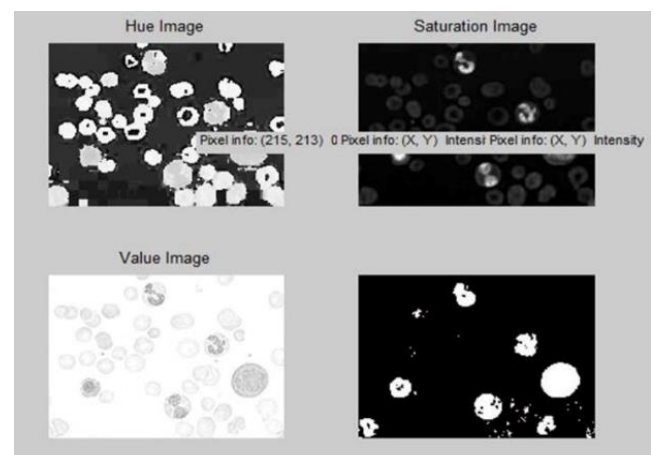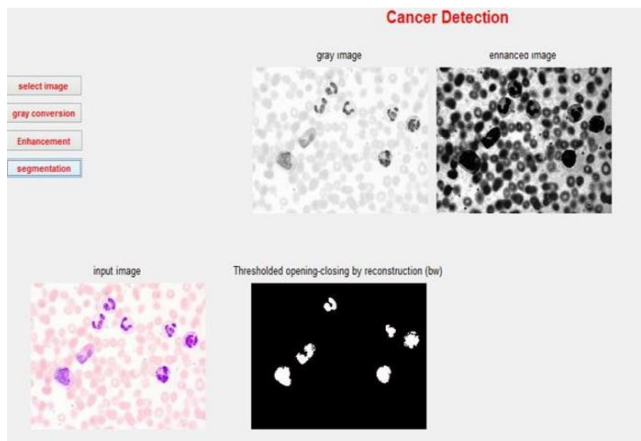

Fig. 4 (a) K-means Clustering algorithm used for the blood smear image
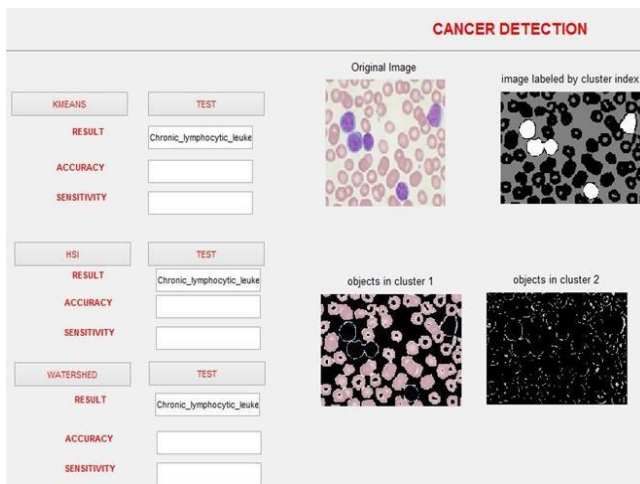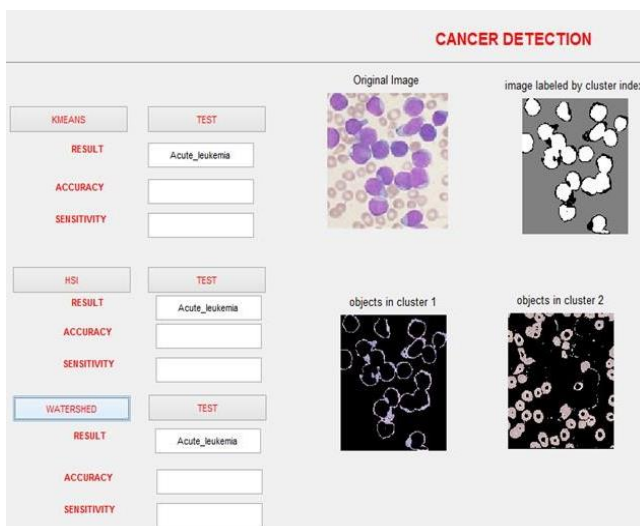

(b)

(c)

Fig. 4 (b) HSV colour-based segmentation algorithm used for the blood smear image, (c) Marker-Controlled Watershed Segmentation algorithm used for the blood smear image

(a)

(b)

Fig. 5 (a), (b) Detection and classification of Leukemia using SVM classifier

Table 1. Features extracted from the images segmented using K-means Clustering Algorithm

Feature Extracted	Blood Smear				
	Normal	AML	ALL	CML	CLL
Size	235.489	218.074	198.89	171.438	220.1755
Mean	22.09	42.7664	49.21	56.1297	42.8669
Variance	15.9375	15.9375	15.9375	15.9375	15.9375
Entropy	-0.921	-0.8449	-0.7684	-0.6718	-0.8537
Standard Deviation	-0.075	-0.0281	-0.0375	-0.0513	-0.0283
Smoothness	-6.0139	-2.6716	-1.1507	-0.1567	-2.9679
Skewness	1.00	1.00	1.00	1.00	1.00
Contrast	0.0053	0.0317	0.0338	0.0687	0.0180
Kurtosis	0.0704	0.0635	0.0607	0.0544	0.0638
Homogeneity	0.2945	0.4095	0.4402	0.4760	0.4160
Correlation	0	0	0	0	0

Table 2. Features extracted from the images segmented using Marker-Controlled Watershed Algorithm

Feature Extracted	Blood Smear				
	Normal	AML	ALL	CML	CLL
Size	235.489	218.074	198.8898	171.431	220.1755
Mean	22.09	42.7664	49.214	56.1297	42.8669
Variance	15.9375	15.9375	15.9375	15.9375	15.9375
Entropy	-0.921	-0.8449	-0.7684	-0.6718	-0.8537
Standard Deviation	-0.075	-0.0281	-0.0375	-0.0513	-0.0283
Smoothness	-6.0139	-2.6716	-1.150	-0.1567	-2.9679
Skewness	1.00	1.00	1.00	1.00	1.00
Contrast	0.0053	0.0317	0.0338	0.0687	0.0180
Kurtosis	0.0704	0.0635	0.0607	0.0544	0.0638
Homogeneity	0.2945	0.4095	0.4402	0.4760	0.4100
Correlation	0	0	0	0	0

Table 3. Features extracted from the images segmented using HSV colour-based segmentation Algorithm

Feature Extracted	Blood Smear				
	Normal	AML	ALL	CML	CLL
Size	235.5257	218.0911	198.8931	171.4466	220.1311
Mean	22.8690	44.0984	50.3336	57.0832	43.5155
Variance	15.9375	15.9375	15.9375	15.9375	15.9375
Entropy	-0.9175	-0.8451	-0.7653	-0.6720	-0.8536
Standard Deviation	-0.080	-0.0299	-0.0390	-0.0530	-0.0291
Smoothness	-6.1458	-3.0507	-1.1541	-0.1727	3.1923
Skewness	1.00	1.00	1.00	1.00	1.00
Contrast	0.0047	0.0317	0.0243	0.0611	0.0186
Kurtosis	0.0164	0.0650	0.0230	0.0274	0.0107
Homogeneity	0.2945	0.0119	0.4443	0.4798	0.4131
Correlation	0	0	0	0	0

V.   CONCLUSION AND FUTURE SCOPE

The paper primarily concentrates on the detection of Leukemia and provides a wider range of classification of Leukemia into its main four types i.e. AML, ALL, CLL or CML. K-means Clustering algorithm, HSV colour-based segmentation algorithm and Marker-Controlled Watershed Segmentation algorithm are the three image segmentation algorithms that were implemented. A number of features were also extracted namely Cell size, Mean, Variance, Standard Deviation, Entropy, Correlation, Skewness, Contrast, Kurtosis, Smoothness and Homogeneity so that the detection process is more precise and accurate. This work can further be expanded by detecting the subtypes of the main types of Leukemia. For example, one of the subtypes of AML is AML M3. The other segmentation algorithms can also be examined to improve the results as compared to the earlier ones.

VI.   REFERENCES

[1] Tathagata Hazra, Mrinal Kumar, Sanjaya Shankar Tripathy, "Automatic Leukemia Detection Using Image Processing Technique", *International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS)*, Vol. **VI**, Issue **IV**, April 2017, pp **42**.

[2] Sachin Paswan, Yogesh Kumar Rathore, "Detection and Classification of Blood Cancer from Microscopic Cell Images Using SVM KNN and NN Classifier", *International Journal of Advanced Research, Ideas and Innovations in Technology*, Vol. **3,** Issue **6**.

[3] Jameela Ali, AbdulRahim Ahmad, Loay E. George, Chen Soong Der, Sherna Aziz, "Red Blood Cell Recognition using Geometrical Features", *IJCSI International Journal of Computer Science Issues*, Vol. **10**, Issue **1**, January **2013**.

[4] Sarmad Shafique, Samabia Tehsin, "Computer-Aided Diagnosis of Acute Lymphoblastic Leukemia", *Hindawi, Computational and Mathematical Methods in Medicine Volume* **2018**, **13** pages.

[5] Chaitali Raje, Jyoti Rangole, "Detection of Leukemia in Microscopic Images Using Image Processing", *International Conference on Communication and Signal Processing*, April 3-5, **2014**, **India**, IEEE.

[6] Nimesh Patel, Ashutosh Mishra, "Automated Leukemia Detection Using Microscopic Images", *Procedia Computer Science 58 (2015)* **635 – 642**, Published by Elsevier B.V.

[7] R.G Bagasjvara, Ika Candradewi, Sri Hartati, Agus Harjoko, "Automated Detection and Classification Techniques of Acute Leukemia using Image Processing: A Review", **2016** *2nd International Conference on Science and Technology-Computer (ICST)*, Yogyakarta, Indonesia, **978-1-5090-4357-6/2016** IEEE.

[8] S.Srimathi, G.Yamuna, "Study of Cancer Detection Techniques Using Various Image Processing Algorithms", *International Journal of Engineering Development and Research (www.ijedr.org)*, IJEDR **2018**, Vol**. 6**, Issue **3**.

[9] Preeti Jagdev**, "**Automated Detection and Classification of Leukemia Using Image Processing and Machine Learning", *International Journal of Advance Research in Science and Engineering,* Vol. **6**, Issue **4**, **April, 2017**.

[10] Berge, Heidi, Dale Taylor, Sriram Krishnan, and Tania S.Douglas. "Improved red blood cell counting in thin blood smears." *In Biomedical Imaging: From Nano to Macro,* **2011** IEEE International Symposium on, pp. **204-207**.

[11] H. Ramoser, V. Laurain, H. Bischof, Rupert Ecker, "Leukocyte segmentation and classification in blood-smear images," *Proc. IEEE-EMBS,* **2005**, pp. **3371-3374.**

[12] M. Subrajeet, D. Patra, and S. Satpathy, "Automated Leukemia detection in blood microscopic images using statistical texture analysis," *Proc. of the 2011 International Conference on Communication*, *Computing & Security*, pp. **184-187**. ACM, **2011**.