Detection of Leukemia and its Types using Image Processing and Machine Learning

P.Rama Rao[1], Mueed Atif[2], Shaik Abdul Muqeet[3]
Department of CSE
G.Pullaiah College of Engineering and Technology

Abstract- *Leukemia (blood cancer) begins in the bone marrow and causes the formation of a large number of abnormal cells. The most common types of leukemia known are Acute lymphoblastic leukemia (ALL), Acute myeloid leukemia (AML), Chronic lymphocytic leukemia (CLL) and Chronic myeloid leukemia (CML). This thesis makes an effort to devise a methodology for the detection of Leukemia using image processing techniques, thus automating the detection process. The dataset used comprises of 220 blood smear images of leukemic and non leukemic patients. The Image segmentation algorithms that have been used are k means clustering algorithm, Marker controlled watershed algorithm and HSV color based segmentation algorithm. The morphological components of normal and Leukemic lymphocytes differ significantly; hence various features have been extracted from the segmented lymphocyte images. The leukemia is further classified into its types and sub types by making use of the SVM classifier, which is a Machine Learning classifier. This thesis aims at detecting leukemia and determine whether it is AML, CML, CLL or ALL; thus taking the classification process one step further in the field of research, because most of the previous methods have been limited to the detection of leukemia or classifying it into one or two subtypes.*

Keywords- K means clustering algorithm, Marker controlled watershed algorithm, HSV colour based segmentation algorithm,SVM Classifier.

I. INTRODUCTION

Many Image processing algorithms have been developed for Leukemia detection. Image segmentation is a fundamental problem in automated hematological analysis and needs to be accurately carried out. Development of computer algorithms for Classification of leukemia is another aspect of research in hematological image processing.

In automated image segmentation the thresholding is done by Otsus method because in Otsus method the threshold value is automatically selected [14]. The Histogram Equalization and Linear Contrast Stretching method is very useful for detecting white cell, for contrast enhancement [1] and for better segmentation [14]. However, it is hard to define boundary of the overlapping cell [1]. Aimi Salihah proposed the use of three contrast enhancement techniques for colour images using RGB components [12].The three contrast enhancement techniques were partial contrast, bright stretching and dark stretching. The results showed that, the partial contrast was the best technique that helped to improve the image visibility while preserving the significant features of acute leukemia images. The Colour images allow for more reliable image segmentation than the grayscale images. Two of the basic models for colour images are the HSI (Hue, Intensity, and Saturation) colour space and the RGB (Red, Green, and Blue) colour space [14]. The three Components of RGB colour are highly correlated, so the chromatic information is not suitable for direct processing. Due to colour space, it is convenient to convert from RGB to HSV (Hue, Saturation and Value) colour space [11].

The feature extraction technique which is carried out after the segmentation process plays a vital role in differentiating the normal cells from the leukemic ones. The features are extracted from the nucleus of the WBC. The commonly extracted features are radius, roundness, standard Deviation, major axis and minor axis of the nucleus of WBC [1].Subrajeet Mohapatra suggested that features such as fractal dimension, shape features including contour signature and texture could be extracted for detection of leukemia [2].Arjun Nelikanti proposed detection of leukemia by extracting features such as mean, maximum and minimum intensity and the area of the cells involved [8].Preety Singh suggested two types of texture descriptors namely LBP and GLCM, wherein LBP is a texture descriptor which is very effective and robust to illumination changes [15]. Once leukemia is detected; it needs to be classified into its types (Acute or Chronic). Subrajeet Mohapatra suggested the use of FLANN-Functional Link Artificial Neural Network for classification purpose [2].

II. PROCESS OVERVIEW

Figure 1 gives a detailed description of the sequence of steps that have been followed for efficient detection and classification of leukemia.

A. Image Acquisition

Totally 220 images of the blood smears of leukemia patients (Acute Lymphoblastic Leukemia, Acute Myeloid Leukemia Chronic Lymphocytic Leukemia and Chronic Myeloid Leukemia patients) and non leukemic patients; have

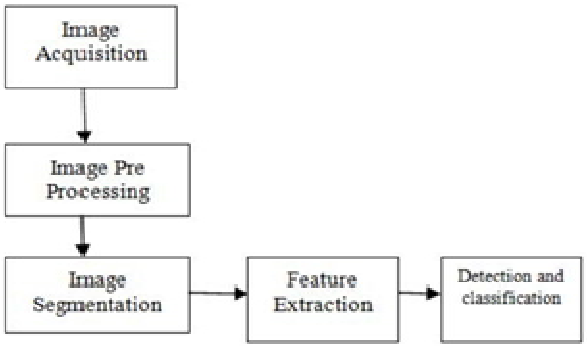

Figure 1. Process flow

been obtained from online databases and from the Goa Medical College.

B. Image Pre-processing

Image pre-processing is a technique implemented to improve the image quality. The images that were obtained were in CMYK form, hence were pre processed and converted to RGB form.

C. Image segmentation algorithm

The main aim of the segmentation process is to simplify and represent the image into something more meaningful and easier to analyze.

1. K means clustering algorithm

In the K means clustering method, the data of the image is grouped in an exclusive way so that if a certain data belongs to a definite cluster, it cannot be included in any other cluster.

Step 1-Input the image from the database, present in RGB form.

Step 2 -The RGB colour space is visually difficult to segment, hence convert the images from RGB to grayscale.

Step 3 -Convert the image from grayscale to L*a*b* Colour Space. Here 'L*' stands for luminosity and contains brightness value of each colour, 'a*' stands for chromaticity layer which indicates where the colour falls along the red-green axis, and 'b*' determines where the colour falls along the blue-yellow axis. All colour information is in the 'a*' and 'b*' layers.

Step 4-Value of k is taken as 3. These clusters correspond to nucleus (high saturation), background (high luminance and low saturation), and other cells (e.g., erythrocytes and leukocyte cytoplasm).

Step 5-Distance between the two clusters is measured using the Euclidean distance metric.

Step 6- Label every pixel in the image with the cluster index.
Step 7-Using the pixel labels, separate objects in the image by colour, which will result in three images.

Step 8-Separate dark blue cluster from light blue cluster using the 'L*' layer in the L*a*b* colour space. The cell nuclei are dark blue.

Step 9- Programmatically determine the index of the cluster containing the blue objects using the cluster centre value. The

blue cluster has the smallest cluster centre value (determined experimentally).

Step 10-Use a mask to label which pixels belong to the blue nuclei. Display the blue nuclei as a separate image.

2. Marker controlled Watershed Algorithm

The major problem faced by the watershed algorithm is over segmentation. This algorithm is sensitive to all the local minima of the image and tends to define the lines of the watershed transform where the local minima give rise to a region. By using the watershed transform; the exact separation of cells cannot be done. Segmentation using this method works well if one can "mark," the fore-ground objects and background locations, so as to detect the boundary lines between the connected cells and successfully separate all the connected nuclei in the image into its individual nuclei.

Marker-controlled watershed segmentation follows this procedure:

Step 1-Read in the Colour Image and Convert it to Greyscale.
Step 2-Use the Gradient Magnitude as the Segmentation Function.
Step 3-Mark the Foreground Objects.
Step 4-Compute the Background Markers.
Step 5-Compute the Watershed Transform of the Segmentation Function and visualize the Result.

3. HSV colour based segmentation algorithm

HSV stands for Hue, Saturation, and Value. This colour model is based on polar coordinates. H component in HSV colour space contains most of the WBC information while the S component contains the structure information of the WBC nucleus. To convert from RGB to HSV scale the colours from [0,255] to [0, 1].HSV separates the image intensity from the colour information. This is very useful in many applications, E.g. If one wants to do histogram equalization of a colour image, it can be done only on the intensity component and leave the colour components alone, also the code for converting between RGB and HSV can be easily implemented. The HSV colour space is used to select the skin tones in an image and apply a filter on the selected region, thus eliminating noise, dust and scratches.

D. Feature extraction

While analysing data, the major problem arises due to the number of variables involved that require a large amount of memory and computation. This problem is overcome by feature extraction. Feature extraction starts with an initial set of data and yields values that are informative and non-redundant, resulting in better human interpretations. The Features that have been extracted are

Cell size- The cell size is the total number of pixels of the nucleus and cytoplasm area.

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 analysing what is normal, extra large or extra small.

Correlation- represents correlation between pixel values and its neighbourhood. It indicates a predictive relationship that can be exploited in practice and shows how strongly pairs of variables are related.

Skewness: is a measure of the lack of symmetry. The zero value indicates that the distribution of the intensity values is relatively equal on both sides of the mean.

Contrast- is a measure of the amount of local variations present in an image.

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

Smoothness- is a function of the colour gradients and enables capturing of important patterns in the data, while leaving out noise in the image.

Variance- is defined as the average of the squared differences from the Mean.

Homogeneity: is a measure of degree of variation for any texture. The larger the changes in grey values, the lower the homogeneity. The range of homogeneity is [0, 1].

E. Classification

To classify cells as leukemic or non leukemic, the SVM classifier has been used. The classification is performed by finding the hyper-plane that differentiates the two classes. The SVMs make use of the kernel trick for non linear classification. The entire data is divided into training and testing data sets. The training data updates the weights while the test data is used for validation of the classifier performance.

III. EXPERIMENTAL RESULTS

The dataset under consideration consists of 220 images of blood smear. Figure 2 shows the segmentation output obtained after implementing the k means clustering algorithm, where the value of K has been taken as 3.

The segmentation output obtained from the implementation of Marker controlled watershed algorithm is given in figure 3. The segmentation output obtained from the implementation of HSV colour based segmentation algorithm is given in figure 4. The features that have been extracted after the implementation of image segmentation using K means clustering segmentation algorithm, Marker controlled watershed algorithm and HSV colour based segmentation algorithm has been tabulated and given in table 1, 2 and 3 respectively. The image segmentation

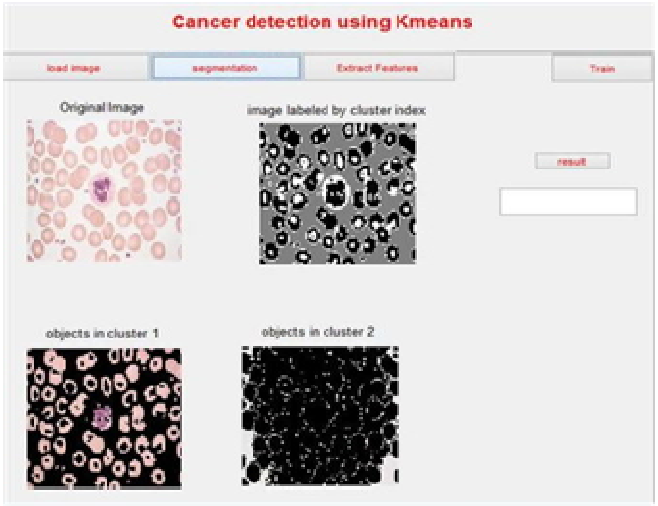

Figure 2. Segmentation of blood smear image using K means clustering

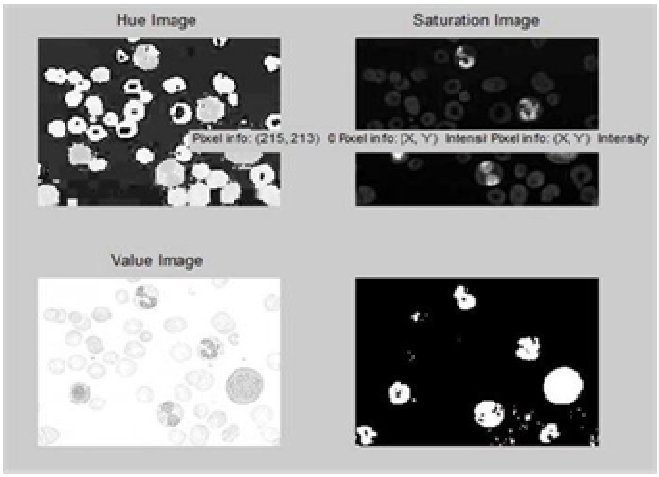

Figure 3. Segmentation of blood smear image using HSV colour based segmentation algorithm

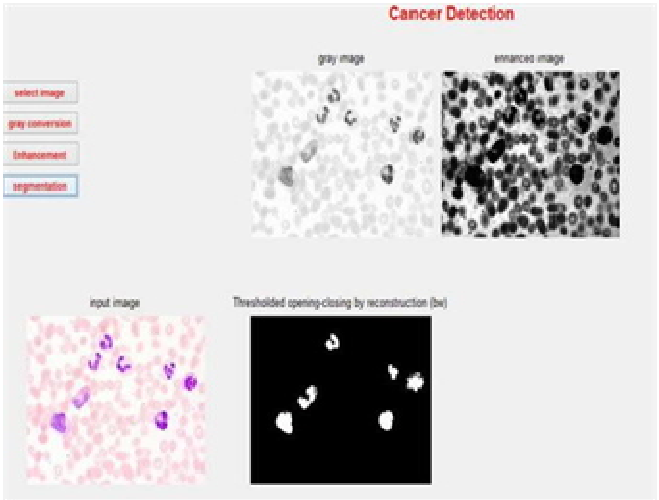

Figure 4. Segmentation of blood smear image using Marker controlled watershed algorithm

Table 1. Feature extraction of images segmented using K means clustering Algorithm

Feature Extracted	Normal Blood smear	AML Blood smear	ALL Blood smear	CML Blood smear	CLL Blood smear
Size	235.489	218.074	198.89	171.438	220.1755
Mean	22.09	42.7664	49.21	56.1297	42.8669
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
Variance	15.9375	15.9375	15.9375	15.9375	15.9375

Table 2. Feature extractions of images segmented using Marker controlled watershed Algorithm

Feature Extracted	Normal Blood smear	AML Blood smear	ALL Blood smear	CML Blood smear	CLL Blood smear
Size	235.489	218.074	198.8898	171.431	220.1755
Mean	22.09	42.7664	49.214	56.1297	42.8669
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
Variance	15.9375	15.9375	15.9375	15.9375	15.9375

Table 3. Feature extraction of images segmented using HSV colour based segmentation Algorithm

Feature Extracted	Normal Blood smear	AML Blood smear	ALL Blood smear	CML Blood smear	CLL Blood smear
Size	235. 5257	218.0 911	198.8931	171.4466	220.1311
Mean	22. 8690	44.0984	50.3336	57.0832	43.5155
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
Variance	15.9375	15.9375	15.9375	15.9375	15.9375

and feature extraction is followed by the detection and classification process. The SVM classifier is used to detect whether the person has leukaemia or not.

If the person is detected with leukaemia, then the type of leukaemia is detected i.e. whether it is ALL, AML, CML or CLL. Figure 5 and 6 show the detection and classification outputs.

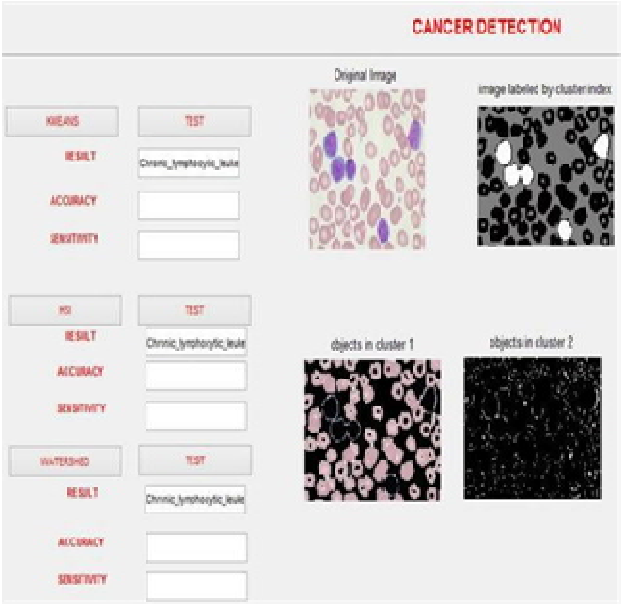

Figure 5. Detection of leukemia and classification using SVM classifier

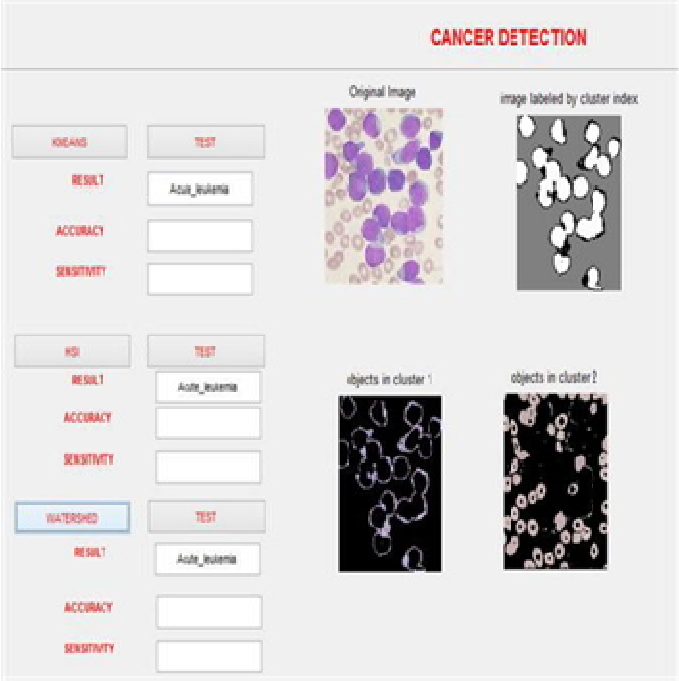

Figure 6. Detection of leukemia and classification using SVM classifier

IV. CONCLUSION AND FUTURE WORK

The paper mainly focuses on the detection of Leukaemia and provides a broader range of Leukaemia classification into its four main types. Three segmentation algorithms were used. A large number of features were extracted to make the detection process more accurate and precise. This work can further be extended by detecting the subtypes of leukaemia types, e.g. AML M3 is a subtype of AML. More segmentation algorithms can be explored, so as to obtain better results as compared to the previous ones.

REFERENCES

[1] Himali P. Vaghela, Hardik Modi, Manoj Pandya and M.B. Potdar Leukaemia Detection using Digital Image Processing Techniques", November 2015 - International Journal of Applied Information Systems (IJAIS)

[2] Subrajeet Mohapatra and Dipti Patra, Automated Leukaemia Detection using Hausdroff Dimension in Blood Microscopic Images ", 2010 IEEE

[3] Emad A. Mohammed, Mostafa M. A. Mohamed, Christopher Naugler and Behrouz. H.Far, Chronic lymphocytic leukaemia cell segmentation from microscopic blood images using watershed algorithm and optimal thresholding", 2013 26th IEEE Canadian Conference of Electrical and Computer

[4] Sos Agaian and Anthony T. Chronopoulos .Automated Screening System for Acute Myeloid Leukaemia Detection in Blood Microscopic Images, IEEE SYSTEMS JOURNAL, VOL. 8, NO. 3, SEPTEMBER 2014

[5] Jyoti Rawat, Kumar, Comparative analysis of segmentation algorithms for leukocyte extraction in the acute lymphoblastic images", 2014 International Conference on Parallel, Distributed and Grid Computing

[6] Adnan Khashman and Hayder Hassan Abbas, Acute Lymphoblastic Leukaemia Identification Using Blood Smear Images and a Neural Classifier", 2014 International Conference on Parallel, Springer-Verlag Berlin Heidelberg 2013

[7] Subrajeet Mohapatra and Dipti Patra, Fuzzy based Blood Image Segmentation for Automated Leukaemia Detection", 2011 IEEE

[8] Arjun Nelikanti Segmentation and Analysis of Cancer Cells in Blood Samples, Indian Journal of Computer Science and Engineering (IJCSE).

[9] R. Hassan, Classification of Blasts in Acute Leukaemia Blood Samples Using K-Nearest Neighbour", 2012 IEEE 8th International Colloquium on Signal Processing and its Applications

[10] A.S.Abdul Nasir, M.Y.Mashor, Unsupervised Colour Segmentation of White Blood Cell for Acute Leukaemia Images", 2011 IEEE

[11] A.S.Abdul Nasir, M.Y.Mashor, Nucleus Segmentation Technique for Acute Leukaemia", 2011 IEEE 7th International Colloquium on Signal Processing and its Applications.

[12] A.S.Abdul Nasir, M.Y.Mashor, Colour Image Enhancement Techniques for Acute Leukaemia Blood Cell Morphological Features", 2010 IEEE

[13] Hayan T. Madhloom, Sameem Abdul Kareem, A Robust Feature Extraction and Selection Method for the Recognition of Lymphocytes versus Acute Lymphoblastic Leukemia"2012International Conference on Advanced Computer Science Applications and Technologies.

[14] Tejashri G. Patil, V. B. Raskar, Blood Microscopic Image Segmentation Acute Leukaemia Detection International Journal of Emerging Research in Management Technology (Volume-4, Issue-9)

[15] Vanika Singhal and Preety Singh, Texture Features for the Detection of Acute Lymphoblastic Leukaemia Proceedings of International Conference on ICT for Sustainable Development Springer