

Automated Screening System for Acute Myelogenous Leukemia Detection in Blood Microscopic Image using K-NN Classifier

Shameem P C

Department of Electronics and Communication
KMCT college of Engineering
University of Calicut

Shyno K G

Department of Applied Electronics and Instrumentation
KMCT college of Engineering
University of Calicut

Abstract— Acute Myelogenous Leukemia (AML) is a subtype of acute leukemia, which is prevalent among adults. The average age of a person with AML is 65 years. The need for automation of leukemia detection arises since current methods involve manual examination of the blood smear as the first step toward diagnosis. This is time-consuming, and its accuracy depends on the operator's ability. In this paper, a simple technique that automatically detects and segments AML in blood smears is presented. The proposed method differs from others in: the simplicity of the developed approach; classification of complete blood smear images as opposed to sub images; and use of these algorithms to segment and detect nucleated cells. Computer simulation involved the following tests: comparing the impact of Hausdorff dimension on the system before and after the influence of local binary pattern, comparing the performance of the proposed algorithms on sub images and whole images, and comparing the results of some of the existing systems with the proposed system.

Keywords—Acute Myelogenous Leukemia, *k*-means with color based thresholding, feature extraction, classification

I. INTRODUCTION

The most important part of any human body is blood as it keeps one alive. It performs many important functions such as to transfer oxygen, carbon dioxide, mineral and etc. to the whole body in order to maintain metabolism. Blood consists of three main components which RBC, WBC and Platelets. Insufficient amount of the blood could affect the metabolism greatly which could be very dangerous if early treatment is not taken. One of the common blood disorders is Leukemia. Leukemia is the common type of cancer in children. All cancers begin in body cells, and leukemia is a cancer that begins in blood cells. Generally, cells grow and multiply to form new cells as the body needs them. When cells grow old, they die and new cells take their place. Sometimes, this cycle does not work correctly. In cancer, new cells are formed when the body does not need them, and old cells do not die when they should.

Leukemia is a cancer that involves the blood-forming tissues of the bone marrow, spleen and lymph nodes. It is characterized by an uncontrolled production of immature blood cells.

Fig.1[1] shows three different images of Myeloblasts from AML patients and non AML patients obtained from American Society of Haematology.

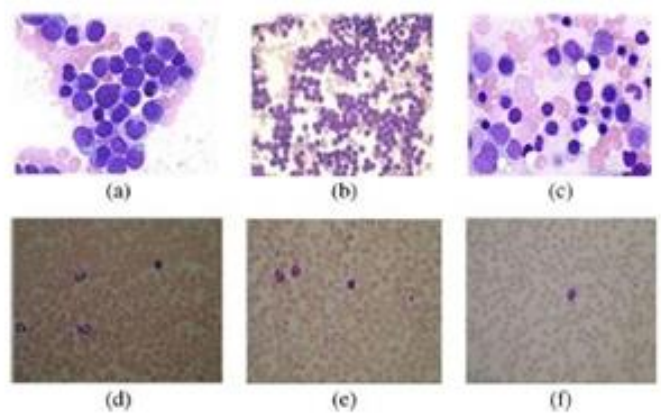


Fig.1. Images from ASH. (a)-(c) Myeloblasts from AML patients. (d)-(f) Healthy cells from non-AML patients.

Leukemia is broadly classified as: 1) *acute leukemia* (which progresses quickly); and 2) *chronic leukemia* (which progresses slowly). Acute myelogenous leukemia (AML) is a heterogeneous clonal disorder of haemopoietic progenitor cells (—blasts), which lose the ability to differentiate normally and to respond to normal regulators of proliferation. This loss leads to fatal infection, bleeding, or organ infiltration, typically, in the absence of treatment, within a year of diagnosis. AML is confirmed when the marrow contains more than 30% blasts. In this paper, only AML is considered.

AML is often difficult to diagnose since the precise cause of AML is still unknown. In addition, the symptoms of the disease are very similar to flu or other common diseases, such as fever, weakness, tiredness, or aches in bones or joints. If the described symptoms are present, blood tests, such as a full blood count, renal function and electrolytes, and liver enzyme and blood count, have to be done. Since there is no staging for AML, choosing the type of treatment can vary from

chemotherapy, radiation therapy, bone marrow transplant, and biological therapy.

II. LITERATURE SURVEY

The diagnosis of leukemia frequently follows a routine blood test that results in an abnormal blood cell count. Once leukemia is suspected, the doctor may take samples of bone marrow and blood to examine cell shape. Samples are also sent to the pathology lab to identify proteins located on the surface and chromosomal and changes. This information is important for diagnosis of individual patients.

A. Previous Methods

For diagnosing leukemia several methods[1], [10] are used. The main five methods are discussed below.

1. *Medical history and physical examination:* The record of present symptoms, and problems a person has had in the past. The medical history of a person's family also helps in diagnose leukemia.

2. *Complete blood count (CBC):* Blood is taken and checked under the microscope for the number of RBCs, WBCs and platelets.

3. *Bone marrow aspiration:* Bone marrow is removed with the help of a needle from breastbone. The removed sample is observed under a microscope to look for abnormal cells.

4. *Cytogenetic analysis:* Cytogenetic test takes blood or bone marrow to help identify individual chromosomes. It shows abnormalities in chromosomes, which help to diagnosis and identify the type of leukemia. Results are usually available within 3 weeks.

5. *Immunohistochemistry:* Blood sample of cells are treated with special antibodies in immunohistochemistry. Under the microscope the change in color can be seen. It helps in determining the types of cells that are present.

B. Existing Methods

The existing methods for leukemia detection is described below

1. *Peripheral Smear:* A blood smear is a diagnostic test used to look for abnormalities within the blood. Giemsa Stain - purple and pink. The cell types are examined under a microscope for unusual shapes or sizes. There are three main cells within the blood that the test focuses on: *Red cells* (which carry oxygen throughout the body). *White cells* (which function as part of the body's immune system) *Platelets* (which are important for blood clotting)

2. *Bone Marrow Aspiration Cytology:* Another method for leukemia detection is Bone Marrow Aspiration Cytology. Since the production of RBC is from Reticulocyte in bone marrow. Here bone marrow aspiration needle is used to take bone marrow and this bone marrow is mixed with stains and examined under microscope. It is a painful method.

C. Limitations

Following are the major limitations of above mentioned methods

Manual method, possibility of human error, not precise difficult to find out initial stage of leukemia, complicated method, well experienced technician is required.

III. PROPOSED METHOD

The system proposed ensures step-by-step processing. The system overview gives a detailed depiction of the sequence of steps that are to be followed for efficient classification of acute leukemia. The first step involves preprocessing the complete images to overcome any background non uniformity due to irregular illumination.. This step is followed by color based thresholding segmentation bring out the nucleus of each cell. Segmentation is followed by feature extraction based on which classification and validation are performed.



Fig.2. Overview of proposed method.

A. Image Acquisition

For AML, we accessed the *American Society of Hematology (ASH)* for their online image bank of leukemia cells. The ASH image bank is a web-based image library that offers comprehensive and growing collections of images relating to a wide range of hematology categories. They provide high-quality images captured using different microscopes in different resolutions. Our database for AML comprised 80 images—40 from AML patients and 40 from non-AML patients. The resolution used for our classification was 184×138 pixels.

IV. NUCLEI SEGMENTATION

The goal of image segmentation[1],[2] is to extract important information from an input image. It plays a key role since the efficiency of subsequent feature extraction and classification relies greatly on the correct identification of the myeloblasts. Many algorithms for segmentation have been developed for gray-level images. Segmentation in this system is performed for extracting the nuclei of the leukocytes using color-based clustering. Cluster analysis is the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity. Cluster analysis does not use category labels that tag objects with prior identifiers, i.e., class labels. *k*-means, which is one of the most popular unsupervised learning algorithm and is also a simple clustering algorithm, here *k*-means clustering algorithm with color based thresholding is used.

The efficiency of the pure color-based *K*-means algorithm can be improved by combining the algorithm by color-based thresholding. The proposed algorithm[3] is carried out in following steps:

1. The very first step is extracting our color bands from the original image into separate 2D arrays, one for each component (Red, Blue and Green).
2. The next step is to compute red, green and blue histogram. Then all axes are set to be of same height and width (averaging of histogram), this makes them easy to compare them.
3. Decide the low and high thresholds for each color band. Choose a value which may suit the image (trial and error method is used to compute the best possible output).
4. Apply these thresholds on their respective color band. Then adding the masks to find where all 3 are —true||.
5. Then we will have the mask of only those parts of image whose threshold has been applied.
6. It may happen that any one band mask may become 0, thus if it becomes all 0" s then set them to 1. Otherwise the entire image will be added with zero and output will be black image.
7. Now use this object mask to mask out the input image. Again concatenate the masked color bands to RGB image. Now here we get the extracted object.
8. But since we don't have exact red objects in image, the red mask can't be applied directly so to obtain land portion of image, the original image is subtracted from the blue and green subparts obtained as result of thresholding operation. The image thus obtained is pure land image. Now these results are given to clustering algorithm.
9. There in water bodies accuracy is perfect so that cluster is untouched, so display it as it as.
10. The other two parts of image viz. forest and land bodies are taken from clustering and displayed as output.

V. FEATURE EXTRACTION

Feature extraction in image processing is a technique of redefining a large set of redundant data into a set of features of reduced dimension. Transforming the input data into the set of features is called feature extraction. Feature selection greatly influences the classifier performance; therefore, a correct choice of features is a very crucial step. In order to construct an effective feature set, several published articles were studied, and their feature selection methodology was observed. It was noted that certain features were widely used as they gave a good classification. We implemented these features on whole images in our system. Those features were considered to boost the classifier performance. Fig. 3 [1] gives the set of features chosen to classify the image database[4],[8]. *HD*: Fractals have been used in medicine and science in the past for various quantitative measurements.

The fractal dimension is a statistical quantity that gives an indication of how completely a fractal appears to fill space. There are many specific definitions of fractal dimension. The most important theoretical fractal dimensions are the dimension, the HD, and the packing dimension. Practically, the box-counting dimension is widely used, partly due to their ease of implementation. In a box counting algorithm, the

number of boxes covering the point set is a power-law function of the box size. Fractal dimension is estimated as the exponent of such power law. All fractal dimensions are real numbers that characterize the fractalness (texture/roughness) of the objects. Myeloblast can be differentiated using perimeter roughness of the nucleus.

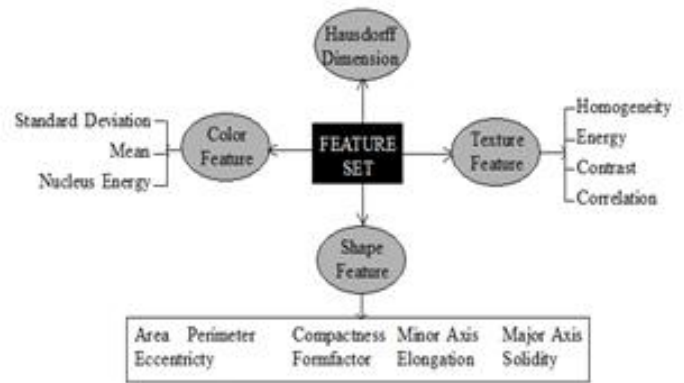


Fig. 3. Feature set developed for the proposed system comprising of shape, color, texture features, and HD.

HD is considered an essential feature considered in our proposed system. The procedure for HD measurement using the box counting method is elaborated below as an algorithm:

- 1) binary image in obtained from the gray-level image of the blood sample;
- 2) edge detection technique is employed to trace out the nucleus boundaries;
- 3) edges are superimposed by a grid of squares;
- 4) the HD may then be defined as follows:

$$HD = \frac{\log(R)}{\log(R(s))} \tag{1}$$

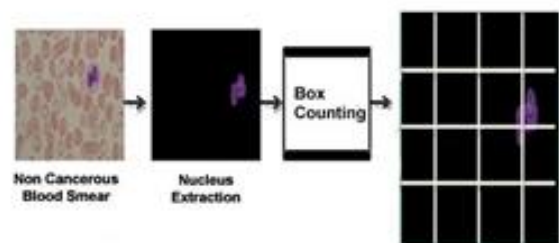


Fig. 4. Superimposing of the nucleus with a grid of squares for box count measure

where *R* is the number of squares in the superimposed grid, and *R(s)* is the number of occupied squares or boxes (box count). Higher HD signifies higher degree of roughness.

Fig. 4 [1] illustrates the given algorithm. It shows how the nucleus from a noncancer cell is superimposed with a grid of squares to perform suitable box counting. The finer the grid

gets, the more accurate is the shape approximated. Fig 5 depicts the results of HD on subimages and complete images. It can be clearly observed that, for subimages, there is only a marginal difference in the HD value, whereas there is a distinct difference for complete images. Thus, HD turned out to be a crucial feature in our system particularly since we considered whole images of the blood sample. In the whole images, the number of nuclei under the field of view was much higher for a cancerous case as opposed to the noncancerous case. This resulted in steep difference in box count between the two cases and thereby proved to be an effective feature.

LBP: The concept of Local Binary Pattern[5] (LBP) was introduced for texture classification. This approach has many advantages. For example, the LBP texture features have the following characteristics: 1) They are robust against illumination changes; 2) they are very fast to compute; 3) they do not require many parameters to be set; 4) they are local features; 5) they are invariant with respect to monotonic grayscale transformations and scaling; and 6) they have performed very well in many computer vision image retrieval applications. The LBP method has proved to outperform many existing methods, including the linear discriminant analysis and the principal component analysis.

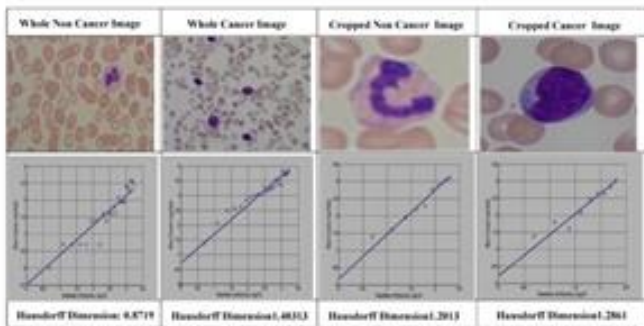


Fig. 5. Result of Hausdorff dimensions

In order to deal with textures at different scales, the LBP operator was later extended to use neighborhoods of different sizes. Defining the local neighborhood as a set of sampling points evenly spaced on a circle centered at the pixel to be labeled allows any radius and number of sampling points. When a sampling point does not fall in the center of a pixel, bilinear interpolation was employed. In the LBP method where each pixel is replaced by a binary pattern that is derived from the pixel's neighborhood. Each grayscale pixel P of an image is used as a center of a circle with radius $R = 1$ or 2 (radius R is usually kept very small). M represents the number of samples that determines the number of points that are taken uniformly from the contour of the circle. If needed, these points are interpolated from adjacent pixels. Each grayscale pixel P is compared with these sample points one by one. If the center point P is larger than the current neighborhood sample point I , the result is a binary zero; otherwise, the result is a binary one. When doing this operation, for example, clockwise from a certain starting point, the result will be a binary pattern with length M . This operation is illustrated in Fig.6. For our database of images, an $(8, 1)$ circular neighborhood was used. The segmented images were extracted using k -means

clustering, and then the LBP operator was applied on them before calculating the HD (see Fig. 6)[1].



LBP of pixel (1,1) is 00111101:61

Fig. 6. LBP operator example.

Two sets of values were extracted: first, HD of the 80 images without applying LBP, and second, HD of the images after applying LBP. When comparing these two data sets, it was observed that the LBP operator enhanced the overall performance by a very high margin. Additionally, the following features have been also chosen in our classification system: shape gray-level co occurrence matrix (GLCM) and color features. The choice features were justified by extensive computer simulations in order to identify the ones that yielded maximum discrimination capability, thus achieving the optimal diagnostic performance. In this paper, we use several features, such as, shape features, GLCM features, and color features.

The diagnostic performance of some new individual features selected in this paper will be analyzed in the following. **Shape features.** One of the shape features that has proven to be a good measure for classifying AML by their shape is compactness.



The shape of the nucleus, according to haematologists, is an essential feature for discrimination of myeloblasts. Region- and boundary-based shape features are extracted for shape analysis of the nucleus. All the features are extracted from the binary-equivalent image of the nucleus where the nucleus region is represented by the nonzero pixels. Table I displays the difference in the values of the shape features for a pair of cancer and non cancer nuclei. GLCM features[6].

Texture is defined as a function of the spatial variation in pixel intensities. The GLCM and associated texture feature calculations are image analysis techniques. Gray-level pixel distribution can be described by second-order statistics such as the probability of two pixels having particular gray levels at particular spatial relationships. This information can be depicted in 2-D gray-level co occurrence matrices, which can be computed for various distances and orientations. In order to use information contained in the GLCM, Haralick defined some statistical measures to extract textual characteristics. Some of these features are the following.

- 1) **Energy:** Also known as uniformity (or angular second moment), it is a measure of homogeneity of image.
- 2) **Contrast:** The contrast feature is a difference moment of the regional cooccurrence matrix and is a measure of the contrast or the amount of local variations present in an image.

- 3) *Entropy*: This parameter measures the disorder of an image. When the image is not texturally uniform, entropy is very large.
- 4) *Correlation*: The correlation feature is a measure of regional-pattern linear dependence in the image.

Table II
Shape Features With Their Illustrations

Features	Cancerous	Normal
Images		
Mean	32.3699	37.1222
Standard Deviation	47.3662	41.0178
Area	6453	1985
Perimeter	256	807.293
Elongation	1.1357	1.522
Eccentricity	0.4740	0.7541
Formfactor	1.2373	0.0383
Solidity	0.5679	0.2117
Copmactness	10.1559	328.32

VI. COMPUTER SIMULATION

The selection of a classification technique for classification is a challenging problem because an appropriate choice given the available data can significantly help improving the accuracy in credit scoring practice. There is a plenty of statistical techniques, which aim at solving binary classification tasks. In this paper k-NN classifier is used for classification process.

A. K-Nearest-Neighbor Classification

k-nearest neighbor algorithm[7] is a method for classifying objects based on closest training examples in the feature space. k-nearest neighbor algorithm is among the simplest of all machine learning algorithms. Training process for this algorithm only consists of storing feature vectors and labels of the training images. In the classification process, the unlabelled query point is simply assigned to the label of its k nearest neighbors.

Typically the object is classified based on the labels of its k nearest neighbors by majority vote. If k=1, the object is simply classified as the class of the object nearest to it. When there are only two classes, k must be a odd integer. However, there can still be ties when k is an odd integer when performing multiclass classification. After we convert each image to a vector of fixed-length with real numbers, we used the most

common distance function for KNN which is Euclidean distance:

$$d(x, y) = \|x - y\| = \sqrt{(x - y) \cdot (x - y)} = (\sum_{i=1}^m ((x_i - y_i)^2))^{1/2} \tag{2}$$

where x and y are histograms in $X = R^m$. Fig. 7 shows visualizes the process of KNN classification[8].

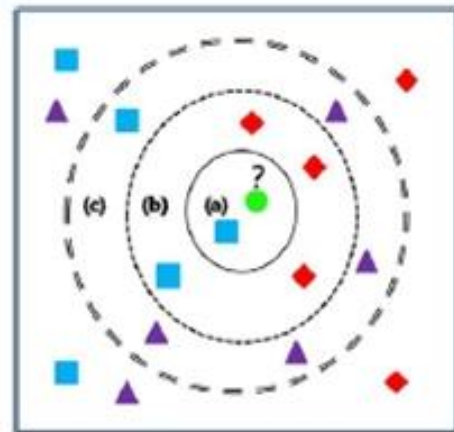


Fig. 7 KNN Classification.

At the query point of the circle depending on the k value of 1, 5, or 10, the query point can be a rectangle at (a), a diamond at (b), and a triangle at (c).

A main advantage of the KNN algorithm is that it performs well with multi-modal2 classes because the basis of its decision is based on a small neighborhood of similar objects. Therefore, even if the target class is multi-modal, the algorithm can still lead to good accuracy. However a major disadvantage of the KNN algorithm is that it uses all the features equally in computing for similarities. This can lead to classification errors, especially when there is only a small subset of features that are useful for classification.

VII. EXPERIMENTAL RESULT

The proposed technique has been applied on peripheral blood smear images obtained from two places, as aforementioned. To evaluate the proposed method, the following four measures of accuracy were used in this paper.

A microscopic blood image of size 184 × 138 is considered for evaluation. The superiority of the scheme is demonstrated with the help of an experiment. Feature extraction with and without the LBP operator presented very interesting results. The system constructed without having to employ the LBP operator gave an efficiency of 93.5%. All the three validation methods were incorporated into our system. However the performance of HD, in particular, after using LBP increased the classifier performance by 4%. By employing LBP, the edges of the nuclei of the myleoblasts were extracted in a very pronounced manner[9].

This effective edge detection enhanced the HD, as the box count for AML was much more than the box count for non-AML images. To see the impact of HD in the feature set, the classifier was run with HD as the only feature. This was done twice, once with applying LBP operator and once without LBP operator. All the parameters for evaluation were extracted for both sets. The results are illustrated in Fig.8.

It was observed that, when LBP was not employed, the HD performance was only around 70%, whereas when LBP was employed, the percentage escalated to 93%. This clearly shows the influence of the LBP operator on the system. In order to see the effectiveness of the developed algorithms, a trial was run comparing the system's performance on subimages and whole images. The obtained results further corroborated the impact of the LBP operator.

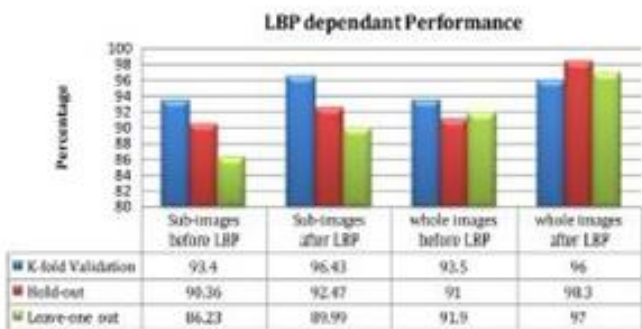


Fig.8 Overall classifier performance with and without LBP code

The efficiency of the pure color-based K-means algorithm can be improved by combining the algorithm by color-based thresholding[5].

The input image is preprocessed using high pass filter for noise removal, followed by Color sharpening which uses a predefined mask that enhances the color intensity of each pixel so that the color separations are distinct and clear. Fig.9[5] shows input image and image after preprocessing.

The K means Clustering algorithm is applied and Classification obtained is shown in Fig.10 [5] Output obtained with proposed algorithm is shown in Fig.11



Fig.9 (a) original input image (b) Output obtained from preprocessing of input image

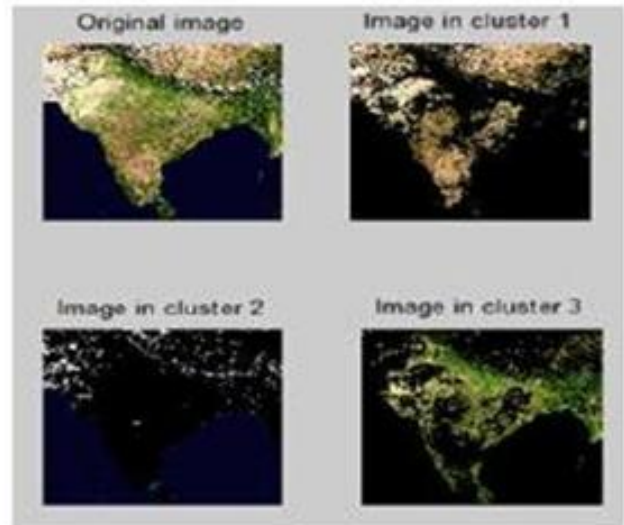


Fig.10 Output image obtained from pure K-means clustering algorithm.

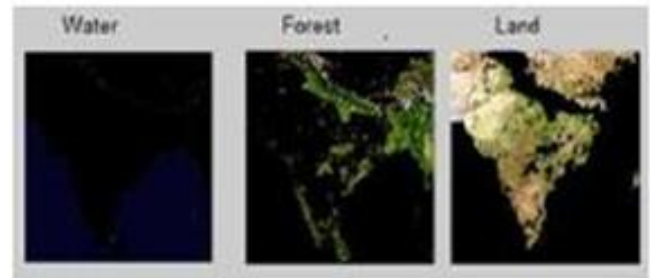


Fig.9 Output obtained from k means clustering with color based thresholding algorithm.

The algorithms are evaluated by using image quality metrics like overall accuracy, user's accuracy, producer's accuracy, average accuracy of user and producer. The proposed algorithm is compared with color based K-Means clustering algorithm using the parameters as shown in Table II[5]

Table II

Comparison of Parameters

Parameters (Accuracy)		K-Means Clustering	Proposed Algorithm
Overall Accuracy		91.85	97.10
User's Accuracy	Water	100	100
	Forest	84.74	95.78
	Land	87.20	93.43
Producer's Accuracy	Water	100	99.9
	Forest	85.13	97.22
	Land	86.85	90.41
Average	User	90.64	96.40
	Producer	90.66	95.84

VIII. CONCLUSION

This paper has reported the design, development, and evaluation of an automated screening system for AML in blood microscopic images. It uses 80 high-quality 184×138 size images obtained from the American Society of Haematology. The presented system performs automated processing, including color correlation, segmentation of the nucleated cells, and effective validation and classification. A feature set exploiting the shape, color, and texture parameters of a cell is constructed to obtain all the information required to perform efficient classification. The impact of the LBP operator on the HD proved to be a promising feature for this analysis. Furthermore, a color feature called cell energy was introduced, and results show that this feature presents a good demarcation between cancer and noncancer cells.

Further research will focus on collection of more samples to yield better performance and building an overall system for cancer classification.

ACKNOWLEDGMENT

I would like to thank Sos Agaian, *Senior Member, IEEE*, Monica Madhukar, and Anthony T. Chronopoulos making their source codes open to public. Also I extend my sincere gratitude to my project guide Asst.Prof. .Shyno.K.G and Head of the Department Asst.Prof. Nishida.T for their guidance and support and also grateful to all the staff members of the Department of Electronics & Communication Engineering of KMCT College of Engineering and Technology, Calicut. I am grateful to Dr.Abhilash MBBS for his guidance. I also thank my family, friends for their support and encouragement.

REFERENCE

- [1] —Automated Screening System for Acute Myelogenous Leukemia Detection in Blood Microscopic Images|| Sos Agaian, *Senior Member, IEEE*, Monica Madhukar, and Anthony T. Chronopoulos, *Senior Member, IEEE*
- [2] F. Scotti, —Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images,|| in *Proc. CIMSAs*, 2005, pp. 96-101.
- [3] V. Piuri and F. Scotti, —Morphological classification of blood leucocytes by microscope images,|| in *Proc. CIMSAs*, 2004, pp. 103-108.
- [4] M. Subrajeet, D. Patra, and S. Satpathy, —Automated leukemia detection in blood microscopic images using statistical texture analysis,|| in *Proc. Int. Conf. Commun., Comput. Security*, 2011, pp. 184-187.
- [5] K-Means Clustering Algorithm with Color-based Thresholding for Satellite Images- Gurudatta V Nayak, Anuja A Rao, NandanaPrabhu.
- [6] H. Ramoser, V. Laurain, H. Bischof, and R. Ecker, —Leukocyte segmentation and classification in blood-smear images,|| in *Proc. IEEE EMBS*, 2006, pp. 3371-3374.
- [7] C. Reta, L. Altamirano, J. A. Gonzalez, R. Diaz, and J. S. Guichard, —Segmentation of bone marrow cell images for morphological classification of acute leukemia,|| in *Proc. 23rd FLAIRS*, 2010, pp. 86-91.
- [8] G. Ongun, U. Halici, K. Leblebicioglu, V. Atalay, M. Beksac, and S. Beksac, —Feature extraction and classification of blood cells for an automated differential blood count system,|| in *Proc. IJCNN*, 2001, vol. 4, pp. 2461-2466.
- [9] S. Mohapatra and D. Patra, —Automated leukemia detection using hausdorff dimension in blood microscopic images,|| in *Proc. Int. Conf. Emerg. Trends Robot Commun. Technol.*, 2010, pp. 64-68.
- [10] S. Mohapatra, S. Samanta, D. Patra, and S. Satpathi, —Fuzzy based blood image segmentation for automated leukemia detection,|| in *Proc. ICDeCom*, 2011, pp. 1-5.
- [11] Detection of Leukemia in Microscopic image using Image Processing - Chaitali Rajee, Jyoti Rangole