New Decision Support Tool for Acute Lymphoblastic Leukemia Classification

Monica Madhukar(mona.madhu@yahoo.co.in), Sos Agaian(sos.agaian@utsa.edu),
Anthony T. Chronopoulos(atc@cs.utsa.edu)
Deptartment of Electrical and Computer Engineering, University of Texas At San Antonio,
One UTSA Circle, San Antonio, Texas 78249

ABSTRACT

In this paper, we build up a new decision support tool to improve treatment intensity choice in childhood ALL. The developed system includes different methods to accurately measure furthermore cell properties in microscope blood film images. The blood images are exposed to series of pre-processing steps which include color correlation, and contrast enhancement. By performing K-means clustering on the resultant images, the nuclei of the cells under consideration are obtained. Shape features and texture features are then extracted for classification. The system is further tested on the classification of spectra measured from the cell nuclei in blood samples in order to distinguish normal cells from those affected by Acute Lymphoblastic Leukemia. The results show that the proposed system robustly segments and classifies acute lymphoblastic leukemia based on complete microscopic blood images.

Keywords: Classification, Acute Lymphoblastic Leukemia, Segmentation, Feature Extraction

1. INTRODUCTION

Blood's major functions are to transport various agents such as oxygen, carbon dioxide, nutrients, wastes, and hormones. Blood cells are composed of erythrocytes (red blood cells, RBCs), leukocytes (white blood cells, WBCs) and thrombocytes (platelets) [37]. White blood cells (WBC) or leukocytes play a significant role in the diagnosis of different diseases, and therefore, extracting information about that is valuable for hematologists (including Leukemia). The pathology is characterized by the uncontrolled accumulation of immature white blood cells.

Leukemia refers to a progressive, malignant disease of the blood-forming organs. It starts in the bone marrow, a soft tissue inside most of the bones, where blood cells are made. There are four main types of leukemia; *Acute Lymphocytic Leukemia (ALL), Acute Myelogenous Leukemia (AML), Chronic Lymphocytic Leukemia (CLL), Chronic Myelogenous Leukemia (CML).* Acute Lymphocytic Leukemia, also known as Acute Lymphoblastic Leukemia is the cancer of the white blood cells characterized by the overproduction and continuous multiplication of the malignant and immature white blood cells (*referred to as lymphoblasts).* Acute leukemia is a progressive disease which appears suddenly and needs to be treated urgently. It is fatal if left untreated due to its rapid spread into the blood stream and other vital organs. Figure 1 shows that ALL is the most common leukemia seen in children, accounting for almost 33% of all the pediatric cases, usually for infants under three years of age. This disease also affects adults, especially those aged 65 and older, but its clinical behavior and treatments are different. Nowadays, the probability of survival for ALL is approximately 75% to 80% in 5 years [11].

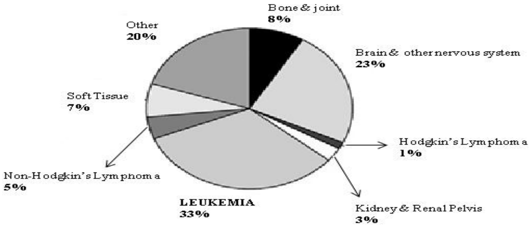


Figure 1. Childhood cancer cases diagnosed in the United States, 2001, Ages 0–19.
Source: U.S. Cancer Statistics Working Group (2004).

Image Processing: Algorithms and Systems X; and Parallel Processing for Imaging Applications II, edited by
Karen O. Egiazarian, Sos S. Agaian, Atanas P. Gotchev, John Recker, Guijin Wang, Proc. of SPIE-IS&T Electronic
Imaging, SPIE Vol. 8295, 829518 • © 2012 SPIE-IS&T • CCC code: 0277-786X/12/$18 • doi: 10.1117/12.905969

Early detection of the disease is required for effective treatment. The primary setback faced by people dealing with ALL cure is that, the *etiology of ALL* still *remains unknown*. Thus, despite of advanced techniques such as flow cytometer, immunophenotyping, molecular probing etc, *microscopic examination of blood slides still remains as the standard leukemia diagnosis technique*. This analysis suffers from time delays and it presents not a standardized accuracy since it depends on operator's capabilities and tiredness. Diagnostic confusion is also posed due to imitation of similar signs by other disorders [8]. Also, blood images obtained by microscopes can be more easily transmitted to clinical centers than liquid blood samples. So there is always a need for a cost effective and robust automated system for leukemia screening which can greatly improve the output without being influenced by operator fatigue [3].

Several attempts of partial/full automated systems for leukemia detection based on image-processing systems are present in literature but they are still at prototype stage [12-14]. Currently five main features are used by automated systems for early ALL detection: cell size, color, shape, density and granularity. For instance, cell segmentation using active contour models is presented in [16]; Color segmentation procedure applied to leukocyte images are described in [17]; a two step segmentation process using HSV color model is used in [15]; The use of shape analysis into WBC segmentation, by considering features such as cell size, granularity, density and shape was introduced in [6]; A cell detection method that utilizes both intensity and shape information of cell to improve the segmentation was proposed by Wang et al. [18]. However, the identification task is usually difficult due to the variety of features and the often unclear images which cause missing out on vital indicators as to which form of leukemia is being observed. Due to complex nature of the blood smear images and variation in slide preparation techniques much work has to be done to meet real clinical demands [10]. Also, most of the existing systems extract features of the sub-images instead of the complete blood smear. The main objective of this paper is to demonstrate that the classification of peripheral blood smear images containing multiple nuclei (and not for cropped images containing one nucleus per image, like most systems do) can be fully automated and can be performed as ancillary/backup service to the physician.

The paper is structured as follows; Section 2 focuses in detail on the proposed system overview and compares it with the literature. Section 3 and Section 4 deal with color correlation and segmentation of the images. Section 5 delves into features extraction of the segmented images, which are subsequently classified in Section 6.

2. ACUTE LYMPHOBLASTIC LEUKEMIA CLASSIFICATION SYSTEM

Many of the image-processing systems for leukemia detection in the literature are still at prototype stage [12-14]. Some systems have been proposed as techniques to refine the segmentation (i.e., to solve a particular cluster of cells such as in [19] or to refine membrane segmentation as in [6]) or to detect incorrect segmentations of white cells as in [20]. Also most of systems developed work on sub-images where only one nucleus per image is presented under the field of view. Table 1 shows methods followed by various systems and the common set-backs faced by all of them. Based on inference made from Table1, it is observed that the most common drawback faced by multiple systems, is that, the features are being extracted only for sub-images instead of whole images. Our goal is to overcome this and also increase the accuracy of the classifier system.

The proposed approach aims to present a more robust system with an efficient segmentation of blood images for high performance. To achieve this goal, the system we propose follows four main processing steps (See Figure 2):

1) *To preprocess the image in order to reduce background non-uniformities and perform color correlation;*
2) *To employ segmentation on whole images by combining different methods in order to exploit all the available a-priori information and thereby achieving a robust identification of the nuclei of the white cells ;*
3) *To extract different sets of features for a database of images*
4) *To run the classifier system and validate the output based on the results obtained.*

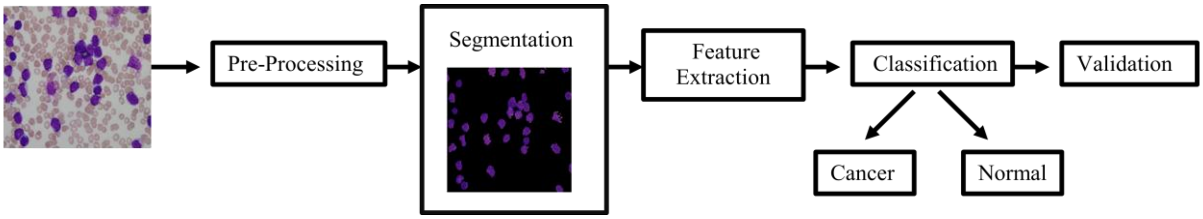


Figure2. System Overview

3. PRE-PROCESSING: COLOR-CORRELATION:

The camera-acquired images of blood sample suffer from irregular background illuminations. Noise may be accumulated during image acquisition and due to excessive staining. The background non-uniformities are smoother than the cells present in the image. Hence it is assumed that background has lower spatial frequencies than the cells. Thus all the test images undergo selective filtering.

Typically images generated by digital microscopes are usually in RGB color space which is difficult to segment. In practice the blood cells and image background varies strongly with respect to color and intensity. This can be caused by multiple reasons such as by camera settings, varying illumination and aging stain. In order to make the cell segmentation robust with respect to these variations an adaptive procedure is used: the RGB input image is converted into the CIELAB color space.

The L*a*b* color space is a color representation technique which is basically used to reduce the color dimension from three to two in comparison to RGB. The L*a*b* space consists of a luminosity layer L*, chromaticity layer a* and chromaticity layer b*. Here the color information is represented in two components i.e. a* and b*. Due to less color dimension L*a*b* color space is mostly employed in color based clustering. In a uniform color scale, the differences between points plotted in the color space correspond to visual differences between the colors plotted. The CIELAB is organized in cube form [21].

Also L*a*b color is designed to approximate human vision and it aspires perceptual uniformity, and its L component closely matches human perception of lightness. In the present work microscopic images are converted from RGB color space to L*a*b* before clustering. Figure3 shows the results of L*a*b space conversion.
Using the X, Y and Z, the CIE Tristimulus Values and X_n, Y_n and Z_n, the tristimulus Values of illuminant, the L*, a* and b* values can be calculated:

If X/X_n , Y/Y_n, and Z/Z_n are all greater than 0.008856, then	If any of X/X_n , Y/Y_n, and Z/Z_n is greater than 0.008856, then
$L *= 116\sqrt[3]{Y/Y_n} - 16$	$L *= 903.3(\frac{Y}{Y_n})$
$a *= 500(\sqrt[3]{\frac{X}{X_n}} - \sqrt[3]{\frac{Y}{Y_n}})$	$a *= 500[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)]$
$b *= 200(\sqrt[3]{\frac{Y}{Y_n}} - \sqrt[3]{\frac{Z}{Z_n}})$	$b *= 200[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)]$

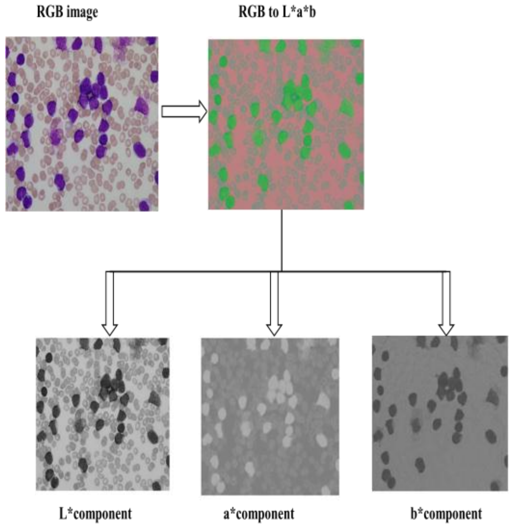


Figure 3 L*a*b Space Correlation Results

Articles	Goal	Algorithms Employed	Accuracy	Drawback
[10]	Automated Cell Nucleus Segmentation	L*a*b Color Correlation, K-means, GLCM	95%	*sub-imaging performed. *Less Robust to touching cells
[29]	Segmentation of the WBC by background suppression and create reference image	L*a*b Color Correlation, Otsu Segmentation, Automated Histogram Thresholding	92%	* segments broken white cells
[9]	Measuring nucleus boundary irregularities and extract features	L*a*b Color Correlation, Box-counting Algorithm, Contour Signature	95%	* Sub-imaging performed
[31]	Automated count of the lymphoblasts	Hue Saturation Value Color Correlation, Morphological Operation such as Erosion	97%	*Not accurate counting of blasts.
[30]	Segment WBC into nucleus & cytoplasm	Canny Edge Detection. GVF snake, Zack Thresholding	92% Nucleus 70% Cytoplasm	*Holds good only for sub-images. *Segmentation is inter-dependent.
[8]	For segregating leukocytes from other blood components	Selective filtering, L*a*b Conversion, bounding box	93%	*Sub-imaging performed *Lymphoblasts subtypes are not considered
[6]	Automated blood cell count is performed and feature extraction is performed for classification using different methods.	L*a*b Color space, Haralick features, Color Histogram	K-NN - 80.76% LVQ -83.33% MLP -89.74% SVM -91.03%	*Needs to be more robust
[32]	2-part approach for identifying cytoplasm and nucleus of WBCs in a color image	HSV Color Space, K-means , EM Algorithm	80%	Only for Sub-images. Does not handle clustered cells

Table1: Comparative Study of different Acute Lymphoblastic Leukemia systems.

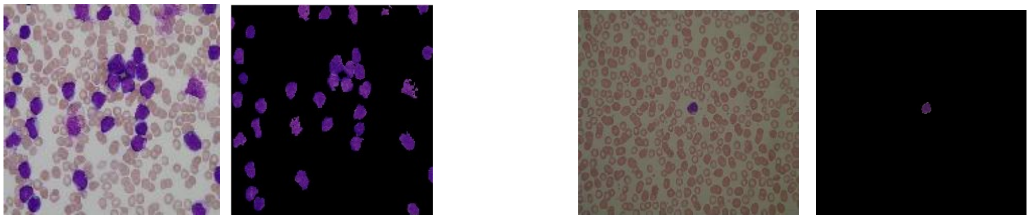
4. SEGMENTATION

Segmentation plays a key role since the efficiency of subsequent feature extraction and classification relies greatly on the correct segmentation of the blasts. Many algorithms for segmentation have been developed for gray level images. Cell segmentation using active contour models is presented in [6]; Color segmentation procedure applied to leukocyte images are described in [17]; a two step segmentation process using HSV color model is used in [15]; The use of shape analysis into WBC segmentation, by considering features such as cell size, granularity, density and shape was introduced in [16]; A watershed-based segmentation is a cell detection method that utilizes both intensity and shape information of cell to improve the segmentation was proposed by Wang et al. [18].

Segmentation is performed for extracting the nuclei of the leukocytes using color based clustering. A k-means clustering procedure is used to assign every pixel to one of the clusters. Each pixel of an object is classified into *k* clusters based on the corresponding *a and *b values in L*a*b color space. These clusters correspond to nucleus (high saturation), background (high luminance, low saturation), and other cells (e. g., erythrocytes and leukocyte cytoplasm). Every pixel is assigned to one of these classes using the properties of the cluster center. Now each pixel in the L*a*b color space is classified into any of the *k* clusters by calculating the Euclidean distance between the pixel and each color indicator. Each pixel of the entire image will be labeled to a particular color depending on the minimum distance from each indicator. We consider only the cluster which contains the blue nucleus, which is required for the feature extraction.
K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. It is composed of following steps:

1. *Place K points (initial group centroids) into the space represented by the objects that are being clustered.*
2. *Assign each object to the group that has the closest centroid.*
3. *Recalculate the positions of the K centroids, when all objects have been assigned,*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*

Several attempts were made to classify *sub-images* of acute lymphoblastic leukemia. We attempt to overcome this drawback by segmenting the complete image comprising of multiple lymphoblasts. We consider microscopic blood images of size 512 x 512. The input images after undergoing L*a*b color conversion are subjected to K-means clustering. The cluster considered contains only the blue nuclei. Figure 4 shows segmented output of the cell nucleus image after applying K-means algorithm.



Cancer Image K-means Output Normal Image K-means output

Figure 4. Examples of Segmentation using K-means Clustering

5. 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*. If the features are carefully extracted it is expected that the feature set will obtain the relevant information from the input data. For constructing an effective feature set, we studied several published articles to observe their feature selection methodology. It was observed that certain features give good results and hence are commonly used. Those features were considered to boost the classifier performance. Figure 5 shows a table representing the set of features including shape features and texture features observed by ten different systems found in the references ([1]...[10]).

From the table it can be observed that shape features such as *area*, *eccentricity and solidity;* texture features such as *energy, entropy* are extensively used for giving robust classification. These most widely used features were considered and implemented on *whole images*. Thus far, the most commonly faced drawback in Acute Lymphoblastic Leukemia Classification, is that, all procedures executed on sub-images of the blood smear consisting of only one nucleus per image. In this work, we *have overcome the drawback of using only cropped images* ***and have extracted features from complete color images*** *of blood smear comprising multiple nuclei.*

Shape Features:

The shape of the nucleus, according to hematologist, is an essential feature for discrimination of blasts. Region and boundary based shape features [10] 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 non-zero pixels.

The features considered are as follows [formulae- refer figure5]:

- *Area*: The total number of none zero pixels within the image region.
- *Perimeter*: Calculating distance between successive boundary pixels gives the perimeter
- *Compactness*: Compactness or roundedness is the measure of a nucleus.
- *Solidity*: An essential feature for blast classification- the ratio of actual area to convex hull area.
- *Eccentricity*: Since lymphocytes are more circular than the blast, eccentricity is an important feature. It is a parameter that is used to measure how much a shape of a nucleus deviates from being circular.
- *Elongation*: The nucleus bulging is measured in terms of a ratio called elongation. This is defined as the ratio between maximum distance (R_{max}) and minimum distance (R_{min}) from the center of gravity to the nucleus boundary.
- *Form-factor*: Dimensionless parameter considered which changes with surface irregularities.

GLCM Features (Harlick [35], [36])*:* Texture is defined as a function of the spatial variation in pixel intensities [1], [10]. The Gray Level Co-occurrence Matrix (GLCM) and associated texture feature calculations are image analysis techniques. Gray level pixel distribution can be described by second-order statistics like the probability of two pixels having particular gray levels at particular spatial relationships. This information can be depicted in 2-dimensional 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:

- *Energy:* Also known as uniformity or ASM, it is a measure of homogeneity of image
- *Contrast:* The contrast feature is a difference moment of the regional co-occurrence matrix and is a measure of the contrast or the amount of local variations present in an image.
- *Entropy*: This parameter measures the disorder of an image. When the image is not texturally uniform, entropy is very large
- *Correlation:* The correlation feature is a measure of regional pattern liner-dependencies in the image.

Fractal Dimension: Fractals have been used in medicine and science earlier for various quantitative measurements [38] [39]. The *fractal dimension*, *D*, 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 Rényi dimension, the Hausdorff dimension 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. Lymphoblast or a mature lymphocyte can be differentiated using perimeter roughness of the nucleus. Hausdorff dimension is considered as an essential feature considered in our proposed system. The procedure for Hausdorff dimension measurement using box counting method [40] is introduced 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 nuclei boundaries.
3. Edges are superimposed by a grid of squares.
4. The Hausdorff Dimension(HD) may then be defined as follows in (1).

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

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 implies higher degree of roughness.

The Hausdorff Dimension, turned out to be a crucial feature in our system, particularly since we considered whole images of the blood sample. In whole images, the number of nuclei under the field of view was much higher for a cancerous case as opposed to the non-cancerous case. This resulted in steep difference in box-count between the two cases and thereby proved to be an effective feature. This is illustrated in the figure 6.

FEATURES	FORMULAE	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
HF- Energy	$Energy = \sum_i \sum_j P^2(i,j)$			★		★	★		★	★	★
HF- Homogeneity	$Homogeneity = \sum_i \sum_j \frac{P_d(i,j)}{1+\lvert i-j\rvert}$			★		★			★	★	★
HF- Entropy	$Entropy = -\sum_i \sum_j p(i,j)\, log(p(i,j))$			★		★			★	★	★
HF - Correlation	$Correlation = \frac{\sum_i \sum_j (i,j)\,(p(i,j)) - \mu_x \mu_y}{\sigma_x \sigma_y}$			★		★			★	★	★
Pr.Axis Ratio	$C_{yy}+C_{xx}-\sqrt{(c_{xx}+c_{yy})^2-4(c_{xx}c_{yy}-c_{xy})^2} / C_{yy}+C_{xx}+\sqrt{(c_{xx}+c_{yy})^2-4(c_{xx}c_{yy}-c_{xy})^2}$				★						
Compactness	$Compactness = \frac{Perimeter^2}{Area}$			★			★	★	★	★	★
Compactness	$Compactness = \frac{2\sqrt{\pi Area(R)}}{Perimeter(R)}$				★						
Circular Variance	$CircVar(R) = \frac{1}{N\mu_r{}^2} \sum_i (\lVert p_i - \mu \rVert - \mu_r)^2$				★						
Elliptic Variance	$EllVar(R) = \frac{1}{N\mu_{rc}} \sum_i (\sqrt{(p_i-\mu)^T C^{-1} (p_i-\mu)} - \mu_{rc}^2)$				★						
SF - Area	*Total number of non-zero pixels within the image region*	★	★	★		★	★	★	★	★	★
SF - Perimeter	*Measured by calculating the distance between Successive boundary pixels*	★	★	★		★	★	★	★	★	★
Contrast	$Contrast = \gamma = \sum_i i^2 d_i^2$					★		★			
Convexity	$Convexity(R) = \frac{Perimeter(ConvexHull(R))}{Perimeter(R)}$				★						
SF- Eccentricity	$Eccentricity = \frac{\sqrt{a^2-b^2}}{a}$	★	★	★		★		★	★	★	★
SF- Solidity	$Solidity = \frac{Area}{Convex\ Area}$	★	★	★		★		★	★	★	★
SF- Form factor	$FormFactor = \frac{4\pi Area}{Perimeter^2}$			★				★	★	★	★
SF- Elongation	$Elongation = \frac{R_{max}}{R_{min}}$			★		★			★	★	★
Major Axis Length	*MajorAxis=a+b, where a & b are distances from each focus to any point on the ellipse*	★	★			★					
Minor Axis Length	$MinorAxis = \sqrt{(a+b)^2 - f^2}$	★				★					
Ratio	Cell Area / Nucleus Area		★			★					
Ratio	Cytoplasm Ar/ Nu Area	★				★		★			
Ratio	Nu Perim/Cell Peimeter					★					
Nucleus Rectangularity	*Perimeter of tightest bounding rectangle /nucleus perimeter*	★	★								
Cell Circularity	*Perimeter of the tightest bounding rectangle / Cell perimeter*	★	★			★		★			
Hausdorff Dimension	*N-nmber of boxes, N(s)– boxcount* **HD = log(N)/log(N(s))**										★

Figure 5. Feature Database listing the features employed by different systems

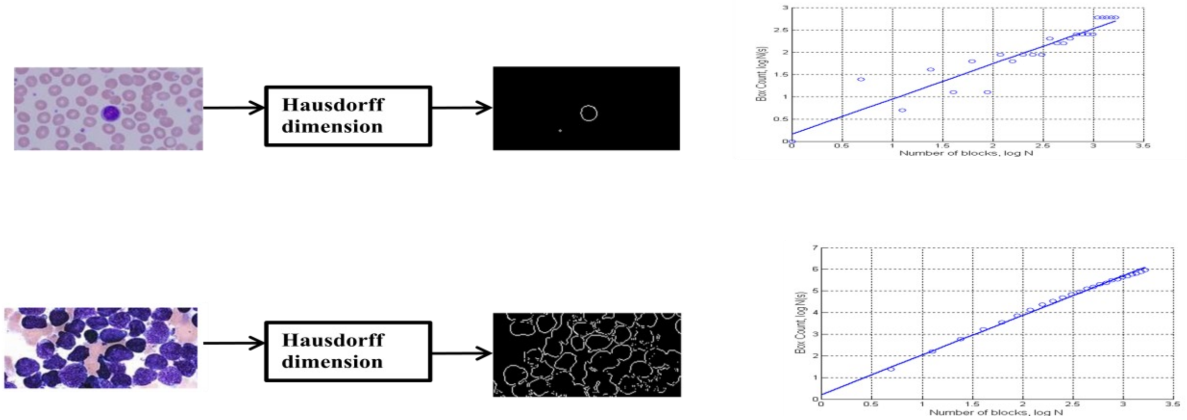


Figure 6. Fractal Dimension Results

6. COMPUTER SIMULATION AND DISCUSSION [33]

The method has been evaluated using a set of 98 images provided by Charles Fabio Scotti, Department of Information Technology - Università degli Studi di Milano. The images of the dataset have been captured with an optical laboratory microscope coupled with a Canon Power Shot G5 camera. All images are in JPG format with 24 bit color depth has a size of 512x512 pixels. The features are extracted from the segmented images and classified using the Support Vector Machine (SVM). Following which, cross-validation of the results is performed. Cross-validation is a technique for judging how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how precisely a predictive model will perform in practice. Three kinds of validation techniques have been used: *K-fold, Hold-Out and Leave-One-Out,* which have been discussed later in this section.

SVM is a classification algorithm that provides state-of-the-art performance in a wide variety of application domains [22]. The support vector classifier distinguishes between two or more classes and does not consider outliers not belonging to any of the classes. A SVM performs pattern recognition for two-class problems by determining the separating hyper plane that has maximum distance to the closest points of the training set [24,25]. These closest points are called support vectors. The SVM performs a nonlinear separation in the input space by using a nonlinear transformation *Φ(.)* that maps the data points **x** of the input space, $\acute{R}^n$, into a higher dimensional space, called kernel space $\acute{R}^p$ ($p > n$). The mapping *Φ(.)* is represented in the SVM classifier by a kernel function $K(x,x_j)$ which defines an inner product in $\acute{R}^p$. Given l samples$\{(x_i, y_i)\}_{i=1}^{l}$, the decision function of the SVM is linear in the feature space and can be written as:

$$g(x) = w\emptyset(x) + b = \sum_{i=1}^{l} \alpha_i^0 y_i K(x, x_i) + b \qquad (2)$$

The optimal hyper plane is the one with the maximal distance (in space $\acute{R}^p$) to the closest points $\Phi(x_i)$ of the training data. Determining the hyper plane requires maximizing the following function with respect to α

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - 1/2 \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i\, \alpha_j y_i y_j K(x_{i,} x_j) \qquad (3)$$

where under constraints $\sum_{i=1}^{l} \alpha_i\, y_i = 0, C \geq \alpha_i \geq 0, i = 1 \ldots l$ and $\alpha = (\alpha_1 \ldots . \alpha_l)$ is the non-negative Lagrange multipliers. The indexes of α_i which have nonzero values when a solution as described below is found to correspond to the support vectors. Writing *W(α)* in matrix notation, incorporating non-negativity of α and constraint, the following dual quadratic program is defined:

$$\textit{Maximize} \qquad W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{\sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_{i,} x_j)}{2} = \alpha.1 - (\alpha H \alpha)/2 \qquad (4)$$

Subject to: $\alpha . y^t = 0$
$\alpha \geq 0$

where $y = (y_1 \dots . y_l)$ and H is a symmetric l x l matrix with elements $H_{ij} = y_i y_j K(x_i, x_j)$, which is a Hessian [33].

An upper bound on the expected error probability P_{error} of a SVM classifier is given by

$$g(x) = w\emptyset(x) + b = \sum_{i=1}^{l} \alpha_i K(x, x_i) + b \qquad (5)$$

where w=$(w_1 \dots . w_p)$ The contribution of a feature x_i to the decision function in equation 10 depends on α_i. In [25,26], $f(\mathbf{x})$ is used to denote the current hypothesis which is determined by the values of the dual variable and the bias, b, at a particular stage of learning. The error, E, is then calculated as the difference between the function output ant the target classification on the training points x_i and the support vectors x_j as follows:

$$E_i = g(x_i) - y_i = \left(\sum_{j=1}^{l} \alpha_j K\left(x_i, x_j\right) + b\right) - y_i \qquad (6)$$

It should be noted that the value of E_i may be large for a correctly classified input feature x_i . To minimize the chance of accepting outliers, the volume of this hyper sphere is minimized [25].

Cross-Validation Methods [34]
Cross-Validation is essentially performed to compare the performance of two or more different algorithms and choose the one that is the best for the given set of data. A statistical method of evaluating and comparing learning algorithms by dividing data into two sets: one used to learn or train a model and the other used to validate the model. Evaluating or comparing learning algorithms using cross-validation is performed as follows: in each iteration, one or more learning algorithms use *k-1* folds of data to learn one or more models, and subsequently the learned models are asked to make predictions about the data in the validation fold. The performance of each learning algorithm on each fold is determined using some predetermined performance metric like efficiency or accuracy. For each algorithm, upon completion, *k* samples of the performance metric will be available.

Hold-Out Validation: An independent test set is used, to avoid over-fitting. This can be achieved by splitting the available data into two non-overlapped parts: one for testing and the other for training. The test data is held out and not looked at during training. Hold-out validation avoids the overlap between training data and test data, yielding a more accurate performance. The drawback is that this procedure does not use all the available data and the results are highly dependent on how the testing and training data are being split. Also, the data in the test set may be valuable for training and if it is held out prediction performance may suffer, again leading to improper results. One solution to overcome this problem is by running the set multiple times and averaging the results, but unless this repetition is performed in a systematic manner, some data may be included in the test set multiple times while others are not included at all, or conversely some data may always fall in the test set and never get a chance to contribute to the learning phase. K-fold cross-validation is used to deal with these challenges and utilize the available data to the max.

K-Fold Cross-Validation: In k-fold cross-validation the data is first partitioned into k equally (or nearly equally) sized sets or folds. Subsequently *k* iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining *k-1* folds are used for learning. Prior to being split into *k* folds, the data is commonly stratified. The process of rearranging the data as to ensure each fold is a good representative of the whole is termed as Stratification.

Leave-One-Out Cross-Validation: Leave-one-out cross-validation (LOOCV) is a special case of *k*-fold cross-validation where *K* represents the number of instances in the data. That is, in every iteration, nearly all the data except for a single observation are used for training and the model is tested on that particular single observation. An accuracy estimate obtained using LOOCV is known to be almost unbiased but it has high variance, leading to unreliable estimates [36].
Figure 9(a) depicts the performance achieved by each of the above cross-validation method.

7. ANALYSIS

Based on validation, it was observed that for most classes the classification accuracy is around 93.5%. The cross-validation method followed was *Leave-M-out cross-validation.* For improved classification accuracy it should be considered to include only “typical” images in the training set. Alternative training procedures which take into account the continuous (rather than discrete) nature of the class membership or hierarchical classifiers could also be considered. Even though SVMs reduce over fitting problems it is still desirable to increase the number of training samples. Feature selection can also be considered for future improvements of the algorithm.

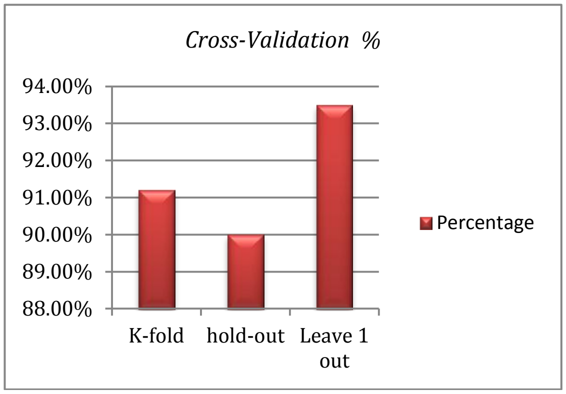


Figure 9(a)

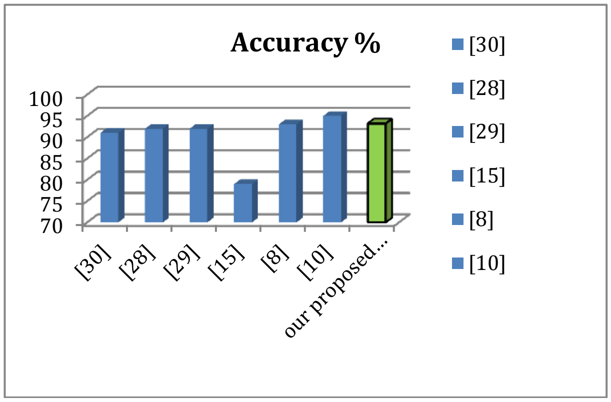


Figure 9(b)

Figure 9(a): Comparative analysis of different Cross-Validation Methods; Figure 9(b): Performance evaluation with existing systems; accuracy obtained by systems implementing on sub-images (as stated in their journal) against our proposed system which implements whole images

8. CONCLUSION & FUTURE WORK

All the discussed features are extracted for 98 complete blood smear images. Our goal is to provide a unique set of features for better classification. We have tried to identify such features which are basically followed by hematologists. The results obtained in terms of features are also verified by an expert. The SVM classifier gives a performance of **93.5%** on 98 different sets of images. The advantage of the proposed scheme over existing schemes is that, *our proposed system robustly classifies the complete blood smear images containing multiple nuclei,* while existing systems mostly consider those images which only have one lymphocyte under the field of view. We extract shape features and texture features of the complete images with multiple nuclei as a whole and classify based on the results obtained. The features extracted in our proposed scheme can be used for training any classifier for further classification. Furthermore the system should be robust to excessive staining and touching cells. Obtained results encourage future works which includes classification of lymphoblast into various subtypes.

9. ACKNOWLEDGEMENT

This research has been partially supported by NSF Grant(HRD-0932339) to the University of Texas San Antonio. We are also thankful to Charles Fabio Scotti, Department of Information Technology - Università degli Studi di Milano, for providing us with high quality image database.

REFERENCES

[1] *Fabio Scotti* , "Automatic Morphological Analysis for Acute Leukemia Identification in Peripheral Blood Microscope Images", in Computational Intelligence for Measurement Systems and Applications, CIMSA .pages 96-101,(2005),http://piurilabs.dti.unimi.it/Papers/01522835.pdf

[2] *Vincenzo Piuri, Fabio Scotti* "Morphological Classification of Blood Leucocytes by Microscope Images", in *Computational Intelligence for Measurement Systems and Applications.* CIMSA. pages 103-108.,(2004)

[3] *Subrajeet Mohapatra, Dipti Patra, Sanghamitra Satpathy "*Automated Leukemia Detection in Blood Microscopic Images using Statistical Texture Analysis", ICCCS '11 Proceedings of the 2011 International Conference on Communication, Computing & Security,(2011)

[4] *Herbert Ramoser, Vincent Laurain, Horst Bischof, Rupert* Ecker, Adv. Comput. Vision GmbH, Wien, "Leukocyte segmentation and classification in blood-smear images", in Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005.pages 3371-3374 **,**(2005).

[5] *Carolina Reta, Leopoldo Altamirano, Jesus A.Gonzalez, Raquel Diaz, Jose S. Guichard "*Segmentation of Bone Marrow Cell Images for Morphological Classification of Acute Leukemia", Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010).

[6] *Guclu Ongun, Ugur Halici, Leblebicioglu, K.; Atalay, V.; Beksac, M.; Beksac, S.;* Dept. of Electr. & Electron.Eng., Middle East Tech.Univ., Ankara, "Feature Extraction and Classification of Blood Cells for an Automated Differential Blood Count System" in Neural Networks, 2001. Proceedings. IJCNN '01. Pages 2461-2466 vol.4.,(2001)

[7] *Ghosh, M.; Das, D.; Mandal, S.; Chakraborty, C.; Pala, M.; Maity, A.K.; Pal, S.K.; Ray, A.K.*; Sch. of Med. Sci. & Technol., IIT Kharagpur, Kharagpur, India , "Statistical Pattern Analysis of White Blood Cell Nuclei Morphometry", in Students' Technology Symposium (TechSym), 2010 IEEE. Pages 59-66. (2010)

[8] *Mohapatra, S.; Samanta, S.S.; Patra, D.; Satpathi, S.*; IPCV Lab., Nat. Inst. of Technol., Rourkela, India, "Fuzzy Based Blood Image Segmentation for Automated Leukemia Detection", in Devices and Communications (ICDeCom), Pages 1-5. (2011)

[9] *Mohapatra, S.; Patra, D.; Satpathi, S.;* Dept. of Electr. Eng., Nat. Inst. of Technol. Rourkela, Rourkela, India, "Image Analysis of Blood Microscopic Images for Acute Leukemia Detection", in Industrial Electronics, Control & Robotics (IECR), 2010. Pages 215-219. (2010)

[10] *Mohapatra, S.; Patra, D.; Satpathi, S.;* Dept. of Electr. Eng., Nat. Inst. of Technol. Rourkela, Rourkela, India, "Automated Cell Nucleus Segmentation and Acute Leukemia Detection in Blood Microscopic Images", in Systems in Medicine and Biology (ICSMB), Pages 49-54. (2010).

[11] *J.N. Jameson, L.K. Dennis, T.R. Harrison, E. Braunwald, A.S. Fauci, S.L. Hauser, D.L.Longo*, Harrison's principles of internal medicine. New York: McGraw-Hill Medical Publishing Division, (2005.)

[12] Serbouti, S.;Duhamel, A.; Harms, H.; Gunzer, U.; Aus, U.M.; Mary, J.Y.;Beuscart, R.; "Image segmentation and classification methods to detect leukemias," in *Proc. International conference of IEEE engineering in Medicine and Biology society*, 1991. Pages 260-261.(1991)

[13] *Foran, D.J.; Comaniciu, D.;Meer, P.; Goodell, L.A.*; Centre for Biomed. Imaging & Inf, UMDNJ, Piscataway, NJ, USA , "Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy.," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 4, pp. 265–273, (2000).

[14] *K.S. Kim, P.K. Kim, J.J. Song, and Y.C. Park,* "Analyzing blood cell image do distinguish its abnormalities," in *Proc. ACM International Conference on Multimedia*, (2002).

[15] *Sinha, N.; Ramakrishnan, A.G.*; Dept. of Electr. Eng., Indian Inst. of Sci., Bangalore, India, "Automation of differential blood count". In *Proceedings Conference on Convergent Technologies for Asia-Pacific Region*, volume 2, pages 547–551, (2003).

[16] *Qingmin Liao; Yingying Deng*; Dept. of Electron. Eng., Tsinghua Univ., Beijing, China "An accurate segmentation method for white blood cell images". in *Proc. IEEE Int. Symp. on Biomedical Imaging*, (2002).

[17] *J. S. Suri, S. K. Setarehdan, and S. Singh,* "Advanced algorithmic approaches to medical image segmentation: state-of-the-art application in cardiology, neurology, mammography and pathology", 541 –558. Springer, (2001).

[18] *Wang Shitong, Korris F.L. Chungb, Fu Duana*, "Applying the improved fuzzy cellular neural network IFCNN to white blood cell detection.", Neurocomputing archive. Vol.70 ,no. 7-9,March (2007).

[19] *Nilsson, B.; Heyden, A.*, "Model-based segmentation of leukocytes clusters," in *Proc. of International Conf. on Pattern Recognition*, pp. 727–730, vol.1,(2002).

[20] *P. Bamford and B. Lovell*, "Method for accurate unsupervised cell nucleus segmentation," in *Proc. of the Engineering in Medicine and Biology Society Conference* , pages 2704-2708, vol.3,(2001).

[21] CIE L*a*b Color Scale – Hunter Lab, http://www.hunterlab.com/appnotes/an07_96a.pdf.

[22] *C.J.C. Burges*, "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery,* vol.2, No.2, pp.121-167,(1998)

[23] *Hearst, M.A.; Dumais, S.T.; Osman, E.; Platt, J.; Scholkopf, B.*, "Support Vector Machines", *IEEE Intelligent Systems*, pp. 18-28, (1998)

[24] *B. Scholkopf, A. J. Smola*, "Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond", The MIT Press, (2002)

[25]. *N. Cristianini, J. Shawe-Taylor*, "An Introduction to Support Vector Machines and other kernel-based learning Methods", Cambridge University Press 2000

[26].*C.-W.Hsu, C.-C. Chang, C.-J. Lin*. "A practical guide to support vector classification", vol.1, pages 1-16, http://www.mendeley.com/research/a-practical-guide-to-support-vector-classi-cation/

[27] *D.M.J. Tax and R.P.W. Duin*, "Support Vector Data Description", Machine Learning, vol. 54, no. 1, pages 45-66,(2004).

[28] *Efron B*,"Estimating the error rate of a prediction rule: Improvement on cross-validation." J. Am. Stat. Assoc., vol.78:316–331,(1983).

[29] *Fabio* Scotti, Dept. of Inf. Technol., Milan Univ., Crema , "Robust Segmentation and Measurements Techniques of white cells on Blood Microscopic Images", Proc. of the IEEE in Instrumentation and Measurement Technology Conference, 2006. IMTC 2006, pages 43-48,(2006).

[30] *Abdul Rahman Ramli*, "A framework for White Blood Cells Classification", vol.11, pages 196-206,(2009).

[31] *Nurul Hazwani Abd Halim, Mohd Yusoff Mashor, and Rosline Hassan*, "Automatic Blasts Counting for Acute Leukemia Based on Blood Samples",vol.2, No.4.

[32]*Neelam Sinha, A.G.Ramakrishnan,"*Blood Cell Segmentation Using EM Algorithm", http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.9955

[33] *Benjamin M. Rodriguez, Gilbert L. Petersona, Sos S. Agaian, "*Steganography Anomaly Detection Using Simple One-Class Classification", Proceedings of SPIE (2007).

[34] *Payam Refaeilzadeh, Lei Tang, Huan Liu*, Arizona State University , "Cross-Validation", in Encyclopedia of Database Systems (EDBS), Editors: Ling Liu and M. Tamer Özsu. Springer, pp6. (2009).

[35] *R. M.Harlick*, "Statistical and structural approaches to texture" *Proceedings of the IEEE*, 67(5), (1979).

[36] *R. M.Harlick, K. Shanmugan, and I. Dinstein*. "Textural features for image classification". *IEEE Transaction on Systems, Man, and Cybernatics*, 29(2), (1973).

[37]*J.Poomcokrak and C.Neatpisarnvanit*, " Red Blood Cells Extraction and Counting", http://www.kmitl.ac.th/ijabme/proceedings/bmeicon08/pdf/Session4/1105.pdf

[38] *B. B. Mandelbrot*, "How long is the coast of Britain? Statistical self similarity and fractional dimension" Science, vol.156:636 – 638, (1967).

[39] *B. T. Milne*, "Measuring the fractal geometry of landscapes." Applied Mathematics and Computation, 27:67 – 79, (1988).

[40] *A. P. Pentland*, "Fractal based description of natural scenes ." IEEE Transactions on Pattern Analysis and Machine Intelligence, 6:661 – 674, (1984).