ALL-IDB: THE ACUTE LYMPHOBLASTIC LEUKEMIA IMAGE DATABASE FOR IMAGE PROCESSING

Ruggero Donida Labati, Vincenzo Piuri, Fabio Scotti

Università degli Studi di Milano, Department of Information Technology,
via Bramante 65, 26013 Crema, Italy

ABSTRACT

The visual analysis of peripheral blood samples is an important test in the procedures for the diagnosis of leukemia. Automated systems based on artificial vision methods can speed up this operation and increase the accuracy and homogeneity of the response also in telemedicine applications. Unfortunately, there are not available public image datasets to test and compare such algorithms. In this paper, we propose a new public dataset of blood samples, specifically designed for the evaluation and the comparison of algorithms for segmentation and classification. For each image in the dataset, the classification of the cells is given, as well as a specific set of figures of merits to fairly compare the performances of different algorithms. This initiative aims to offer a new test tool to the image processing and pattern matching communities, direct to stimulating new studies in this important field of research.

Index Terms— Acute lymphoblastic leukemia, public image database, image segmentation, image classification.

1. INTRODUCTION AND RELATED WORKS

Acute Lymphocytic Leukemia (ALL), also known as acute lymphoblastic leukemia, is an important hematic diseases. It is fatal if left untreated due to its rapid spread into the bloodstream and other vital organs, and it mainly affects young children and adults over 50. Early diagnosis of the disease is crucial for the recovery of patients, especially in the case of children. The symptoms of ALL are common also in other diseases and for this reason, the diagnosis is very difficult. One of the steps in the diagnostic procedures encompasses the microscope inspection of peripheral blood. The inspection consists in the search of white cells with malformations due to the presence of a cancer. For decades, this task has been performed by experienced operators, which basically perform two main analyses: the cell classification and counting (now performed by cytometers). Interestingly, the morphological analysis does not require a blood sample because can be performed by using a single image. For this reason, this analysis is suitable for low-cost, homogenous accuracy, remote screening systems.

Only few attempts of partial/full automated systems for leukemia detection based on image-processing systems are present in the literature [1 - 3]. In particular, some works have been proposed to segment white cells [4, 5], to refine the segmentation (i.e., to correctly segment clusters of cells) [6, 7], or to detect incorrect segmentations of white cells [8]. A system for the classification of single white cells is presented in [9]. A complete classification system to detect the acute leukemia from blood images is proposed in [10], based on some morphological features of the gray level images. The cell type classification by using artificial neural networks and morphological operators has been treated more in particular in [11]. The work presented in [12] discusses methods to enhance the microscope images by removing the undesired microscope background components, a method for a robust estimation of the mean cell diameter and a new fully self-adaptive segmentation strategy to robustly identify white cells. Results indicated that the morphological analysis of white cells offers remarkable classification accuracy (about 92%). In the literature, there are other works for the ALL recognition based on different approaches such as the analysis of gene expression [13], hemocytometer statistics [14] and holographic microscope images [15].

At the best of our knowledge, there are not available public supervised image datasets to test and fairly compare algorithms for cell segmentation and classification of the ALL disease. In this paper, we present the ALL-IDB, a public image dataset of peripheral blood samples of normal individuals and leukemic patients, which provides a supervised classification and segmentation of the data. The samples have been collected by experts of M. Tettamanti Research Center for childhood leukemia and hematological diseases, Monza, Italy. The paper is structured as follows. Section 2 presents the process for the acquisition of the blood samples and the classification of the cells. Section 3 describes the two versions of dataset, and the characteristics of the images. Section 4 proposes metrics to evaluate the accuracy of systems for the automatic detection of blasts.

2. MORPHOLOGICAL ANALYSIS OF CELLS

A typical blood microscope image is plotted in Fig. 1. The principal cells present in the peripheral blood are red

blood cells, and white cells (leucocytes). Leucocyte cells containing granules are called granulocytes (composed by neutrophil, basophil, eosinophil). Cells without granules are called agranulocytes (lymphocyte and monocyte). The percentage of leucocytes in human blood typically ranges between the following values: neutrophils 50-70%. eosinophils 1-5%. basophils 0-1%. monocytes 2-10%, lymphocytes 20.45% [16].

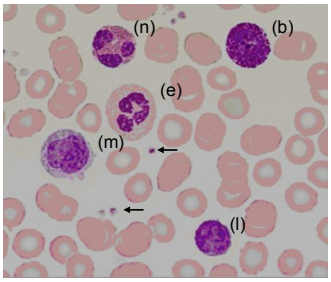

Fig. 1. Blood white cells marked with colorant: basophil (b), eosinophil (e), lymphocyte (l), monocyte (m), and neutrophil (n). Arrows indicate platelets. Others elements are red cells.

The ALL disease is related to the lymphocytes in the bone marrow and in the peripheral blood. The colorant used in the preparation of the blood tends to concentrate only in white cells, in particular in their nuclei that are typically center-positioned (the darker elements in Fig. 1). In most of the cases, the white cells are also bigger than the red cells. The most common leukemia classification by morphological analysis is the FAB method [17], even if nowadays it has been updated with the immunologic classification [18], which it is not image-based. Differently from the FAB method (requiring only a microscope), the immunologic classification needs a more sophisticated setup for the procedure.

Usually, an automatic method for the detection of lymphoblasts in microscopically color images can be divided in the sequent steps.

- **Segmentation** - The cells are separated from the background by using algorithms based on different characteristics of the cells (e.g. shape, color, inner intensity).
- **Identification of white cells** - The cells are classified in white cells and red cells. The classifier can search the presence of the nucleus by using color information.
- **Identification of lymphocytes** - The lymphocytes can be distinguished from the other white cells by analyzing the shape of the nucleus (e.g., a deeply staining nucleus which may be eccentric in location, or a small amount of cytoplasm).
- **Identification of candidate lymphoblasts** - Candidate lymphoblasts can be identified in a set of lymphocytes by the analysis of morphological deformations of the cells.

In particular, lymphocytes present a regular shape, and a compact nucleus with regular and continuous edges. Instead, lymphoblasts present shape irregularities. Concerning the ALL, the candidate lymphoblasts are analyzed by using the FAB classification as follows.

- **L1** - Blasts are small and homogeneous. The nuclei are round and regular with little clefting and inconspicuous nucleoli. Cytoplasm is scanty and usually without vacuoles.
- **L2** - Blasts are large and heterogeneous. The nuclei are irregular and often clefted. Usually, one or more large nucleoli are present. The volume of the cytoplasm is variable, but often abundant and may contain vacuoles.
- **L3** - Blasts are moderate-large in size and homogeneous. The nuclei are regular and round-oval in shape. One or more prominent nucleoli are present. The volume of cytoplasm is moderate and contains prominent vacuoles.

Fig. 2 shows the great variability in shape and pattern of the blast cells, according to the FAB classification.

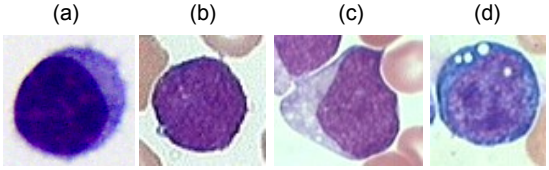

Fig. 2. Morphological variability associated to the blast cells according to the FAB classification: (a) a healthy lymphocyte cell from a non-ALL patient, (b-d) lymphoblasts from ALL patients where (b), (c) and (d) are L1, L2 and L3 respectively.

3. THE DATASET

All images in the datasets are in JPG format with 24 bit color depth, and a native resolution equal to 2592 × 1944, captured with a PowerShot G5 camera. The images are related to different magnifications of the microscope (ranging from 300 to 500). The ALL-IDB database has two distinct versions (ALL-IDB1 and ALL-IDB2) focused on segmentation and classification, which can be freely downloaded from [19]. Table 1 shows the characteristics of the images stored in the two datasets.

Table 1: Characteristics of the dataset.

Image Acquisition Setup		
Camera: Canon PowerShot G5		
Magnification of the microscope: 300 to 500		
Image format: JPG		
Color depth: 24 bit		
	ALL-IDB1	**ALL-IDB2**
Images:	109	260
Resolution:	2592 × 1944	257 × 257
Elements:	39000	260
Candidate lymphoblasts:	510	130

3.1. Description of ALL-IDB1

The ALL-IDB1 can be used for testing the segmentation capability of algorithms, as well as the accuracy of classification systems. This dataset is composed by 108 images collected during September 2005.

It contains about 39000 blood elements, where the lymphocytes have been labeled by expert oncologists. The number of candidate lymphoblasts present in the ALL-IDB1 is equal to 510. Only the lymphoblasts that are completely described in the image have been considered and classified.

Fig. 3 shows two example images belonging to the ALL-IDB1. The blood in the first three images was taken from healthy people and the blood of the last three images was taken from people affected by ALL.

The annotation of ALL-IDB1 is as follows. The ALL-IDB1 image files are named with the notation "ImXXX_Y.jpg" where XXX is a 3-digit integer counter and Y is a Boolean digit equal to 0 if no blast cells are present, and equal to 1 if at least one blast cell is present in the image. Please note that all images labeled with Y=0 are from healthy individuals, and all images labeled with Y=1 are from ALL patients.

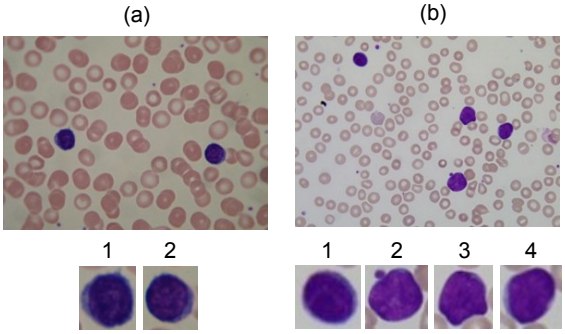

Fig. 3. Examples of the images: healthy blood (a), blood with ALL blasts (b). (a1-2) and (b1-4) are zoomed subplots of the (a) and (b) images centered on lymphocytes and lymphoblasts, respectively.

For each image "ImXXX_Y.jpg", a corresponding classification text file "ImXXX_Y.xyc" is given. This file contains the centroid coordinates of each candidate lymphoblast. The centroid is manually estimated by a skilled operator. Each row of this file is related to a single cell. Fig. 4 shows a region of interest of an image and the corresponding classification.

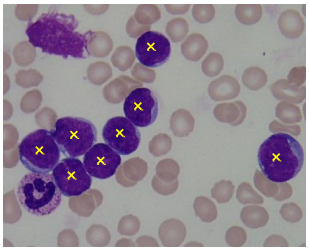

Fig. 4. Examples of a classified blood image portion. Each cross represents the centroid of a lymphoblast stored in the corresponding classification file.

The ALL-IDB1 images of blood suffer from a typical non-uniform background illumination (visible in Fig. 5). Even if the images still remain intelligible, segmentation methods based on basic thresholding can heavily suffer of this condition.

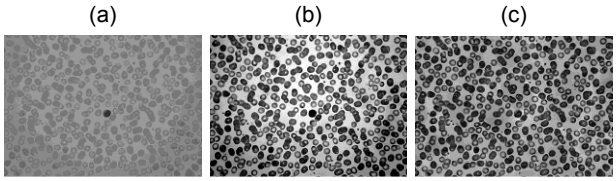

Fig. 5. Images of the ALL-IDB1 are typically bimodal and a non-uniform background can be present: (a) an example image; (b) the effect of a histogram stretching and the relative worsening of the vignetting effect present in the microscope image; (c) the result obtained by applying a dedicated image enhancement method.

3.2. Description of ALL-IDB2

This image set has been designed for testing the performances of classification systems. The ALL-IDB2 is a collection of cropped area of interest of normal and blast cells that belongs to the ALL-IDB1 dataset. It contains 260 images and the 50% of these represent lymphoblasts. ALL-IDB2 images have similar gray level properties to the images of the ALL-IDB1. The dataset is public and free available. Fig. 6 shows an example of the ALL-IDB2 images plotting 4 normal white blood cells and 4 probable blast cells. The images of the ALL-IDB2 dataset are named "ImXYZ_1.jpg" if the central cell is a probable blast, and "ImXYZ_0.jpg" in the other cases.

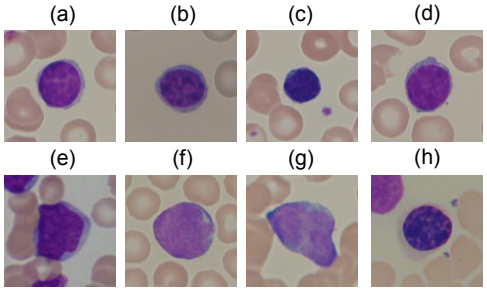

Fig. 6. Example images contained in the ALL-IDB2: healthy cells from non-ALL patients (a-d), probable lymphoblasts from ALL patients (e-h).

4. ACCURACY OF ALGORITHMS ON ALL-IDB

A system capable to identify the presence of blast cells in the input images can work with different structures of modules. For example, it can process the following steps: (i) the identification of white cells in the image, (ii) the selection of lymphocytes, and (iii) the classification of tumor cells. Each single step typically contains segmentation/ classification algorithms. In order to measure and fairly compare the identification accuracy of different structures of modules, we propose a benchmark approach partitioned in two different tests, as follows.

- **Cell test** – The benchmark accounts for the classification of single cells is blast or not (the test is positive if the considered cell is blast cell or not).
- **Image test** – The whole image is classified (the test is positive if the considered image contains at least one blast cell or not).

For each level of the benchmark, it can be processed the confusion matrix. The term "elements" refers to the cells/images of the corresponding level. Hence, it is possible to evaluate: True Positives (*TP*) as the number of elements correctly classified as positive by the test; True Negatives (*TN*) as the number of elements correctly classified as negative by the test; False Positive (*FP*) as the number of elements classified as positive by the test, but they are not; False Negative (*FN*) as the number of elements classified as negative by the test, but they are not.

Using these definitions, it is possible to process the following standard parameters: *Sensitivity* as the probability of correctly classifying elements with ALL, equals to $TP\ /\ (TP + FN)$; *Specificity* as the probability of correctly classifying elements without ALL, computable as $TN\ /\ (\ TN + FP)$; *Classification Error* as the total error in an analysis layer, $CE\ = (TP + TN)\ /\ (FP\ +\ FN)$. Table 2 is an example that can be used to report the performances of an image processing method tested with the two ALL-IDB levels of analysis (cells and images).

Table 2: Proposed set of figures of merit for the ALL-IDB 1 and 2.

Figures of Merit	Classified Element	
	Cells	Images
TP%		
TN%		
FP%		
FN%		
Classification error %		
Specificity %		
Sensitivity %		

For instance, a trivial classifier, which considers all the population as appertaining to the most frequent class, obtains a classification error equal to 45.37% by performing the image test on the ALL-IB1. The classification errors obtained by performing the cell test on the ALL-IDB1 and ALL-IDB2 are about 1.3% and equal to 50% respectively.

If the tested method requires the use of calibration/ training data, it is necessary to evaluate the obtained results by using the remaining data of ALL-IDB (e.g., the N-fold validation technique). In case of repeated tests, it is important to report the standard deviation of the obtained classification error and figures of merit.

4. CONCLUSIONS

In this paper, we proposed a public dataset of blood samples, specifically designed for the evaluation and comparison of the performances of algorithms for segmentation and image classification. We have also examined the actual state of the art related to the automatic systems for the detection of ALL, and proposed a metric to evaluate the performances of these algorithms.

We strongly discourage the use of the ALL-IDB content for diagnostic or different activities than the purpose of this initiative. ALL-IDB must be considered as an image processing dataset. We hope that the presented dataset could help to give birth to new studies in this important field of research under a fair comparative presentation of the results.

5. REFERENCES

[1] D.J. Foran, et. al., "Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy", *IEEE Transactions on Information Technology in Biomedicine*, pp. 265-73 vol. 4, 2000.

[2] K.S. Kim, P.K. Kim, J.J. Song, and Y.C. Park, "Analyzing Blood Cell Image do Distinguish its Abnormalities" *ACM InternationalConference on Multimedia*, Los Angeles, CA, USA, pp. 395-397, 2002.

[3] V.A. Kovalev, A.Y. Grigoriev, and H. Ahn, "Robust recognition of White Blood Cell Images", *Proceedingsof the InernationalConferenceon Pattern Recognition*, Vienna, pp. 371-375 vol. 4, 1996.

[4] H. Nor Hazlyna, et al., "Comparison of acute leukemia Image segmentation using HSI and RGB color space," *10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, pp.749-752, May 2010.

[5] F. Sadeghian, et al.,"A Framework for White Blood Cell Segmentation in Microscopic Blood Images Using Digital Image Processing", *Biological Procedures Online*, Springer New York, pp. 196 - 206, 2009.

[6] B. Nilsson, and A. Heyden, "Model-based segmentation of leukocytes clusters", *International Conference on Pattern Recognition*, pp. 727- 730 vol. 1, 2002.

[7] B. Prasad, J.S. Iris Choi, and W. Badawy, "A High Throughput Screening Algorithm for Leukemia Cells", *IEEE Canadian Conference on Electrical and Computer Engineering* , Ottawa, pp. 2094-2097, 2006.

[8] P. Bamford, B. Lovell, "Method for Accurate Unsupervised Cell Nucleus Segmentation", *Proceedingsof the 23rd Annual Conferenceof IEEE Engineering in Medicine and Biology Society*, pp. 2704-2708, 2001.

[9] C. Reta, L. A. Robles, J. A. Gonzalez, R. Diaz, and J. S. Guichard, "Segmentation of Bone Marrow Cell Images for Morphological Classification of Acute Leukemia", *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference*, May 2010.

[10] V. Piuri, and F. Scotti, "Morphological Classification of Blood Leucocytes by Microscope Images", *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, pp. 103-108, 2004.

[11] F. Scotti, "Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images", *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, pp. 96-101, 2005.

[12] F. Scotti, "Robust Segmentation and Measurements Techniques of White Cells in Blood Microscope Images", *Proceedingsof the IEEE Instrumentation and Measurement Technology Conference*, pp. 43-48, 2006.

[13] G. Ji, Z. Yang, W. You, W, "PLS-Based Gene Selection and Identification of Tumor-Specific Genes," *IEEE Transactions onSystems, Man, and Cybernetics, Part C: Applications and Reviews*, no.99, pp.1-12, October 2010.

[14]B. Prasad, and W. Badwy, "High Throughput Algorithm for Leukemia Cell Population Statistics on a Hemocytometer*", IEEE Biomedical Circuits and Systems Conference*, pp. 27-30, 2007.

[15] M. R. Asadi, et al. "Leukemia Cell Recognition with Zernike Moments of Holographic Images", *IEEE Proceedings of 7th Nordic Signal Processing Symposium*, Iceland, pp. 214-217, 2006.

[16] Keith Breden Taylor, and Julian B. Schorr, "Blood Vol. 4", Collier's Encyclopedia, 1978.

[17] J.M. Bennett et al. "Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group", Br J Haematol, Vol. 4, No. 33, pp. 451-458, August, 1976.

[18] Biondi, G. Cimino, R. Pieters, and C. H. Pui, "Biological and therapeutic aspects of infant leukemia", *Blood*, Vol. 96, No. 1, pp. 24-33, July, 2000.

[19] ALL-IDB web site, R. Donida Labati, V.Piuri, and F.Scotti, UniversitàdegliStudi di Milano, Departement of InformationTechnology, http://www.dti.unimi.it/fscotti/all.