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

Ruggero Donida Labati IEEE Member, Vincenzo Piuri IEEE Fellow, Fabio Scotti IEEE Member

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

ABSTRACT

The visual analysis of pheripheral 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 they can increase the accuracy 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 cell is given, and it is provided a specific set of figures of merits to be processed in order to fairly compare different algorithms when working with the proposed dataset. We hope that this initiative could give a new test tool to the image processing and pattern matching communities, aiming at stimulate new studies in this important field of research.

Index Terms— Acute lymphoblastic leukemia, Public Image Database, image segmentation, image classification

1. INTRODUCTION

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 disease 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 on the research of white cells with malformation due to the presence of a cancer. From decades, this operation is performed by experienced operators, which basically perform two main analyses: the cell classification and counting (now performed by cytometers). Interestingly, the morphological analysis just requires an image, not a blood sample and hence is suitable for low-cost, standard-accurate, and remote screening systems.

Only few attempts of partial/full automated systems for leukemia detection based on image-processing systems are present in literature [1, 2, 3]. In particular, some works have been proposed to segment [4, 5], to refine the segmentation (i.e., to correctly segment clusters of cells) [6, 7] or to detect incorrect segmentations of white cells as the method proposed in [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 image working based only on morphological features and using only gray level images is proposed in [10]. The cell type classification by using artificial neural networks and morphological operators is more in particular treated in [11]. The work presented in [12] exposes methods to enhance the microscope images by removing the undesired microscope background components, a method for the 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 blood's white cells can offer remarkable classification accuracy (about 92%). In the literature, there are present 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 ALL-IDB, a public image dataset of peripheral blood samples of normal individuals and leukemic patients and the relative supervised classification and segmentation data. These samples have been collected by experts of M. Tettamanti Research Center for childhood leukemias and hematological diseases, Monza, Italy.

The paper is structured as follows. Section 2 describes the two versions of dataset, the images characteristics, the process for the acquisition of the blood samples and the classification of the cells. Section 3 proposes the metrics for the evaluation of the automatic detection of the 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

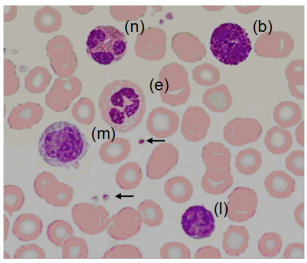

Fig. 1. Blood's white cells marked with colorant: basophil (b), eosinophil (e), lymphocyte (l), monocyte (m), and neutrophil (n). Arrows indicate platelets. Others elements are red cells.

cells, and the 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].

The ALL disease is related to the lymphocytes in the bone marrow and into 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 case, 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 than 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 classifiers 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, and 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 cell.

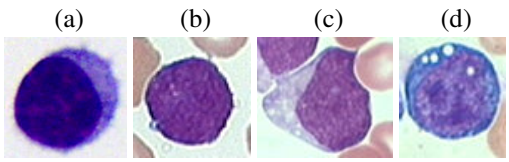

Fig. 2. Morphological variability associated to the blast cells according to the FAB classification: (a) healthy lymphocytes cell from non-ALL patients, (b-d) lymphoblasts from ALL patients where (b), (c) and (d) are L1, L2 and L3 respectively.

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. One or more, usually large nucleoli are present. The volume of 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.

3. THE DATASET

The images of the dataset has been captured with an optical laboratory microscope coupled with a Canon PowerShot G5 camera. All images are in JPG format with 24 bit color depth, resolution 2592×1944. The images are taken with different magnifications of the microscope ranging from 300 to 500. The ALL-IDB database has two distinct versions(ALL-IDB1 and ALL-IDB2) which can be freely downloaded from [19].

Table 1 shows a short description of the images.

3.1. Description of ALL-IDB1

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

It contains about 39000 blood elements, where the lymphocytes has been labeled by expert oncologists. The number

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-IDB1**
Images:	109	260
Resolution:	2592 × 1944	257 × 257
Elements:	39000	260
Candidate lymphoblasts:	510	130

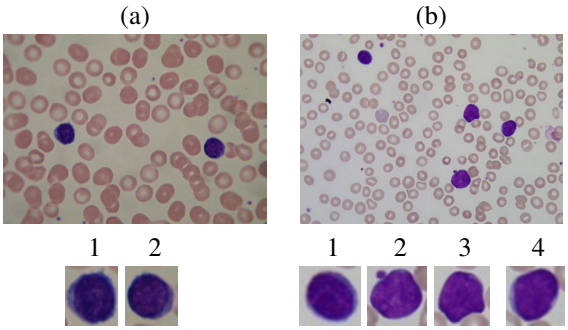

Fig. 3. Examples of the images: healthy blood (a), blood with ALL blasts (b). (a1-2) and (b1-4) are zoomed subplots of the (a) and (b) images centered on lymphocytes and lymphoblasts respectively.

of candidate lymphoblasts present in the ALL-IDB1 is equal to 510. Only the lymphoblasts that are completely described in the image are 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 is 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 for healthy individuals, and all images labeled with Y=1 are from ALL patients. Each image file ImXXX_Y.jpg is associated with a text file ImXXX_Y.xyc reporting the coordinates of the centroids of the blast cells, if any.

For each image, a corresponding classification text file 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.

The ALL-IDB1 images of blood suffer from a typical non-uniform background illuminations (visible in Fig. 5). Even

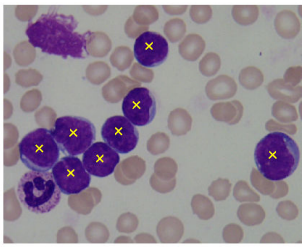

Fig. 4. Examples of a classified blood image portion. Each cross represents the centroid of a lymphoblast stored in the corresponding classification file.

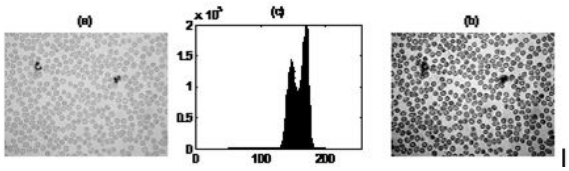

Fig. 5. Images of the ALL IDB1 are typically bimodal and a non uniform background can be present: (a) an example image; (b) the corresponding histogram; (c) histogram stretching of (a) showing the native microscope vignetting effect.

if the images still remain intelligible, segmentation methods based on basic thresholding can heavily suffer of this condition.

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, except the image dimensions. 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_0.jpg" if the central cell is a probable blast, and "ImXYZ_1.jpg" in the other cases.

4. ACCURACY OF ALGORITHMS ON ALL-IDB

A system capable to identify the presence of blast cells in the input image can work with different structures of modules, for example, it can processes the following steps: (i) the identification of white cells in the image, (ii) the selection of Lymphocytes, (iii) the classification of tumor cell. 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

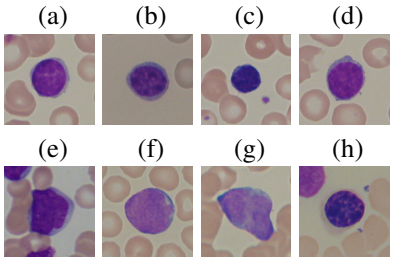

Fig. 6. Examples of the images contained in ALL-IDB2: healthy cells from non-ALL patients (a-d), probabile lymphoblasts from ALL patients (e-h).

a benchmark approach partitioned in three different tests, as follows:

- **Cell test** - the benchmark account for the classification of single cells is blast or not (the test is positive if the considered cell is blast cell or not);
- **Image level** - 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 of each single test where the term elements refers to the cells/images of the corresponding level:

- **True positives (TP)** - the number of elements correctly classified as positive by the test;
- **True negatives (TN)** - the number of elements correctly classified as negative by the test;
- **False positive (FP)** - also known as type I error, is the number of elements classified as positive by the test, but they are not;
- **True positive (FN)** - also known as type II error, is 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* (the probability of correctly classifying elements with ALL equals to $TP/\left(TP+FN\right)$, *Specificity* (the probability of correctly classifying elements without ALL computable as $TN/\left(TN+FP\right)$ and the *Classification error* (where the total error in an analysis layer is defined by $CE=FP+FN$. 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 (cell and image).

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.

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

Figures of Merit	**Classified Element**	
	Cells	Images
TP%		
TN%		
FP%		
FN%		
Misclassification %		
Specificity %		
Sensitivity %		

5. CONCLUSIONS

In this paper, we have proposed a public dataset of blood samples, specifically designed for the evaluation and comparison of the performances algorithms of segmentation and image classification. We have also examined the actual state of art of the actual automatic systems for the detection of ALL and proposed a metric for evaluate the performance 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-DB 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 approach based on a common dataset and figures of merits.

6. REFERENCES

[1] D.J. Foran, D. Comaniciu, P. Meer, and L.A. Goodell, "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, December 2000.

[2] K. S. Kim, P. K. Kim, J. J. Song, and Y. C. Park, "Analyzing blood cell image to distinguish its abnormalities (poster session)," in *Proceedings of the eighth ACM international conference on Multimedia*, 2000, pp. 395–397.

[3] A. Y. Grigoriev and H-S. Ahn, "Robust recognition of white blood cell images," in *Proceedings of the International Conference on Pattern Recognition (ICPR '96)*, 1996, pp. 371–375.

[4] H. Nor Hazlyna, M.Y. Mashor, N.R. Mokhtar, A.N. Aimi Salihah, R. Hassan, R.A.A. Raof, and M.K. Osman, "Comparison of acute leukemia image segmentation using hsi and rgb color space," in *10th International Conference on Information Sciences Signal Pro-*

cessing and their Applications (ISSPA 2010), May 2010, pp. 749–752.

[5] F. Sadeghian, Z. Seman, A. Ramli, B. Abdul Kahar, and M-Iqbal Saripan, "A framework for white blood cell segmentation in microscopic blood images using digital image processing," *Biological Procedures Online*, vol. 11, no. 1, pp. 196–206, 2009.

[6] B. Nilsson and A. Heyden, "Model-based segmentation of leukocytes clusters," in *16th International Conference on Pattern Recognition*, 2002, vol. 1, pp. 727–730.

[7] B. Prasad, J. S. Iris Choi, and W. M. Badawy, "A high throughput screening algorithm for leukemia cells," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE 2006)*, 2006, pp. 2094–2097.

[8] P. Bamford and B. Lovell, "Method for accurate unsupervised cell nucleus segmentation," in *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2001, vol. 3, pp. 2704–2708.

[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," in *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference*, 2010.

[10] V. Piuri and F. Scotti, "Morphological classification of blood leucocytes by microscope images," in *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA)*, july 2004, pp. 103–108.

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

[12] F. Scotti, "Robust segmentation and measurements techniques of white cells in blood microscope images," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference (IMTC)*, April 2006, pp. 43 –48.

[13] G. Ji, Z. Yang, and W. You, "Pls-based gene selection and identification of tumor-specific genes," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. PP, no. 99, pp. 1–12, 2010.

[14] B. Prasad and W. Badawy, "High throughput algorithm for leukemia cell population statistics on a hemocytometer," in *IEEE Biomedical Circuits and Systems Conference (BIOCAS)*, November 2007, pp. 142–145.

[15] M.R. Asadi, A. Vahedi, and H. Amindavar, "Leukemia cell recognition with zernike moments of holographic images," in *Proceedings of the 7th Nordic Signal Processing Symposium (NORSIG 2006)*, June 2006, pp. 214 –217.

[16] K. B. Taylor and J. B. Schorr, "Blood vol.4," 1978.

[17] J. M. Bennett, D. Catovsky, Marie-Theregse Daniel, G. Flandrin, D. A. G. Galton, H. R. Gralnick, and C. Sultan, "Proposals for the classification of the acute leukaemias french-american-british (fab) co-operative group," *British Journal of Haematology*, vol. 33, no. 4, pp. 451–458, 1976.

[18] A. Biondi, G. Cimino, R. Pieters, and Ching-Hon Pui, "Biological and therapeutic aspects of infant leukemia," *Blood*, vol. 96, no. 1, pp. 24–33, July 2000.

[19] R. Donida Labati, V. Piuri, and F. Scotti, "All-idb web site," University of Milan, Departement of Information Technologies, http://www.dti.unimi.it/fscotti/all.