A Multiple Classifier Learning by Sampling System for White Blood Cells Segmentation

Cecilia Di Ruberto, Andrea Loddo, Lorenzo Putzu

Department of Mathematics and Computer Science, University of Cagliari,
via Ospedale 72, 09124 Cagliari, Italy
dirubert@unica.it, lorenzo.putzu@unica.it

Abstract. The visual analysis and the counting of white blood cells in microscopic peripheral blood smears is a very important procedure in the medical field. It can provide useful information concerning the health of the patients, e.g., the diagnosis of Acute Lymphatic Leukaemia or other important diseases. Blood experts in clinical centres traditionally use these methods in order to perform a manual analysis. The main issues of the traditional human analysis are certainly related to the difficulties encountered during this type of procedure: generally, the process is not rapid and it is strongly influenced by the operator's capabilities and tiredness. The main purpose of this work is to realize a reliable automated multiple classifier system based on Nearest Neighbour and Support Vector Machine in order to manage all the regions of immediate interests inside a blood smear: white blood cells nucleus and cytoplasm, erythrocytes and background. The experimental results demonstrate that the proposed method is very accurate and robust being able to reach an accuracy in segmentation of 99%, indicating the possibility to tune this approach to each couple of microscope and camera.

Keywords: Automatic detection, Biomedical image processing, Segmentation, Machine Learning, White blood cell analysis.

1 Introduction

The main purpose of this work was to develop an automatic system able to extract appropriate information from blood cell images taken by microscopes in order to easily perform a useful activity on them, e.g., white blood cells count. Ideally, there are several useful operations for what this method can be used for: essentially, identify, analyse, classify or count the white blood cells held in one or more microscopic images. Nevertheless we can say that the most important and key step of the entire automatic process is, certainly, the image segmentation, which differentiates meaningful objects from the background. It is a crucial step because its accuracy greatly affects both the computational speed and the overall accuracy of the whole system. However, it is also a very difficult problem to manage because of the complex nature of the cells, low resolution of microscopic images and complex scenes, e.g. cells can overlap each other or cells can have

different sizes or shapes. On the other hand, the colour and contrast between the cells and the background can vary so often according to the frequent, inconsistent staining technique, thickness of smear and illumination. Although standardization is useful to avoid superfluous differences in the features of similar images, a robust segmentation approach can cope with the described issues. One natural way for colour image segmentation is to perform pixels clustering or classification in colour space. Unsupervised and supervised schemes [1], such as k-means, neural network et al., have been widely used for this purpose even if there are many disadvantages to deal with. Generally, the biggest problem of an unsupervised clustering scheme is how to determine the number of clusters, which is known as cluster validity. And as for a colour image, the selection of colour space is quite critical. The supervised scheme needs training. The training set and initialization may affect the results, and overtraining should be avoided. So a supervised clustering/classification algorithm with good generalization property is most appealing. Our method aims to solve the segmentation problem in a non-linear feature space obtained by kernel methods in order to overcome the non-linearity of data distribution and the shift/offset of colour representing the different regions of interest inside a blood sample: mature erythrocytes, nuclei and cytoplasm of white blood cells. We wanted to develop an automatic machine learning to perform image segmentation of blood and bone marrow cells images. SVM (Support Vector Machines) and ANN (Artificial Neural Network) are machine learning models with excellent performances in classification, but their main drawbacks are that a training phase is absolutely necessary to make them work and the training phase could be computationally hard with large datasets.Our solution has been developed following the suggestions of [2–4]. We used the ALL-IDB dataset [5], a public and free dataset that contains microscopic images of blood samples, specifically designed for the evaluation and the comparison of algorithms for segmentation and image classification. Our idea is to use a part of this dataset as a training set for our learning by sampling algorithm. The first step of the algorithm is to apply a classic segmentation method to obtain pure samples related to the regions of white blood cells nucleus and cytoplasm, mature erythrocytes and background. The pixels obtained from this region have been reduced in number through a Nearest Neighbour Search (NNS) by removing any duplicates or elements with distance next to zero. Then we prepared the training samples by adapting sampling from the regions obtained from the classic segmentation phase so as to perform the training process of a multiclass SVM in order to correctly classify all the pixels of a given image. Finally, the SVM is used to segment the image for extracting whole white cells, using a classification phase by means of a classification model. Since the size of training set could be controlled and reduced in sampling, SVM training is really fast. Section 2 introduces some background concepts about peripheral blood analysis and the dataset for our purposes. Section 3 illustrates a brief summary about the classic segmentation methods, known in literature. Section 4 presents SVM basis theory and how it can be used for our purposes. Section 5 shows the proposed solution and some experimental results. Section 6 describes how to tune the pro-

posed approach to each dataset. Discussions, conclusions and future aspects are given in Section 7.

2 Background

A typical blood image usually consists of three components: red blood cells (RBCs) or erythrocytes, leukocytes, and platelets. Leukocytes are easily identifiable, as their nucleus appears darker than the background. However, the analysis and the processing of data related to the WBCs are complicated due to wide variations in cell shape, dimensions and edges. The generic term leukocyte refers to a set of cells that are quite different from each other (Fig. 1). Leukocyte cells containing granules are called granulocytes, and they include neutrophils, basophils and eosinophils. Cells without granules are called mononuclear, and they include lymphocytes and monocytes. Furthermore, lymphocytes suffering from ALL, called lymphoblasts, have additional morphological changes that increment with increasing severity of the disease. In particular, lymphocytes are regularly shaped and have a compact nucleus with regular and continuous edges, whereas lymphoblasts are irregularly shaped and contain small cavities in the cytoplasm, termed vacuoles, and spherical particles within the nucleus, termed nucleoli [5]. In Fig. 1 some images examples of healthy and sick WBCs. As we previously said our idea is to use a part of ALL-IDB dataset as a training set for our learning by sampling algorithm, while the other part of this dataset will be used to test the method. The ALL-IDB is a public image dataset of peripheral blood samples from normal individuals and leukaemia patients. These samples were collected by the experts at the M. Tettamanti Research Centre for childhood leukaemia and haematological diseases, Monza, Italy. The ALL-IDB database has two distinct version (ALL-IDB1 and ALL-IDB2). The ALL-IDB1 can be used both for testing segmentation capability of algorithms, as well as the classification systems and image pre processing methods. This dataset is composed

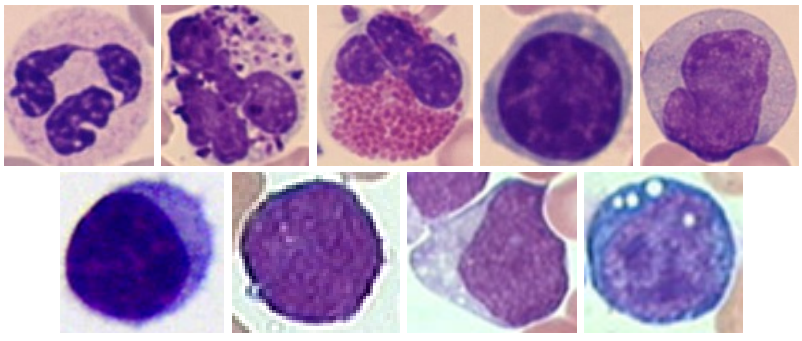

Fig. 1. (Top) The different types of WBCs: neutrophils, basophils, eosinophils, lymphocytes and monocytes. (Bottom) A comparison between a healthy lymphocyte (left) and lymphocytes suffering from ALL of class L1, L2 and L3, respectively [6].

of 108 images captured with an optical laboratory microscope coupled with an Olympus Optical C2500L camera or a Canon PowerShot G5 camera. All images are in JPG format with 24-bit colour depth. The first 33 have 1712×1368 resolution, the remaining have 2592×1944 resolution. The images are taken with different magnifications of the microscope ranging from 300 to 500 which brings the colour differences that we managed grouping the images with same brightness characteristics together. The ALL-IDB2 is a collection of cropped area of interest of normal and blast cells that have been extracted from the ALL-IDB1 dataset. It contains 260 images and the 50% of these represent lymphoblasts. Some images example belonging to the ALL-IDB are showed in Fig. 2

3 Related works

According to the literature, few examples of automated systems are able to analyse and classify WBCs from microscopic images, and the existing systems are only partially automated. In particular, a considerable amount of work has been performed to achieve leukocytes segmentation. For example, Madhloom [7] developed an automated system to localise and segment WBC nuclei based on image arithmetical operations and threshold operations. Sinha [8] and Kovalev [9] attempted to differentiate the five types of leukocytes in cell images. Sinha used k-means clustering on the HSV colour space for WBCs segmentation and different classification models for cell differentiation. Kovalev first identified the nuclei and then detected the entire membrane by region growing techniques. Few papers sought to achieve robust segmentation performance under uneven lighting conditions. However, a study by Scotti [10], used a low-pass filter to remove background, different threshold operations and image clustering to segment WBCs. Moreover, other authors proposed methods for automated disease classification. In particular, Piuri [11] proposed an approach based on edge de-

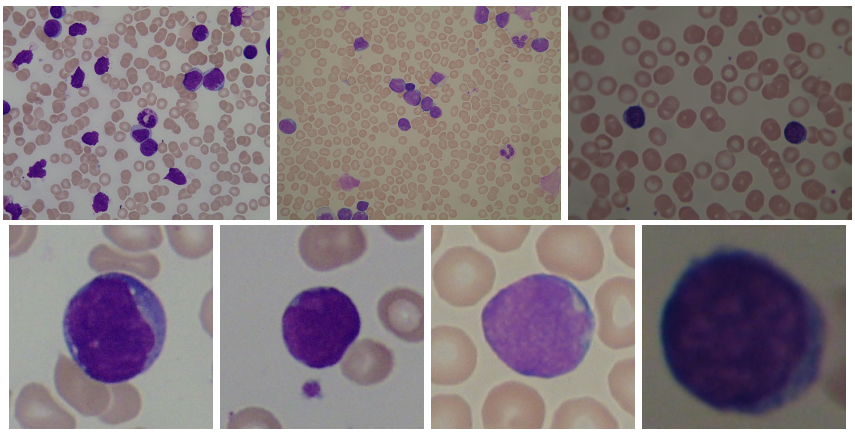

Fig. 2. Original images from the ALL-IDB1 and original images from the ALL-IDB2

tection for WBC segmentation, and they used morphological features to train a neural network to recognise lymphoblasts. Halim [12] proposed an automated blast counting method to detect acute leukaemia in blood microscopic images that identifies WBCs through a thresholding operation performed on the S component of the HSV colour space, followed by morphological erosion for image segmentation. Although the results of this study seem very encouraging, there is no method to determine the optimum threshold for segmentation, and no feature or classifiers were presented. Mohapatra [13] investigated the use of an ensemble classifier system for the early diagnosis of ALL in blood microscopic images. The identification and segmentation of WBCs is realised through image clustering followed by the extraction of different types of features, such as shape, contour, fractal, texture, colour and Fourier descriptors, from the sub-image. Finally an ensemble of classifiers is trained to recognise ALL. The results of this method were good, but they were obtained by using a proprietary dataset, so the reproducibility of the experiment and comparisons with other methods are not possible.

4 SVM for segmentation

Remembering that our starting objective was to segment blood cells images, now we explain how this classification method can be used to reach our segmentation purposes and targets. SVM has been chosen in order to perform a classification of every single pixel belonging to the images we have to segment, following the method proposed in [1]. Once the manually segmented images have been obtained on a certain training set, we used a set of these produced images to train the different SVM we realized. The **first strategy** works as a normal binary SVM classifier, hence we have exactly two classes in which the pixels will be classified: the positive class groups together the white blood cell nuclei and cytoplasm pixels, instead the negative class represents pixels belonging to erythrocytes or background. Fig. 3 shows the segmentation result using this solution, in which the WBC is exactly recognized and segmented, but the lighter region of erythrocytes are misclassified as WBC region. The **second strategy** substantially works like the first one, with the main difference that we exclude from the training samples all the pixels belonging to the cytoplasm, in order to avoid misclassification due to similarities with the lighter region of erythrocytes. Fig. 3 shows the segmentation result using this solution. Again the nucleus is well detected but for determined classes of WBCs the cytoplasm is not well detected. The **third strategy** is based on the results obtained with the two previous version. In fact the classifier needs more valid training samples for cytoplasm only. So, in this version we performed a three-class SVM, using both the pixels belonging from WBCs nuclei (class 1) and both pixels belonging to the WBCs cytoplasm (class 2). Thus, pixels belonging to erythrocytes or background are labelled with class 3. Fig. 3 shows the segmentation result using this solution in which both nucleus and cytoplasm is well detected.

5 System implementation

For each strategy present, the training set is formed by sampling pixels from the images belonging to the ALL-IDB2 presenting healthy WBCs chosen to make part of the available training images. On the other hand, the test set is formed of the first 33 images of ALL-IDB1, acquired in the same lighting conditions and with the same camera. In order to provide to the SVM the most accurate pixels related to white blood cell nuclei and cytoplasm all the images belonging to the ALL-IDB2 have been manually segmented by skilled operators, creating two ground-truth images for each sample. The first one contains the white blood cells segmented in their entirety while the second one contains only the white blood cells nuclei. From these images, the segmented cytoplasm region could be easily obtained performing a difference operation between the first image and the second one and remembering that the cytoplasm region is always placed around the white blood cell nucleus. The obtained cytoplasm and nucleus images are used as masks to extract the pixels from the original images.

5.1 Classifiers setting

Now our interest became to perform a properly training phase over the given pixels just obtained, as we already proposed in [17]. But, since the chosen pixels must be the most various possible all over the regions, this time we used the kNN with Euclidean distance in order to provide to the SVM a smaller but more effective set of pixels. In this way the SVM should realize a more robust classification model during the training phase. In particular we performed a NNS all over the pixels belonging from the same region, in order to remove duplicates (pixels with distance = 0) or too close values (pixels with distance close to 0). Also pixels that presents a distance higher than the others have been removed, in fact they can be considered as outliers. The distances computed with the NNS

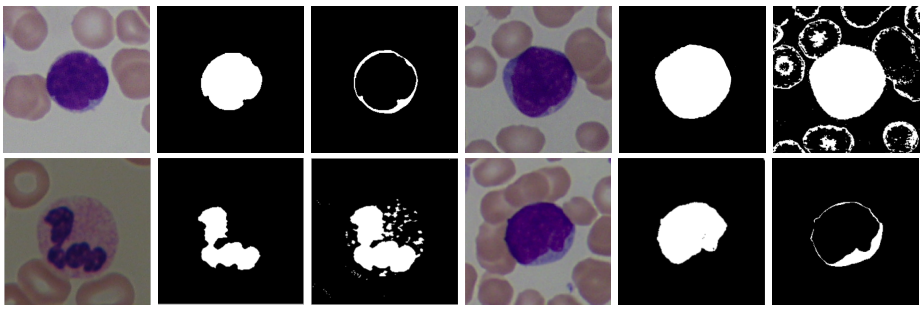

Fig. 3. (Top) From left to right: training original image from ALL-IDB2, manually segmented nucleus and cytoplasm; test original image, segmentation result for nucleus an cytoplasm with the first strategy. (Bottom) From left to right: test original image, segmentation result for nucleus an cytoplasm with the second strategy; test original image, segmentation result for nucleus an cytoplasm with the third strategy

have also been used to select pixels that compose our training set. In fact not all the pixels will be used, but only a small portion that permits to obtain a fast but accurate segmentation. Pixel selection uses distances so as to consider all the possible variations in colour inside a region. So, a uniform sampling have been made. Once obtained our training set we performed two main experiments. The first one have been realised to verify our implementation performances over single WBCs and in order to identify the most suitable parameters for the SVM. Thus, through a 10 fold cross-validation each time we divided the original training set in two subsets, the first was used to train the SVM and the second one was used to test the obtained model. The parameters that permitted to obtain an ideal average accuracy value was c parameter equal to $1e3$ and γ equal to $1e1$. The second and final experiment have been realised to verify the segmentation performances of the proposed method. Thus the whole original training set was used to create the SVM model. The first 33 native resolution images belonging to ALL-IDB1 were used as test set and to check the method applied to a natural image composed of several white blood cells of many different classes.

5.2 Experimentation

Once the first (visual) results have been obtained we started experimenting with various features that can be used to train the classifier. In fact, even though we are talking of a segmentation technique, pixels are used as features for the SVM classifier. Until now the only one descriptors that we used a the colour values. Although in many cases these features are enough to reach a good segmentation result, in other cases a poor feature set like this is not able to discriminate pixels belonging to regions with wide variations in colours. Thus the first intuition was to add the average colour values of each pixel neighbourhood. These average values have been tested for neighbourhood of size 3×3, 5×5 and 7×7. For the same neighbourhood we have also computed other statistical features that are often used for segmentation purposes: standard deviation, uniformity and entropy. While the segmentation accuracy highly benefits from the use of these new features, the overall system became to slow, both in training and segmentation phase. Furthermore, the step of samples selection, used to train the classifier, became too complex, due to a higher number of samples with different values. For all these reasons the features previously mentioned have been extracted only for neighbourhood of size 3×3, showing excellent performances as showed in fig. 4, outperforming previous results. After the segmentation, all the images have been automatically cleaned, as we have already proposed in [18, 19], in order to remove small artefacts from the background and to give to the reader an idea about the goodness of the results. In order to evaluate the segmentation performances of the proposed method, also a subset of images (10 random samples) belonging to the ALL-IDB1 have been manually segmented by skilled operators, creating two ground-truth images for each sample. These images display respectively the WBCs and the WBCs nuclei present in the image. Fig. 4 shows an image belonging to the ALL-IDB1, its relative ground-truth images and our segmentation result. Using the manually segmented images we have

also computed the segmentation accuracy of the proposed method comparing pixel-wise our segmentation results that often reaches the 99%.

6 System extension

Since, getting manually segmented images is not so simple and cheap, we propose an extension of our system to be applied to each dataset of peripheral blood images, acquired in each illumination condition and with different combinations of cameras and microscopes. Our propose is based on ROI (Region of Interest) selection. Thus, making use of few original images the object of interest (WBCs) could be selected and used as positive example for our multiple classifier. Considering that we are talking of a segmentation method based on classifiers also negative instances are needed, so the background region, that comprises red blood cells and plasma, must be selected. An example of ROI selection for positive and negative example is showed in Fig. 5.

Obviously for the negative example the selected regions mustn't present WBCs. In fact in this case the NNS is performed also over pixels belonging from different region, in order to avoid errors committed during the ROI selection and in order to remove pixels with close values. In this way the obtained training set should present again uniform pixel values. Note that with this approach the WBC cytoplasm and nucleus are managed as a unique region, both because the ROI selection is not so suitable for adjacent region and both because they can be easily separated in a further step by using a simple threshold. Differently with this approach we are able to take into account also RBCs, considering them as a

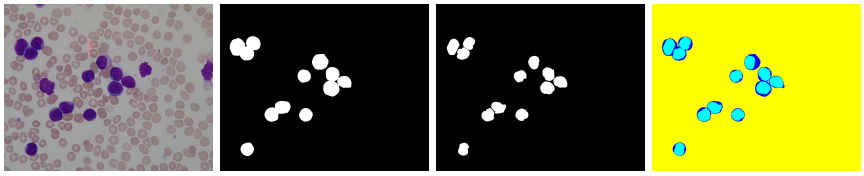

Fig. 4. Original images from the ALL-IDB1 database, ground-truth for whole leukocyte, ground-truth for leukocyte nuclei and final segmentation result.

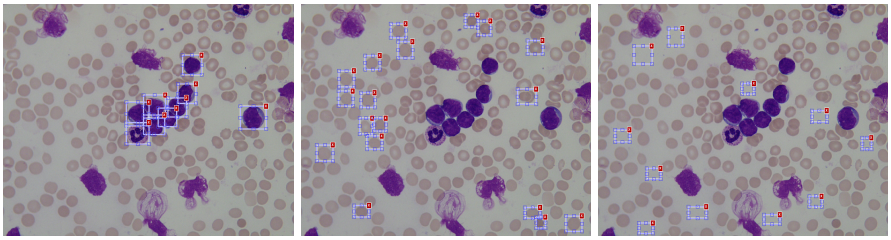

Fig. 5. Examples of ROI selection for WBCs, RBCs and plasma.

different class. Thus it is possible to perform a binary segmentation or a multiple segmentation as showed in Fig. 6.

As it can be seen also in this case the segmentation is really accurate, being able to properly segment WBCs and also RBCs. Using the manually segmented images we have computed the segmentation accuracy of this version that again reaches the 99%.

7 Conclusions

This work proposed and investigated a new automated white blood cell recognition method that can be applied to support some existing medical methods, like the WBCC, White Blood Cells Counting. It is realized using lots of notions already known in literature but combining them to build an essentially brand new method in which the major innovation is brought by the use of a multiple classifier approach that makes use of the Nearest Neighbour and Support Vector Machine. Whereas the aim of a fully automated cell analysis and diagnosis of with blood cell has not yet reached, many important steps in the image segmentation using Learning by Sampling method have been realized. We proposed a segmentation approach using several variations in the schemes. The experimental results demonstrate that the new approach is very accurate and robust in relation to some traditional methods. The performances achieved with the ALL-IDB often reaches the 99% of accuracy, but in particular we want to highlight the possibility to tune this approach to each couple of microscope and camera using only few image samples. Despite the good results, we do not consider the development of our project totally concluded. Our purposes and hopes are certainly to continue the work in order to experiment several new investigations that could potentially bring to better results. Among the future works we can indicate the extension to different colour spaces in which segmentation process could be easily and more effective. A further step will include analysis and recognition of the different types of healthy and blasted white blood cells. Finally, our idea is to export the whole procedure to bone marrow images, in which the first segmentation phase is, usually, more difficult than in the peripheral blood images, since the brightness conditions could be very different and large clusters of cells can exist.

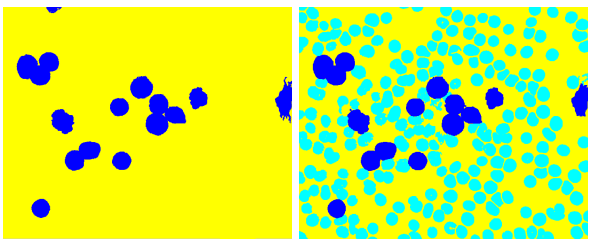

Fig. 6. Segmentation results after ROI selection for two and three classes.

Acknowledgments

This work has been funded by Regione Autonoma della Sardegna (R.A.S.) Project CRP-17615 DENIS: Dataspace Enhancing Next Internet in Sardinia. Lorenzo Putzu gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.).

References

1. P. Chen, L. Huijuan, C. Feilong, Segmentation of blood and bone marrow cell images via learning by sampling. In Emerging Intelligent Computing Technology and Applications, pp. 336-345, 2009.
2. K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory, vol. 21, no. 1, pp. 32-40, 1975.
3. L. G. Shapiro, G. C. Stockman, Computer Vision, chap. 12, pp. 279-325, New Jersey, Prentice Hall, 2001.
4. Gonzalez, R. C., Woods, R. E., Digital Image Processing, Prentice Hall Pearson Education, Inc.. New Jersey, USA, ed. 3, 2008
5. Donida Labati, R., Piuri, V., Scotti, F.: ALL-IDB: the Acute Lymphoblastic Leukemia Image DataBase for Image Processing. In Proceedings of the 18th IEEE ICIP International Conference on Image Processing. Editors: Benot Macq, Peter Schelkens. IEEE Publisher, pp. 2045-2048, Brussels, Belgium, September 11-14, 2011.
6. Bennett, J.M., Catovsky, D., Daniel, M.T., Flandrin, G., Galton, D.A., Gralnick, H.R., Sultan, C.: Proposals for the classification of the acute leukemias. French-American-British (FAB) co-operative group. British Journal of Hematology, vol. 33, no. 4, pp. 451-458, August 1976
7. Madhloom, H. T., Kareem, S. A., Ariffin, H., Zaidan, A. A., Alanazi, H. O., Zaidan, B. B.: An Automated White Blood Cell Nucleus Localization and Segmentation using Image Arithmetic and Automated Threshold. Journal of Applied Sciences, vol. 10, no. 11, pp. 959-966, 2010
8. Sinha, N., Ramakrishnan, A. G.: Automation of Differential Blood Count. In Proceedings of the Conference on Convergent Technologies for the Asia-Pacific Region, Editors: A. Chockalingam. IEEE Publisher, vol. 2, pp. 547-551, Taj Residency, Bangalore, October 15-17, 2003
9. Kovalev, V. A., Grigoriev, A. Y., Ahn, H.: Robust Recognition of White Blood Cell Images. In Proceedings of the 13th International Conference on Pattern Recognition. Editors: M.E. Kavanaugh and B. Werner. IEEE Publisher, pp. 371-375, Vienna, Austria, August 25-29, 1996
10. Scotti, F.: Robust Segmentation and Measurements Techniques of White Cells in Blood Microscope Images. In Proceedings of the IEEE Instrumentation and Measurement Technology Conference. Editors: P. Daponte and T. Linnenbrink. IEEE Publisher, pp. 43-48, Sorrento, Italy, 24-27 April, 2006
11. Piuri, V., Scotti, F.: Morphological Classification of Blood Leucocytes by Microscope Images. In Proceedings of the IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, IEEE Publisher, pp. 103-108, Boston, MA, USA, 14-16 July, 2004

12. Halim, N. H. A., Mashor, M. Y., Hassan, R.: Automatic Blasts Counting for Acute Leukemia Based on Blood Samples. International Journal of Research and Reviews in Computer Science, vol. 2, no. 4, August, 2011
13. Mohapatra, S., Patra, D., Satpathy, S.: An Ensemble Classifier System for Early Diagnosis of Acute Lymphoblastic Leukemia in Blood Microscopic Images. Journal of Neural Computing and Applications, Article in Press, 2013
14. J. F. David, D. Comaniciu, P. Meer, Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy. IEEE Transaction on Information Technology in Biomedicine, vol. 4, no. 4, pp. 12-22, 2000.
15. O. Lezoray, A. Elmoataz, H. Cardot, G. Gougeon, M. Lecluse, H. Elie, Hubert M. Revenu, Segmentation of Color Images from Serous Cytology for Automated Cell Classification. In journal of Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology, vol. 22, no. 4, pp. 311-322, 2000.
16. V. N. Vapnik, V. Vapnik, Statistical learning theory, vol. 1, Wiley New York, 1998.
17. C. Di Ruberto, A. Loddo, L. Putzu, Learning by Sampling for White Blood Cells Segmentation, In LNCS of the International Conference on Image Analysis and Processing (ICIAP), 2015 (In Press).
18. L. Putzu, C. Di Ruberto, Investigation of Different Classification Models to Determine the Presence of Leukemia in Peripheral Blood Image. In LNCS of the International Conference on Image Analysis and Processing (ICIAP), vol. 8156, pp. 612-621, 2013.
19. L. Putzu, G. Caocci, C. Di Ruberto, Leucocyte Classification for Leukaemia Detection using Image Processing Technique. Artificial Intelligence in Medicine, vol. 62 no. 3, pp. 179-191, 2014.