Leucocyte classification for leukaemia detection using image processing techniques

Lorenzo Putzu[a], Giovanni Caocci[b], Cecilia Di Ruberto[a]

[a]*Department of Mathematics and Computer Science, University of Cagliari, via Ospedale 72, 09124 Cagliari, Italy.*
[b]*Hematology, Department of Medical Sciences, University of Cagliari, via Is Guadazzonis 2, 09126 Cagliari, Italy.*

Abstract

Introduction: The counting and classification of blood cells allows for the evaluation and diagnosis of a vast number of diseases. The analysis of white blood cells (WBCs) allows for the detection of acute lymphoblastic leukaemia (ALL), a blood cancer that can be fatal if left untreated. Currently, the morphological analysis of blood cells is performed manually by skilled operators. However, this method has numerous drawbacks, such as slow analysis, non-standard accuracy, and dependences on the operator's skill. Few examples of automated systems that can analyse and classify blood cells have been reported in the literature, and most of these systems are only partially developed. This paper presents a complete and fully automated method for WBC identification and classification using microscopic images.

Methods: In contrast to other approaches that identify the nuclei first, which are more prominent than other components, the proposed approach isolates the whole leukocyte and then separates the nucleus and cytoplasm. This approach is necessary to analyse each cell component in detail. From each cell component, different features, such as shape, colour and texture, are extracted using a new approach for background pixel removal. This feature set was used to train different classification models in order to determine which one is most suitable for the detection of leukaemia.

Results: Using our method, 245 of 267 total leukocytes were properly identified (92% accuracy) from 33 images taken with the same camera and under the same lighting conditions. Performing this evaluation using different classification models allowed us to establish that the support vector machine with a Gaussian radial basis kernel is the most suitable model for the

identification of ALL, with an accuracy of 93% and a sensitivity of 98%. Furthermore, we evaluated the goodness of our new feature set, which displayed better performance with each evaluated classification model.

*Conclusions:*The proposed method permits the analysis of blood cells automatically via image processing techniques, and it represents a medical tool to avoid the numerous drawbacks associated with manual observation. This process could also be used for counting, as it provides excellent performance and allows for early diagnostic suspicion, which can then be confirmed by a haematologist through specialised techniques.

Keywords: Image processing, Microscopic image segmentation, Cell analysis, White blood cell detection, Leukaemia classification

1. Introduction

The observation of blood cells from microscopic images allows for the evaluation and diagnosis of many diseases. Leukaemia is a blood cancer that can be detected through the analysis of white blood cells (WBCs) or leukocytes. There are two types of leukaemia: acute and chronic. According to the French-American-British (FAB) classification model [1], acute leukaemia is classified into two subtypes: acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). Here, we consider only ALL, which affects a group of leukocytes called lymphocytes. ALL primarily affects children and adults over 50 years of age. The risk of developing ALL is highest in children younger than 5 years of age, and it declines and begins to rise again after age 50. Due to its rapid expansion into the bloodstream and vital organs, ALL can be fatal if left untreated [2]. Therefore, early diagnosis of this disease is crucial for a patients' recovery, especially for children. Diagnosis of ALL is based on the morphological identification of lymphoblasts by microscopy and the immunophenotypic assessment of lineage commitment and developmental stage by flow cytometry [3]. The observation of blood samples by skilled operators is one diagnostic procedure available to initially recognise different diseases. Human visual inspection is tedious, lengthy and repetitive, and it suffers from the presence of a non-standard precision because it depends on the operator's skill; these disadvantage limit its statistical reliability. Various systems for the automatic quantification of blood cells exist on the market that count the numbers of different types of cells within a blood smear. These counters use flow cytometry to measure the physical

characteristics and chemical properties of the blood cells using a light detector that uses fluorescence or electrical impedance to identify cell types. Although the results of quantification are very precise, the instrument does not detect morphological abnormalities of the cells; therefore, a subsequent complementary blood analysis via the microscope is required. The use of image processing techniques can help count the cells in human blood to provide information about cell morphology. These techniques require only one image, making them less expensive. Moreover, they are more scrupulous in providing more accurate standards. The main goal of this work is to analyse microscopic images to provide a fully automatic procedure to support medical activity. This procedure will count and classify WBCs affected by ALL. Thus, the main contribution of this work is the development of a fully automated system for the detection and segmentation of WBCs. After the feature extraction step, the detected WBCs can be recognised as suffering from ALL or not. The various phases of the proposed method are shown in Fig. 1. The method is presented in detail in the following sections, and it is applied to two sample images (with a scale factor of 0.30) and compared with other approaches described in the literature. The paper is organised as follows: after presenting background and related works in Section 2, Section 3 describes the identification of the leukocytes. This step includes the identification and separation of grouped leukocytes and terminates with an image cleaning, which removes all of the abnormal components from the image. The second step selects the nucleus and cytoplasm of each leukocyte (Section 4). The third step deals with the feature extraction (Section 5). The last phase aims to the classify WBCs (Section 6). The database used and the experimental evaluation of our system are presented in Section 7. Section 8 is devoted to conclusions and potential future directions.

2. Background and related works

A typical blood image usually consists of three components: red blood cells (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. 2). Leukocyte cells containing granules are called granulocytes, and they include neutrophils, basophils and eosinophils. Cells without granules

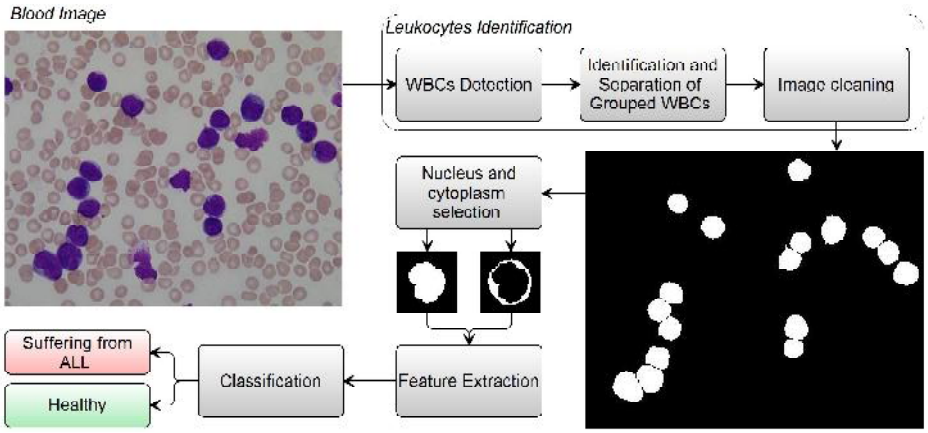

Figure 1: Diagram of the proposed method; from blood image to ALL classification via the identification of WBCs (see text for details).

are called mononuclear, and they include lymphocytes and monocytes. Thus, we can distinguish between these cells according to their shape or size, the presence of granules in the cytoplasm and the number of lobes in the nucleus. The lobes are the most substantial part of the nucleus, and thin filaments connect them to each other. Neutrophils are mainly present in human blood at a percentage ranging between 50 and 70%, and they range in size form 10-12 microns. They are distinguishable due to the number of lobes present in the nucleus, which can range from 1 to 6 according to the cell maturation. Basophils represent only 0-1% of all lymphocytes in human blood, and they have a diameter of approximately 10 microns. Generally, basophils have an irregular, plurilobated nucleus that is obscured by dark granules. Eosinophils are present at 1-5% in human blood, and they are round, 10-12 microns in size, and have a nucleus with two lobes. Eosinophils differ from other WBCs due to the presence of granules, which include para-crystalline structures in the shape of a coffee bean. Lymphocytes are very common in human blood, with a percentage of 20-45% and a size of 7-15 microns. They are characterised as having a rounded nucleus and a poor cytoplasm. Monocytes are the most voluminous WBCs, with a diameter of 12-18 microns, and they represent 3-9% of circulating leukocytes. Their nucleus is large and curved, often in the shape of a kidney. Furthermore, lymphocytes suffering from ALL, called lymphoblasts, have additional morphological changes that increment with increasing severity of the disease. In particular, lymphocytes are

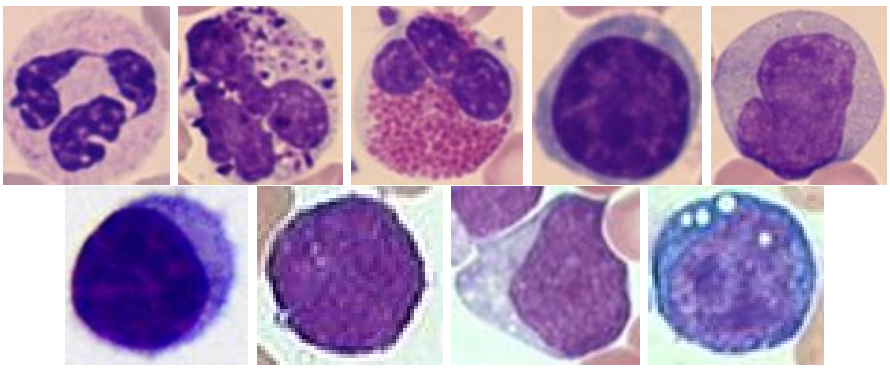

Figure 2: (Top) A comparison between different types of WBCs: neutrophils, basophils, eosinophils, lymphocytes and monocytes. (Bottom) A comparison between lymphocytes suffering from ALL: a healthy lymphocyte, followed by lymphoblasts classified as L1, L2 and L3, respectively, according to the FAB [1].

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 [4] (Fig. 2).

According to the literature, few examples of automated systems that 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 [5] developed an automated system to localise and segment WBC nuclei based on image arithmetical operations and threshold operations. Sinha [6] and Kovalev [7] 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 [8], 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 [9] proposed an approach based on edge detection for WBC segmentation, and they used morphological features to train a neural network to recognise lymphoblasts. Halim [10] 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 [11] investigated the use of an ensemble classifier system for the early diagnosis of ALL in blood microscopic images. The identification and segmentation of WBCs 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.

3. Leukocytes identification

In contrast to other reported methods, our method does not require separate steps of pre-processing and segmentation; it uses pre-processing steps inserted among the various stages of segmentation to make the latter simpler and more robust. Other methods aim first to identify the nuclei, which are more prominent than other components [5], and then to detect the entire membrane (i.e., by region growing [7, 12, 13]). In contrast, our method detects the membrane first, in order to separate the adjacent cells more accurately.

3.1. WBC detection

WBC identification consists of several phases:

- Conversion from RGB to CMYK colour model
- Histogram equalization or contrast stretching operations
- Segmentation by threshold using Zack algorithm
- Background removal operation

WBC identification was made possible by conversion to the CMYK colour model. In fact, leukocytes are more contrasted in the Y component of CMYK colour model because the yellow colour is present in all elements of the image, except leukocytes (Fig. 4 shows two examples). Redistribution of image

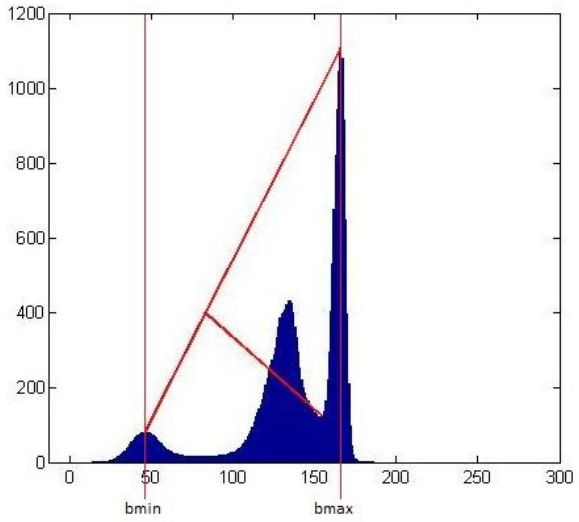

Figure 3: Example Zack algorithm.

grey levels is necessary to make the subsequent segmentation process easier. Then, a histogram equalization or contrast stretching can be used (see Fig. 4). Segmentation is achieved using an automatically calculated threshold. Many threshold techniques have been described in the literature [14, 15]. Here, we use the threshold value based on the triangle method or Zack algorithm [16]. The triangle method is applied to the image histogram (h), resulting in a straight line that connects the highest histogram value ($h[b_{max}]$) and the lowest histogram value ($h[b_{min}]$), where b_{max} and b_{min} indicate the values of the grey levels where the histogram $h[x]$ reaches its maximum and minimum, respectively. Then, the distance (d) between the marked line and the histogram values between b_{min} and b_{max} is calculated. The intensity value where the distance (d) reaches its maximum defines the threshold value (see Fig. 3).

This algorithm is particularly effective when the histogram displays clear valleys between high and weak peaks present in the Y component histograms generated from leukocytes and red blood cells. The threshold results are displayed in Fig. 4. The complement image is then calculated to obtain WBCs on a dark background.

To improve our results, we removed the image background. Some approaches for background extraction have been described in the literature. For example, Scotti [8] used a collection of images to estimate background

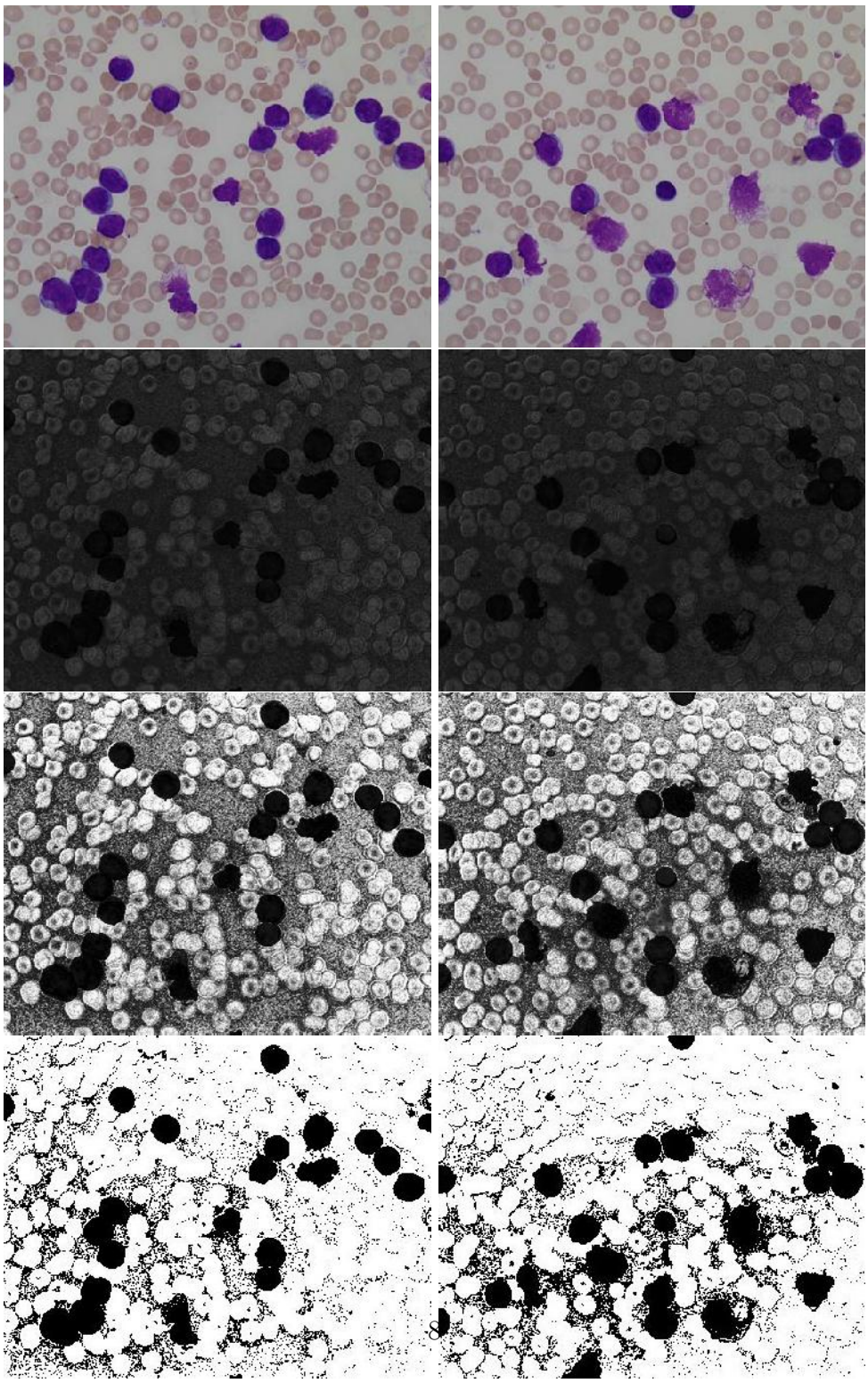

Figure 4: Top to bottom: original blood sample images, Y component images, histogram equalization results and segmentation results.

pixels, while other approaches have a very high computational cost that is unnecessary for this application. The proposed approach involves the use of an automatic threshold to the original grey level image or along the green component of the RGB colour space. The threshold value is calculated again using the triangle method. Fig. 5 shows how the result may not be accurate in all aspects. For example, the centre of red blood cells can be detected as background. However, this does not preclude the achievement of effective background removal because our goal is to preserve only the WBCs present in the image. Background removal can be performed using simple arithmetical operations. Obviously, the background removal process does not produce a clean result for the whole image. To clean up the image, we used area opening [17, 18], which allows us to delete all of the objects with a size smaller than the structuring element. The structuring element used has a circular shape, and its size is calculated based on the average size of the objects in the image (see Fig. 5).

3.2. Identification and separation of grouped leukocytes

An important problem for the analysis of blood images is the presence of adjacent cells and in this case, the presence of leukocyte agglomerates. Only in this phase we can detect and separate leukocyte agglomerates, because in the previous phase we produced an image containing only the WBCs. This process can be summarised in the following basic steps:

- Agglomerate identification through roundness analysis
- Distance transform image calculation
- Watershed segmentation operation
- Watershed line refining

Several methods can be used to verify the presence of adjacent leukocytes [14]. In this work, we used the roundness value because we were able to identify the presence of adjacent cells through the analysis of shape, which is, in the absence of anomalies, almost round. Roundness (1) is a measure of circularity (area-to-perimeter ratio) that excludes local irregularities and can be obtained as the ratio of the area of an object to the area of a circle with the same perimeter of the convex hull of the object:

$$Roundness = \frac{4 * \pi * area}{convex\_perimeter^2} \tag{1}$$

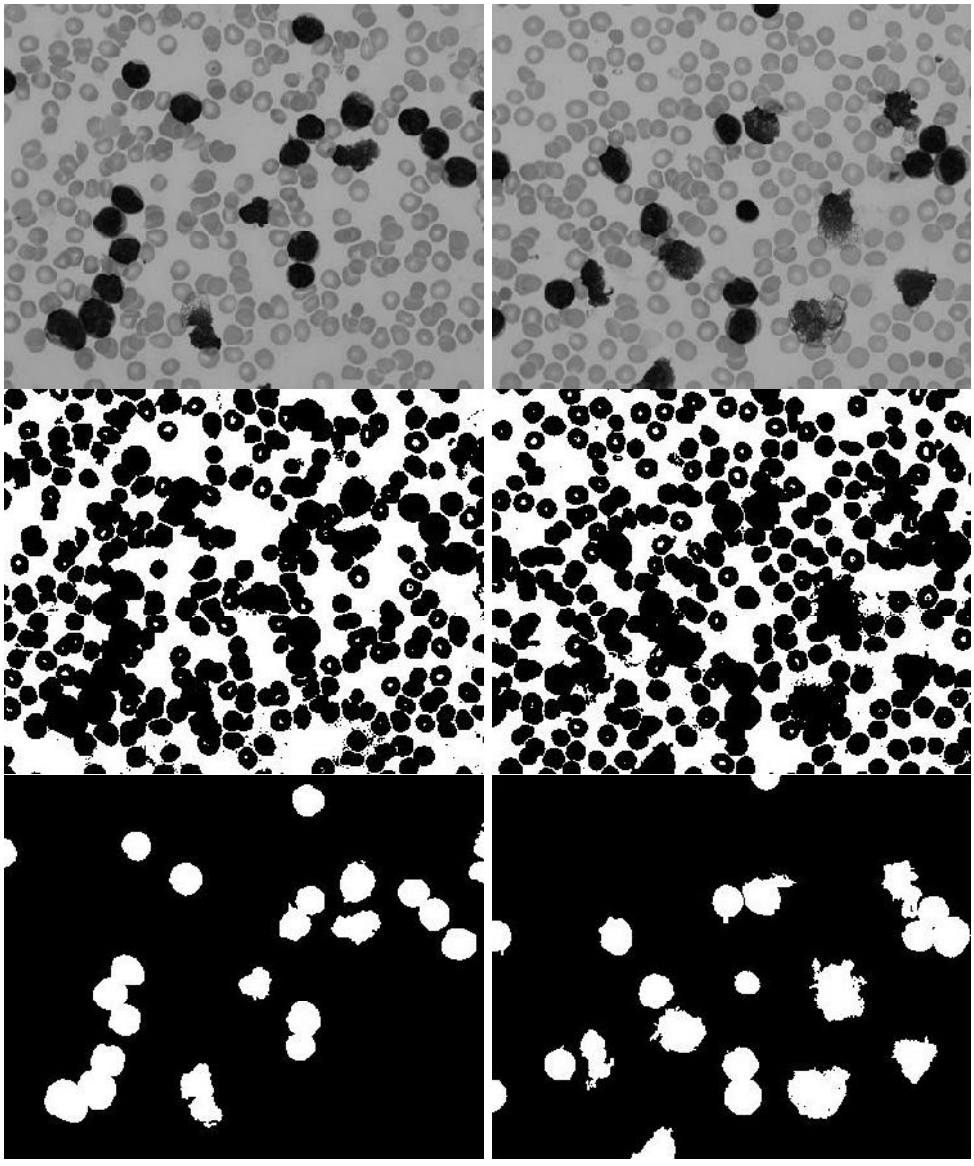

Figure 5: Top to bottom: grey level images, background identification results and background removal results.

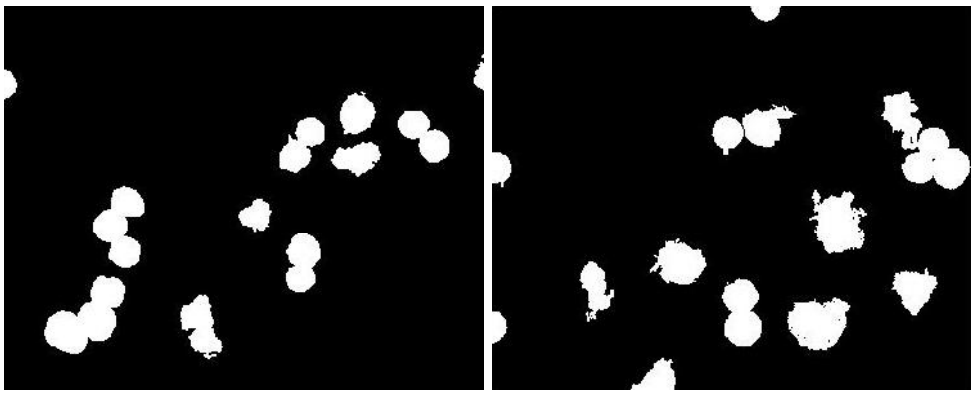

Figure 6: Leukocytes identified as grouped.

Roundness equals 1 for a circular object and is less than 1 for an object that departs from circularity; this measure is relatively insensitive to irregular boundaries. We observed that a roundness value of 0.80 can be used to adequately discriminate single leukocytes from groups of leukocytes, so we adopted this value as the threshold. Connected components with a roundness value greater than the threshold are classified as individual leukocytes, and they proceed directly to the next step of the analysis, while connected components having a roundness value smaller than the threshold are classified as grouped leukocytes and proceed to the separation process. This process creates two different images, and only the image containing grouped leukocytes (Fig. 6) is taken into account. In some cases, this image may be empty, so the phase of separation of the leukocytes will not take place.

Many approaches have been proposed to separate adjacent cells, some of which are included in the process of segmentation, while others are specifically dedicated to separating overlapping cells. For example, the approaches used by Sinha and Ramakrishnan [6] and Kovalev [7] utilise sub-images extracted from the original image by cutting a square around the previously segmented nucleus. Then, assuming that each sub-image has a single WBC, clustering around the nucleus is performed using shape and colour information. The proposed approach is divided into two parts. The first part utilises the method proposed by Lindblad [19], which uses the distance transform [20]. A watershed segmentation [21] is then applied to the distance transform to yield a rough separation between adjacent leukocytes. In this way, the separation tends to be inaccurate because it uses the distance transform

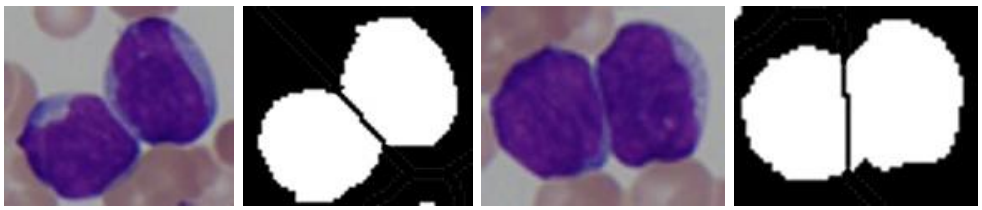

Figure 7: Two original blood sample sub-images and their respective watershed results.

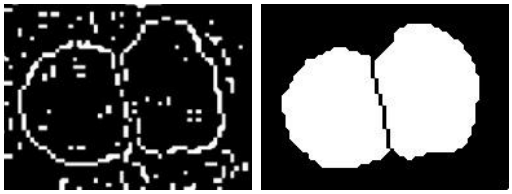

Figure 8: Local maxima image and final separation results.

as a form delimiter. Thus, it only performs well in the presence of adjacent leukocytes with a nearly rounded shape, but it does not perform well in the presence of multiple complex forms, as displayed in the last image of Fig. 7.

Therefore, a second part is needed to refine the contours extracted through the watershed transform. Then, all of the pixels of the component under examination that are located at a distance not greater than a predetermined value from the watershed line are taken into consideration. These pixels are used to derive the deepest concavity for which the line of exact separation must pass. Therefore, by exploiting the information regarding the points of concavity and the points of maximum image in grey tones, it is possible to obtain a cutting line that best fits the contour of the leukocytes. Fig. 8 illustrates the final separation of two adjacent leukocytes, and Fig. 9 shows the final separation results for the whole sample image.

3.3. Image cleaning

Image cleaning requires the removal of all of the leukocytes located on the edge of the image and all abnormal components (non leukocytes), which prevents errors in the later stages of the analysis process. Cleaning the image edge is a simple operation, whereas the removal of abnormal components is a more complex process because it is necessary to determine the number of leukocytes present in the image. First, the size of the area and the size of the convex area are computed for each leukocyte. The size of the area is used

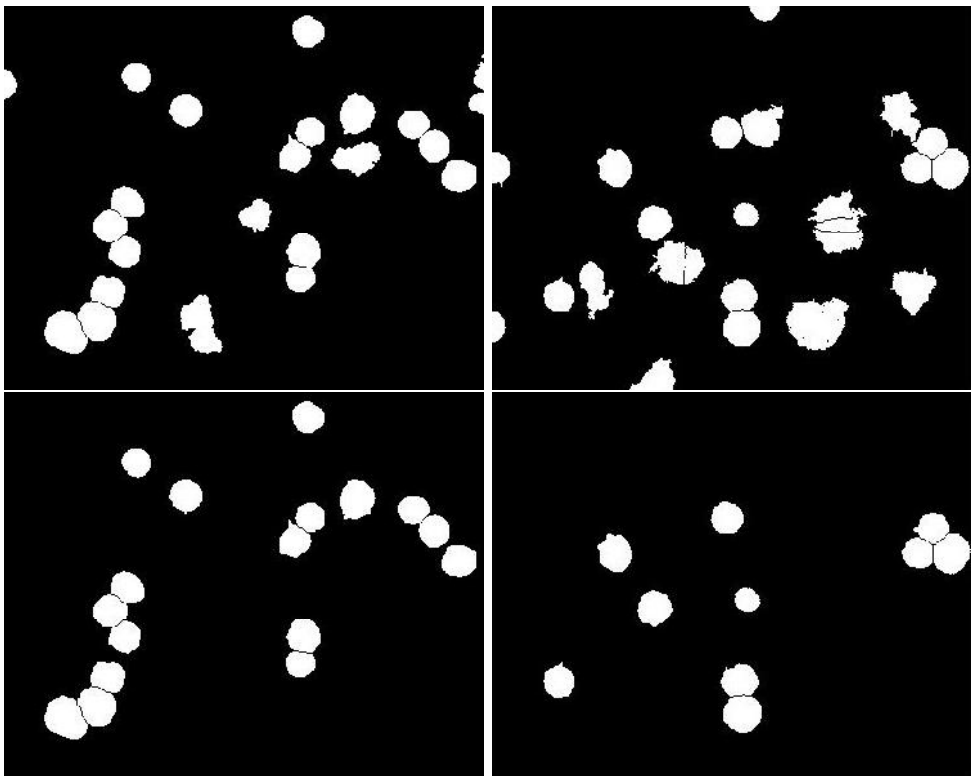

Figure 9: Final separation results and image cleaning results.

to calculate the mean area, which is necessary to determine and eliminate components with irregular dimensions. For example, a very small area might indicate the presence of an artefact that was not removed. Alternatively, a very large area may indicate the presence of adjacent leukocytes that were not adequately. The area and convex area are then used in combination to calculate the solidity value, which we used to discriminate the abnormal component. Solidity measures the density of an object. Solidity is defined as the ratio of the area of an object to the area of a convex hull of the object:

$$Solidity = \frac{area}{convex\_area} \tag{2}$$

A solidity value of 1 signifies a solid object, and a value less than 1 signifies an object with an irregular boundary (or containing holes). The solidity value used for the threshold is calculated directly from the image containing only the individual leukocytes, and when this image is empty, a default value of

0.90 is used. In this case, the default value was identified by experimental results indicating that a solidity value equal to 0.90 adequately discriminates abnormal components; thus, we used this value as the threshold. All objects with a solidity value below the threshold are discarded. In fact, a solidity value below the threshold indicates the presence of artefacts that were not adequately removed. Fig. 9 shows the final results after border cleaning and the removal of abnormal components.

4. Nucleus and cytoplasm selection

Once the leukocytes have been identified, we can proceed to the second segmentation level, which selects for the nucleus and cytoplasm. This step can be simplified by performing an automatic image crop using the bounding box size, which is the smallest rectangle that completely contains a connected component, to isolate a single leukocyte in each sub-image (Fig. 10). Another border cleaning operation is necessary to preserve only the WBC under examination. By definition, the leukocyte nucleus is inside the membrane, making it possible to further simplify this step by cropping the entire portion of the image outside the leukocyte under examination (Fig. 10). This procedure allows for more robust nucleus selection because it completely excludes artefacts of the selection. The nucleus selection approach takes advantage of Cseke's [22] observations, which demonstrated that WBC nuclei are more in contrast on the green component of the RGB colour space. However, in this colour space, the threshold operation described by Otsu [23] does not produce clean results, especially in the presence of granulocytes, because granules are selected erroneously as part of the nucleus. To avoid this issue, the binary image obtained from the green component is combined with the binary image obtained from the a* component of the CIELab colour space via a threshold operation. The mask obtained allows us to clearly extract the leukocyte nucleus. Finally, to obtain the cytoplasm, a subtraction operation is performed between the binary image containing the whole leukocyte and the image containing only the nucleus (Fig. 10).

5. Feature extraction

In this phase, the goal is to transform the images into data and then to extract information reflecting the visual patterns that pathologists refer to, while simultaneously extracting the descriptors that are most relevant to the

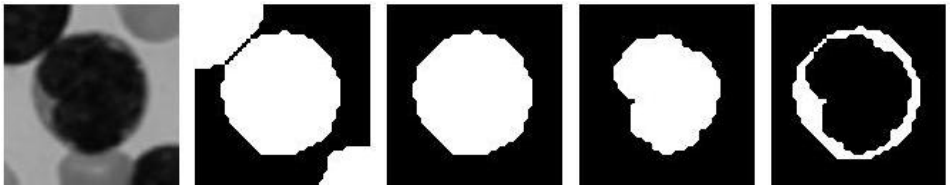

Figure 10: Left to right: grey level sub-image, binary sub-image, whole leukocyte sub-image, nucleus sub-image and cytoplasm sub-image.

subsequent classification process. To this end, we extracted three different types of descriptors from the previously calculated sub-images: shape features, colour features and texture features. Starting from binary sub-images of the nucleus and cytoplasm, we extracted shape descriptors, such as *area*, *perimeter*, *convex area*, *convex perimeter*, *major axis*, *minor axis* and *orientation*. These measures are also used to calculate *elongation*, *eccentricity*, *rectangularity*, *compactness*, *convexity*, *roundness* (1) and *solidity* (2). Elongation (3) measures how an object is elongated. Eccentricity (4) is the ratio of the distance between the foci of the ellipse and its major axis length; this value is between 0 and 1. Rectangularity (5) represents how rectangular a shape is (i.e., how well it fills its minimum bounding box). Compactness (6) is defined as the ratio between the area of an object and the area of a circle with the same perimeter; the maximum value is 1 for a circle. Convexity (7) is the relative amount that an object differs from a convex object, and this value represents the ratio of the perimeter of an object's convex hull to the perimeter of the object itself; the value is 1 for a convex object and less than 1 if the object is not convex, such as an object with an irregular boundary. These measures can be defined as:

$$Elongation = 1 - \frac{minoraxis}{majoraxis} \tag{3}$$

$$Eccentricity = \frac{\sqrt{(majoraxis^2 - minoraxis^2)}}{majoraxis} \tag{4}$$

$$Rectangularity = \frac{area}{majoraxis * minoraxis} \tag{5}$$

$$Compactness = \frac{4 * \Pi * area}{perimeter^2} \tag{6}$$

Figure 11: The binary image of the nucleus and the result of the extraction of the number of lobes obtained through iterative erosion and through ultimate erosion.

$$Convexity = \frac{perimeter_{convex}}{perimeter} \tag{7}$$

where *minoraxis* and *majoraxis* are the width and the height, respectively, of the bounding box (i.e. the smallest rectangle containing the shape). We added two specific measures to these classical measures to analyse leukocytes: the ratio between the area of the cytoplasm and the nucleus and the number of nuclear lobes. To extract the number of nuclear lobes, Scotti [8] proposed an approach using repeated erosions until the correct number of lobes is reached. In a similar manner, our approach makes use of the ultimate erosion of the binary image [17, 18], which consists of the regional maxima of the Euclidean distance transform of the complement of the binary images (Fig. 11). Notably, the number of lobes remains unchanged.

The main disadvantage of shape features is that they are susceptible to errors in segmentation. Thus, these descriptors are used together with regional descriptors, which are less susceptible to errors. Among these are colour descriptors, which are the most discriminatory features of blood cells. The colour descriptors used are mean, standard deviation, smoothness, skewness, kurtosis, uniformity and entropy, which are calculated from sub-images in shades of grey. These descriptors are extracted from images in shades of grey, from which the entire portion of the image outside the leukocyte under examination was cropped. As was observed with previous segmentations, removing the outer portion of the image does not improvement the extracted feature. In fact, even pixels set to zero are considered when calculating feature values that are very different from the real ones. To overcome this problem, we used a new method for calculating the feature that uses the previously calculated binary mask, allowing us to calculate the number of background pixels and discard them from the histogram (Fig. 12) and the calculation of the features themselves. Table 1 demonstrates that feature

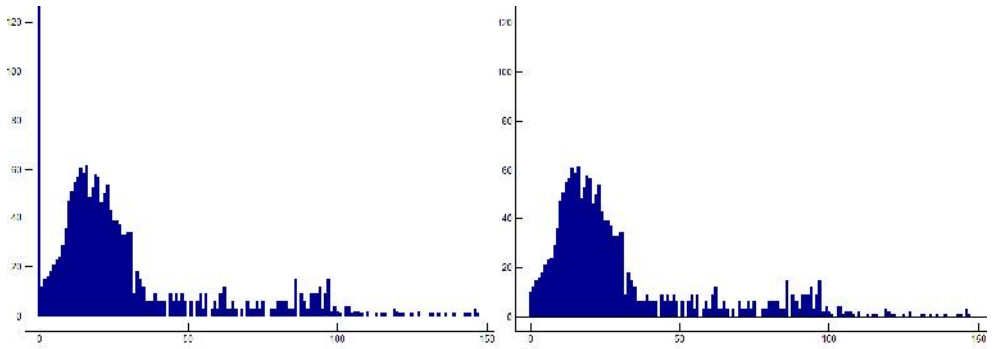

Figure 12: Cropped image histograms before and after the background pixel removal process.

Mean	Standard deviation	Smoothness	Skewness	Kurtosis	Uniformity	Entropy
23,89	41,12	0,0253	1,5093	3,7708	0,3837	3,5006
14,005	26,599	0,0107	2,6114	9,9624	0,3832	3,2066
9,599	20,278	0,0062	3,6318	19,3476	0,4521	2,693
14,26	28,26	0,0121	2,7642	10,6452	0,4067	3,1306
7,64	16,599	0,0042	4,2961	27,9502	0,4712	2,5044
10,68	24,341	0,0090	3,5394	16,6557	0,4480	2,8034
13,35	31,074	0,0146	3,0688	11,8181	0,4451	2,83
10,39	24,44	0,0091	3,6075	16,8258	0,4459	2,7165
Mean	Standard deviation	Smoothness	Skewness	Kurtosis	Uniformity	Entropy
62,52	44,835	0,0299	0,0672	1,6173	0,0142	6,6484
36,44	31,989	0,0154	1,5918	4,6788	0,0276	5,8443
29,05	26,06	0,0103	2,6247	9,9489	0,0348	5,38
39,1	34,907	0,0183	1,6456	4,5183	0,0236	5,9884
24,14	21,72	0,0072	3,4348	15,872	0,0425	5,0643
32,06	33,058	0,0165	2,099	6,5707	0,0301	5,6592
39,79	42,7298	0,0273	1,4854	3,7196	0,0321	5,6916
30,954	33,7949	0,0172	2,0716	6,3014	0,0436	5,346

Table 1: Colour features extracted before and after the background pixel removal process.

values change significantly after background pixel removal.

This approach allows for the extraction of chromatic features not only for whole WBCs but also for the nucleus and the cytoplasm, for 21 colour descriptors.

However, the descriptors based only on histograms frequently have drawbacks, as they do not provide information regarding the mutual position of the pixels. Some objects have a repeating pattern as the primary visual

characteristic, so it is necessary to consider both the intensity distribution and the position of the pixels having a similar grey level, as indicated by Haralick [24]. Then, we evaluated the descriptors applied to the grey level co-occurrence matrix (GLCM) calculated starting from sub-images in grey level. The following descriptors are evaluated: autocorrelation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, sum of squares (variance), sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation1, information measure of correlation2, inverse difference normalised and inverse difference moment normalised. These features are calculated for angles of 0, 45, 90 and 135 degrees.

Also these features are extracted from the images in grey tones for which the outer portion of the membrane has been cut off, and it is again necessary to exclude from the calculation of the features all of those pixels belong to the background. However, in this case, the binary mask is not used to calculate the number of pixels equal to zero; rather, it is used to determine the occurrences of adjacent pixels of value (0,0). Thus, given the nature of the GLCM, we must subtract the value in cell (0,0) of the GLCM calculated from the binary image from the value in the cell (0,0) in the GLCM calculated from the image in shades of grey. The result of this operation yield feature values that are very different from the previous ones (Fig. 13).

The total number of extracted features is 131: 30 shape descriptors, 21 colour descriptors and 80 texture descriptors.

6. Classification

Previously [25], the classification process was carried out using only 50 features using the support vector machine (SVM) classifier because this model is particularly suitable for binary classification problems for which the separation between classes depends on a large number of variables. In [25], the SVM was used with the standard configuration suggested by Hsu, Chang, and Lin in [26], providing SVM novices with a recipe to rapidly obtain acceptable results. However, the proposed approach is not good enough in some situations. In this work, the number of features is even higher, so we decided to test the SVM with the most common kernel: linear (L), quadratic (Q), polynomial (P) and Gaussian radial basis (R). For each kernel function, the parameters were tuned using optimization techniques in order to find the maximum accuracy value. The parameters obtained were as follows: for the

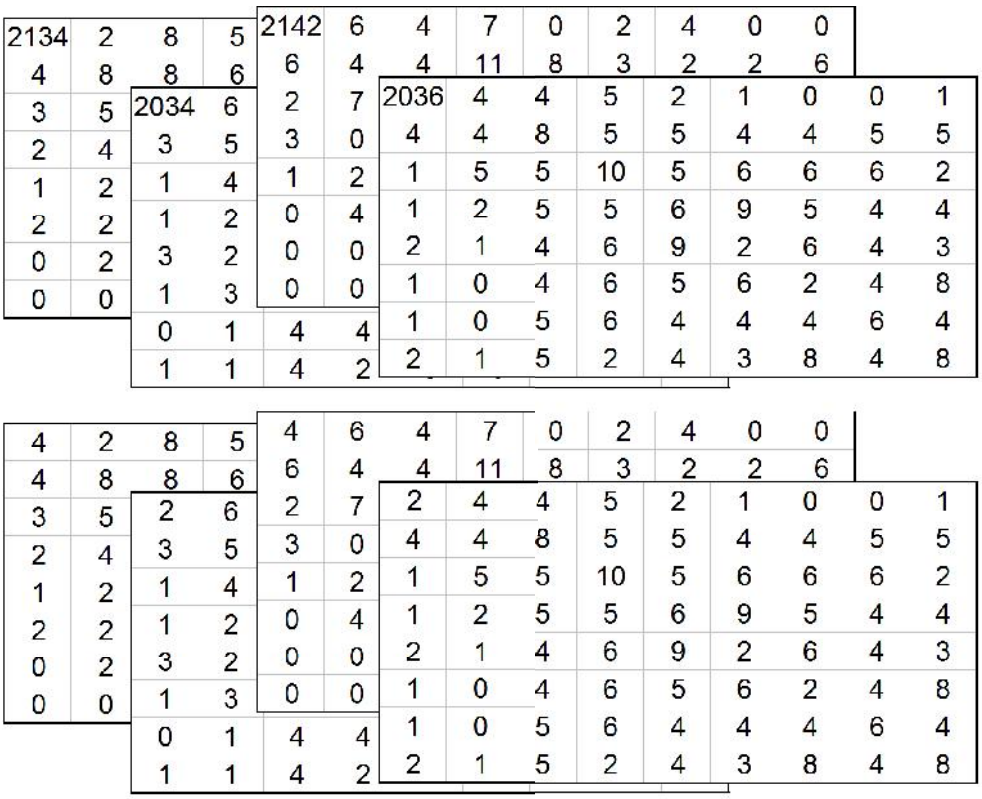

Figure 13: Grey level co-occurrence matrix for angles of 0, 45, 90 and 135 degrees before and after the (0,0) occurrence removal process.

L kernel, c = 1e-2; for the Q kernel, c = 1e2; for the P kernel c = 1e4 with polynomial order 3; and for the R kernel, c = 1e1 and γ= 1e1. To evaluate the goodness of SVM models, the results were compared to k-Nearest Neighbour (k-NN) [27] using the Euclidean distance measure with different values of k, Naive Bayes (NB) [28], [29] by a Gaussian (G) and kernel data distribution (K) and Decision Trees [30]. In addition to the type of algorithm used to induce the model, the performance of a model also depends on the size of the training and the test set. In particular, as the size of the training and the test sets decrease, the performance of the model depends on their specific composition, resulting in higher variance. Therefore, given the small size of the dataset used, the performance of the models is evaluated using a k-fold, cross-validation, re-sampling technique. Considering k = 10,

the whole dataset is randomly divided into 10 folds. The cross-validation process is repeated 10 times, using a different sub-sample as the validation data for testing the model and the remaining k-1 sub-samples as the training data each time. Finally, the 10 performances from the folds are averaged to achieve a single estimation. Once the estimate of instances predicted by the model of classification is obtained, it is possible to evaluate performance by comparing with the real class of instances, which, in this case, compares the class predicted by the classification model for a certain WBC with the class assigned to it by an expert haematologist. In a binary problem, as in our case, the instances are subdivided in positive and negative. For this particular problem, we defined the instances as positive when the WBCs were affected by leukaemia and negative when the WBCs were not suffering from leukaemia, and based on this definition, we calculated the *accuracy* value. Although accuracy is the most widely used metric, it considers each class of equal importance. Often, as in our case, it is more appropriate to use a metric that places the most importance on the correct classification of positive instances (we want to be sure to correctly classify WBCs with leukaemia). For this purpose, the *sensitivity* value is used.

7. Experimental evaluation

7.1. Image database

One problem we encountered while testing of our method was the absence of publicly available datasets. In fact, many authors tested their system with only a few sample images, or with their own datasets, which are not publicly available. Thus, we could not directly compare our findings with the results obtained by various proposed systems, limiting the reproducibility of the innovations proposed by similar systems. This is also the main reason why this study focused on ALL. Here, we used the acute lymphoblastic leukaemia image database ALL-IDB proposed by Donida Labati [4]. ALL-IDB is a public image dataset of peripheral blood samples from normal individuals and leukaemia patients, and it contains the relative supervised classification and segmentation data. 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 versions: in the first version (ALL-IDB1) can be used for both testing the segmentation capability of algorithms, as well as the classification systems and image pre-processing methods, and the second version (ALL-IDB2) is a collection

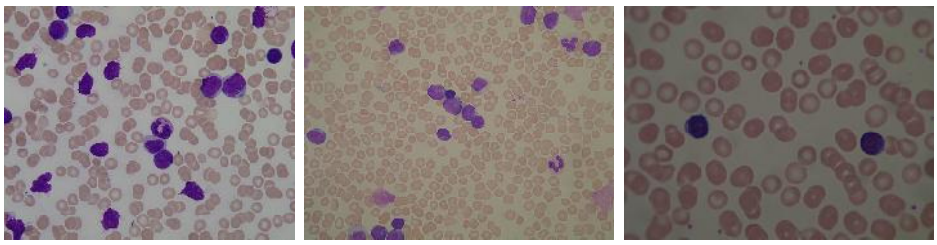

Figure 14: Original images from the ALL-IDB1

of cropped areas of interest from normal and blast cells that belong to the ALL-IDB1 dataset, so it can be used only for testing the performance of classification systems. In both versions of the dataset, each image has an associated text file containing the coordinates of the centroid of each candidate lymphoblast, which was manually labelled by a skilled operator and can be used as a ground truth. ALL-IDB1 (the version that we used for testing) includes 108 images in JPG format with 24-bit colour depth. Most of the images in this dataset were captured with an optical laboratory microscope at different magnifications, ranging from 300 to 500, coupled with a Canon PowerShot G5 camera (resolution is 2592x1944). The remaining images were acquired with a microscope at a constant magnification, coupled with an Olympus C2500L camera (resolution is 1712x1368). Images belonging to the ALL-IDB1 are shown in Fig. 14.

7.2. Results

There are many differences among the images in ALL-IDB1 in terms of resolution, magnification and lighting (Fig. 14). Therefore, testing was carried out using a subset of 33 images acquired from the same camera and under the same lighting conditions, so we could evaluate the performance of the proposed method for the detection and classification steps. In the early stages of the analysis we properly individuated 245 of 267 leukocytes, for an average accuracy of 92% (Fig. 15). The performance of the proposed method for WBC identification (shown in detail in Table 2) is excellent in most cases. The worst results were achieved in images with significant overlapping between leukocytes, which are difficult for human experts.

From sub-images containing individual leukocytes, we extracted a classification vector with a size of 1x245 and two different feature matrices, one with a size of 131x245 containing the previously described features and one with a size of 50x245 containing features used in previous work [25]. Using

Image	Manual count	Auto count	Accuracy
Image1	9	5	55%
Image2	10	10	100%
Image3	12	11	91%
Image4	7	4	57%
Image5	24	19	79%
Image6	18	18	100%
Image7	7	7	100%
Image8	17	16	94%
Image9	7	7	100%
Image10	12	12	100%
Image11	15	12	80%
Image12	12	12	100%
Image13	10	7	70%
Image14	5	3	60%
Image15	17	17	100%
Image16	16	16	100%
Image17	3	3	100%
Image18	8	8	100%
Image19	12	12	100%
Image20	2	2	100%
Image21	3	3	100%
Image22	5	5	100%
Image23	6	6	100%
Image24	4	4	100%
Image25	3	3	100%
Image26	5	5	100%
Image27	3	3	100%
Image28	2	2	100%
Image29	4	4	100%
Image30	3	3	100%
Image31	2	2	100%
Image32	2	2	100%
Image33	2	2	100%

Table 2: Performance of the proposed method for WBCs identification.

	4 colour feature [25]		7 colour feature		21 colour feature	
Classifier	ACC	S D	ACC	S D	ACC	S D
SVM-L	0,788	0,006	0,839	0,01	0,856	0,006
SVM-Q	0,787	0,009	0,858	0,02	0,813	0,018
SVM-P	0,793	0,009	0,801	0,016	0,83	0,01
SVM-R	0,792	0,009	0,87	0,006	0,884	0,006
k-NN	k=10 0,779	0,006	k=6 0,847	0,01	k=3 0,871	0,006
NB-G	0,759	0,002	0,758	0,004	0,834	0,006
NB-K	0,793	0,004	0,824	0,005	0,844	0,006
tree	0,718	0,008	0,813	0,011	0,866	0,014
Mean	0,776	0,006	0,813	0,010	0,852	0,007

Table 3: Experimental results using chromatic features: accuracy (ACC) and standard deviation (SD) are reported.

the same classifier, we aimed to demonstrate not only how the features proposed in this work are better but also that the new feature extraction process for chromatic and texture features increases the overall accuracy. First, we selected and tested only chromatic features. Then, we selected and tested texture features. Finally, we calculated the overall accuracy achieved with the new set of features. In all cases, the performance of the models was evaluated using a 10-fold Cross-Validation.

In the first case, for the 131x245 feature matrix, we selected seven chromatic features for whole WBC and all 21 chromatic features for whole WBC, nucleus and cytoplasm, whereas for the 50x245 feature matrix, we selected all four (the only ones present) chromatic features [25]. Table 3 displays the experimental results and (for the Nearest Neighbour model only the best results are shown).

In the second case, for the 131x245 feature matrix, we selected 80 texture features, whereas for the 50x245 feature matrix, we selected 16 texture feature [25]. Table 4 displays the experimental results (for the Nearest Neighbour model only the best results are shown).

In general, the chromatic and texture feature proposed in this work are more discriminant, and the new feature extraction process increases the overall accuracy of the process; in particular for the texture feature, which reached an accuracy of 0.906 vs. 0.773 for the previous method of feature extraction.

Finally, we compared the whole feature matrix (131x245 and 50x245) to show the accuracy achieved with the new feature set. Table 5 displays the

	16 texture feature [25]		80 texture feature	
Classifier	ACC	S D	ACC	S D
SVM-L	0,755	0,01	0,887	0,03
SVM-Q	0,753	0,008	0,858	0,011
SVM-P	0,741	0,008	0,856	0,009
SVM-R	0,747	0,01	0,906	0,011
k-NN	k=9 0,701	0,01	k=10 0,724	0,014
NB-G	0,773	0,004	0,852	0,002
NB-K	0,771	0,003	0,864	0,002
tree	0,74	0,028	0,806	0,019
Mean	0,747	0,011	0,844	0,009

Table 4: Experimental results using texture features: accuracy (ACC) and standard deviation (SD) are reported.

	Feat matrix 50x245 old feature set [25]		Feat matrix 131x245 new feature set	
Classifier	ACC	S D	ACC	S D
SVM-L	0,894	0,009	0,901	0,006
SVM-Q	0,887	0,011	0,9	0,005
SVM-P	0,895	0,011	0,901	0,007
SVM-R	0,906	0,007	0,932	0,008
k-NN	k=18 0,749	0,006	k=19 0,755	0,009
NB-G	0,809	0,005	0,85	0,003
NB-K	0,854	0,004	0,885	0,006
tree	0,87	0,015	0,863	0,02
Mean	0,86	0,01	0,873	0,009

Table 5: Experimental results using different feature set: accuracy (ACC) and standard deviation (SD) are reported.

experimental results (for the Nearest Neighbour model only the best results are shown).

Even for this example, the performance of the classification models benefited from the new feature set, reaching an accuracy of 0.932 using the SVM-R classification model vs. 0.906 using the previous set of features. Additionally, it is important to note that the decision to test the SVM classifier using different configurations and compare the results to each other and with

other classifiers allowed us to identify models based on different kernels that are suitable for this application. In all cases, our accuracy was above 0.9, while in our previous work [25], the accuracy achieved by the SVM classifier with a standard configuration reached a maximum value of 0.75. Moreover, the sensitivity value obtained in the test phase using the SVM classifier was never below 0.95 and reached a maximum value of 0.987 with the SVM-R classifier. The results obtained are comparable to those obtained by Deore [31], one of the few authors who used the ALL-IDB to test their method of image processing for the identification and classification of leukocytes with leukaemia. In fact, at the end of the classification stage, their accuracy value reached 0.9363. Unfortunately, this work does not provide any detail about the segmentation method used for WBCs, nor does it provide an accuracy value for their identification, so it is not possible to determine the number of samples used to train the classifier.

8. Conclusions

Here, we propose an innovative method for the automatic identification and classification of leukocytes using microscopic images, providing an automated procedure to support the recognition of ALL. Our results indicate that the proposed method is able to efficiently identify the WBCs present in an image and to properly classify leukoblasts with great accuracy. We have also proposed a new method for feature extraction from a cropped image that is excellent both for chromatic and texture features, and it can be used for feature extraction in many fields and for many applications.

The next step for this work will include further development of the identification phase. These improvements are necessary to increase the accuracy of counting WBCs and increase the overall accuracy of segmentation because accurate segmentation leads to a more robust extraction of shape features, which is essential for this type of problem. Moreover, it will be important to study and analyse the use of new features that may be decisive for this type of analysis. Then, the selection of the most discriminatory features will provide the highest level of accuracy. Further development of the proposed method could affect the separation of adjacent leukocytes, which is of considerable importance to account for all leukocytes in the image.

The proposed method for WBC identification also showed good results for images taken with different cameras and in different lighting conditions (Fig. 15). The decision to perform the test using only 33 images acquired

from the same camera and with the same illumination conditions was made because the extracted features are strictly related to the image resolution and lighting. To expand the size of the dataset and provide a greater number of useful examples to the classification model in the training phase, a feature extraction method that accounts for differences between images, avoids descriptors that are heavily dependent on image resolution and illumination, and scales and converts images into a representation that is less influenced by such characteristics is necessary. Finally, to increase the level of overall accuracy, the use of a multi-class classification model for the identification of various types of leukocytes and lymphoblasts is required.

In conclusion, the automatic method proposed in this study represents a rapid and low-cost technique to clarify ALL suspicion, complementary to immunophenotypic assessment of lineage commitment and developmental stage by flow cytometry. Further prospective studies are required to validate the sensibility and specificity of this method, especially in different lightning and resolution conditions.

Acknowledgments.

This work has been funded by Regione Autonoma della Sardegna (RAS) 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] 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

[2] Biondi, A., Cimino, G., Pieters,R., Pui, C. H.: Biological and Therapeutic Aspects of Infant Leukemia. Blood, vol. 96, no. 1, pp. 24-33, July 2000

[3] Inaba, H., Greaves, M., Mullighan, C. G.: Acute lymphoblastic leukaemia. The Lancet, vol. 381, no. 9881, pp. 1943-1955 June 2013

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] Mohapatra, S., Patra, D., Satpathy, S.: An Ensemble Classifier System for Early Diagnosis of Acute Lymphoblastic Leukemia in Blood Micro-

scopic Images. Journal of Neural Computing and Applications, Article in Press, 2013

[12] Cheewatanon, J., Leauhatong, T., Airpaiboon, S., and Sangwarasilp, M.: A New White Blood Cell Segmentation Using Mean Shift Filter and Region Growing Algorithm. International Journal of Applied Biomedical Engineering, vol. 4, pp. 30-35, 2011

[13] Lezoray, O., Cardot, H.: Cooperation of Color Pixel Classification Schemes and Color Watershed: a Study for Microscopic Images. IEEE Transactions on Image Processing vol. 11 no. 7 , pp. 783-789, 2002

[14] Gonzalez, R. C., Woods, R. E., Digital Image Processing, Prentice Hall Pearson Education, Inc.. New Jersey, USA, ed. 3, 2008

[15] Gonzalez, R. C., Woods, R. E., Eddins, S.L.: Digital Image Processing Using MATLAB, Pearson Prentice Hall Pearson Education, Inc., New Jersey, USA, ed.2, 2009

[16] Zack, G., Rogers,W., Latt, S.: Automatic Measurement of Sister Chromatid Exchange Frequency. Journal of Histochemistry and Cytochemistry vol. 25, pp. 741-753, 1977

[17] Serra, J.: Image Analysis and Mathematical Morphology, vol. I, Academic Press, London, 1982

[18] Serra, J.: Image Analysis and Mathematical Morphology, vol. II, Theoretical Advances, Academic Press, London, 1988

[19] Lindblad, J.: Development of algorithms for digital image cytometry. PhD thesis, Uppsala University, Faculty of Science and Technology, 2002

[20] Maurer, C. R., Rensheng, Q., Raghavan, V.: A Linear Time Algorithm for Computing Exact Euclidean Distance Transforms of Binary Images in Arbitrary Dimensions. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 2, pp. 265-270, February 2003

[21] Meyer, F.: Topographic distance and watershed lines. Signal Processing, vol. 38, pp. 113-125, July 1994

[22] Cseke, I.: A Fast Segmentation Scheme for White Blood Cell Images. In Proceedings of the IAPR International Conference on Image, Speech and Signal Analysis. IEEE Publisher, vol. 3, pp. 530-533, The Hague, Netherlands, 30 August-1 September, 1992

[23] Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, 1979

[24] Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics. vol. SMC-3, no. 6, pp. 610-621, November, 1979

[25] Putzu, L., Di Ruberto, C.: White Blood Cells Identification and Classification from Leukemic Blood Image. In Proceedings of the IWBBIO International Work-Conference on Bioinformatics and Biomedical Engineering. Editors: Ignacio Rojas, Francisco M. Ortuo Guzman. Copicentro Editorial, pp. 99-106, Granada, Spain, 18-20 March, 2013.

[26] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. A Practical Guide to Support Vector Classification. Available at http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf.

[27] Cover, T. M., Hart, P. E.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January, 1967

[28] Duda, R. O., Hart., P. E.: Pattern Classification and Scene Analysis. John Willey & Sons, New York, 1973

[29] Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. In Proceedings of the Tenth National Conference on Artificial Intelligence. Editors: Paul Rosenbloom and Peter Szolovits. Published by The AAAI Press, Menlo Park, California, pp. 223-228, San Jose, California, 12-16 July, 1992

[30] Quinlan, J. R.: Induction of decision trees. Machine Learning, vol. 1, no. 1, pp.81-106, 1986

[31] Deore, S. G., Nemade, N.: Image Analysis Framework for Automatic Extraction of the Progress of an Infection. International Journal of Advanced Research in Computer Science and Software Engineering, vol3, n. 6, pp. 703-707, June, 2013

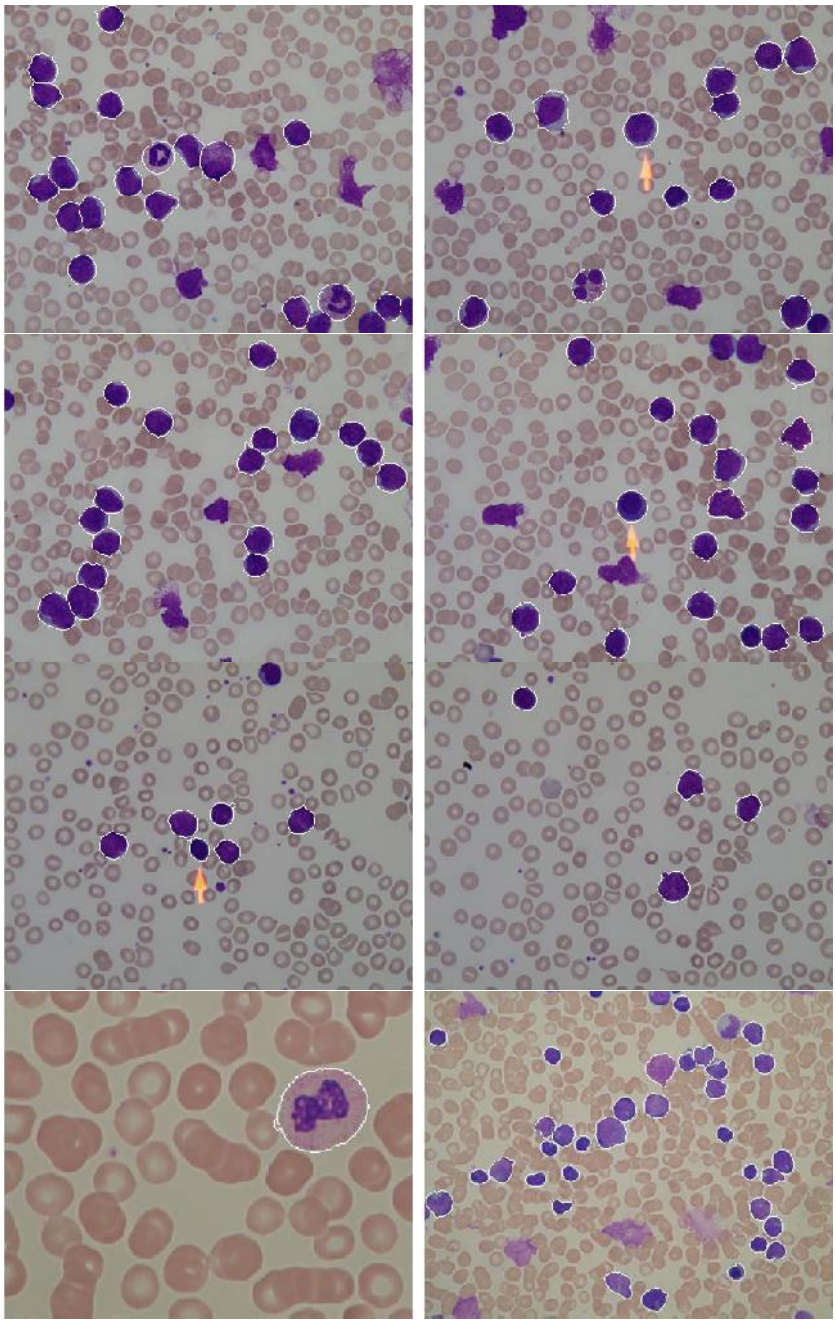

Figure 15: Original blood images after the leukocyte identification process (border highlighted).