Lymphoblast cell morphology identification to detect Acute Lymphoblastic Leukemia (ALL) using various color segmentation

Syadia Nabilah Mohd Safuan, Mohd Razali Md Tomari, Wan Nurshazwani Wan Zakaria, Nurmiza Othman

Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia
mdrazali@uthm.edu.my

Abstract. Acute Lymphoblastic Leukemia (ALL) is a disease that is defined by the uncontrollable growth of abnormal cell of lymphocyte, which is called lymphoblast. ALL patients cannot be left untreated as it can be fatal, and hence, early detection is very crucial for proper treatment suggestion. Conventionally, ALL analysis is manually done, which is time-consuming and highly dependent on the pathologist’s skills. Furthermore, it will be hard to pathologist as the number of sample increases, and also there are other cells inside the blood smear image, which will create confusion to them. Haematology counter is great to assist the process, but unfortunately, the cost is unbearable for some countries. For that reason, in this project, a computer-aided ALL detection with fast response is proposed. The system consists of five main modules which are image acquisition, pre-processing, segmentation, feature extraction and classification. For the process flow, firstly the color space correction based on l*a*b* color space is applied to standardize the color of the input image. Next, WBC segmentation is employed to locate the WBC region and consequently divided it into two parts which are the nucleus and cytoplasm based on a combination of color space analysis and Otsu thresholding. However, the segmented image contains noises and hence, is eliminated by using a combination of morphological filter and Connected Component Labelling (CCL). Then, the feature extraction process is made to study the nature of each individual cell using features derive from color, texture and geometrical properties. Lastly, lymphoblast classification is incurred to categorized the lymphoblast and the non-lymphoblast cell by employing Support Vector Machine (SVM) with a linear kernel function. As mentioned before, WBC segmentation is divided into two parts, and nucleus detection accuracy is higher than cytoplasm detection accuracy, which is 98.87% and 74.12% respectively. The presence of color space correction is analyzed, and the result is better with 96.92% accuracy for the presence of color space correction compared to 93.55% accuracy for without color space correction result. Classification performance is able to achieve 98.72% of accuracy.

1. Introduction
There are three major components in the human body which are red blood cell, white blood cell and the platelets, as shown in Figure 1. Every cell has its own function. To illustrate, RBC helps to transport oxygen from the lung to each part of the body [1], and WBC fights diseases, viruses and bacteria [2]. White blood cell is reported to be closely related to human body immunity. The immune system is important to help the body fights bacteria, viruses and any other elements that will give a bad effect to
our body system. Since we are surrounded by different kinds of people every day and also exposed to the possibility of getting viruses, a strong antibody system is crucial, and a big number of WBC can greatly help to determine such condition. The normal WBC count is in the range of 4500-10,000 (µƖ) while the abnormal count is beyond that range. A patient with low WBC count can potentially be diagnosed with diseases such as HIV and Lymphoma while high WBC count can trigger diseases such as Anemia, Leukemia and tissue damage. Basically, if a patient complains about their continuous sickness, the first test that will be performed by a doctor is the blood test since it will definitely explain a lot about the patient’s health condition. Necessary steps can, therefore, be taken to prevent serious illness. Other than that, by counting the amount of WBC in the body, the level of a person’s immunity system can be determined. Doctors can take further actions regarding the diseases and prevent the viruses from spreading all over the body.

![Sample of blood smear image](image)

**Figure 1.** Sample of blood smear image

White blood cell analysis includes white blood cell detection, and the count is crucial to identify the patient health condition. One of the diseases that can be related to white blood cell analysis is Leukaemia disease. Leukaemia is one of the top causes of death in the world in recent time [3]. Acute Lymphoblastic Leukaemia (ALL) is one type of leukaemia that needs to be taken seriously, which commonly attack 25% of young children [4]. The most important subject in detecting the disease is the presence of lymphoblast in blood smear image. Overproduction and continuous growth of this cell can lead to multiplication of abnormal and immature cells in the human body, thus contribute to the growth of leukaemia disease. ALL mostly attacks young children below 15 years old and adult over 50 years old. There are five types of WBC in blood smear which can be categorized into two categories which are granulocytes and agranulocytes. Granulocytes consist of Basophil, Eosinophil and Neutrophil while agranulocytes consist of Lymphocyte and Monocyte. Basically, Lymphoblast is an abnormal cell of Lymphocyte. Lymphoblast can be differentiated by its shape irregularities, a small cavity in the cytoplasm, spherical particles within the nucleus and the number of lobes in the nucleus [5].

Conventional ways in detecting WBC and classifying ALL is by manual analysis which is done by the pathologist. However, this practice can contribute to inconsistent result as it completely depends on the pathologist’s skills and experiences. It is also can create confusion as for some cases; different pathologist has a different result. Next, as the number of sample increases, it will be more difficult for the pathologist to analyse it manually. The process is time-consuming and tedious. However, the industry has come out with a haematology counter, which is automated, fast and accurate. But unfortunately, this machine cannot be afforded by some countries as it is costly.

This project proposed a Computer-Aided System (CAS) that can automatically detect WBC and classify Lymphoblast, which can work as good as the human eye. In this paper, the CAS framework consists of two main parts which are WBC identification and ALL classification. The system is able to prune out the WBC region and classify Lymphoblast cell to detect ALL disease accurately. WBC
identification and Lymphoblast classification accuracies are highly dependent on the WBC segmentation, which makes the process is very significant to the project. Pruning out the region of interest called segmentation is a tough process depending on the image itself. Segmentation is reported as a very important task as the accuracy of WBC or lymphoblast classification is highly dependent on this process [6]. This process can be divided into five main categories that are threshold-based methods, learning-based methods, active-contour-based methods, metaheuristic-based methods and saliency-based methods [2]. All these categories can be used for segmentation purposes based on different kinds of image type. This section focuses on the threshold-based method as for the uniform type of image like blood image, a threshold-based is reported to be the best and have a reliable performance with high running speed [2]. It is also reported in [7] that threshold method for segmentation has the highest number of references compared to other techniques. One of the methods is Otsu thresholding, and it is applied on the green channel of RGB to segment the RBC region as in [8]. Otsu thresholding also can be done on Y color band of CMYK to segment the cell region [9] and the result images of Otsu thresholding. Other than that, [10] proposed histogram thresholding to find the optimal values of threshold to segment the region of WBC. While in [11], K-mean clustering is applied for an initial segmentation of the cell nucleus. As for the work reported in [12], the RGB image was converted to CMYK to take only the Y color component as the region of interest had more contrast in it and threshold the image. To prune out the clumped region, work in [13] proposed the watershed image for segmentation purposes but to other images as well. It was also applied to identify lung cancer cell by setting the marker location to be regional minima which helped to emphasize the image information [14].

The classification process is the most crucial part of the system as the accuracy of the whole system performance is highly dependent on the accuracy of the classified cell. Classifiers such as Neural Network (NN) and Support Vector Machine (SVM) are widely used in this process. [15] compared three classifiers which are the nearest neighbour classifiers (kNN), feed-forward neural network (FF-NN) and linear Bayes Normal classifier. It was found that FF-NN showed a mean error of 0.0133, which was lower than the other two classifiers. However, the execution time was higher. Two classifiers which were Artificial Neural Network (ANN) and Support Vector Machine (SVM) were compared to each other [16]. ANN classifier had more fluctuation in the overall accuracy, while SVM was more stable. SVM also was superior compared to ANN. Another work used SVM as well, and they achieved a classification accuracy of 93% [17] and 95% [11].

2. System overview

The system contains seven main blocks which are color space correction, nucleus and cytoplasm segmentation, morphological filter, feature extraction, lymphoblast classification and lastly, assessment procedure. The input image was obtained from a light microscope and also known as the blood smear image. Initially, to cater to the issue regarding the image acquisition variability condition, the image’s color intensity needed to be transformed into one a standard color characteristic. The color correction process using L*a*b* color space correction [18] was done to the original image. Next, the WBC segmentation was done by removing RBC and background using Otsu thresholding and combination of RGB, CMYK and HSV color space analysis. Morphological filter was used to remove the remaining noises after the segmentation process. Features extraction of shape feature and texture feature were taken to collect the data on WBC nucleus and cytoplasm. The data were fed to a classifier, which was SVM to classify and differentiate lymphoblast from any other cells.

There were 108 samples of blood smear images which were taken from ALL-IDDB1 database as depicted in Figure 2. The images were public blood sample which consisted of two different conditions which were normal blood sample and ALL detected blood sample. The images were saved in the JPEG files. Deep colored purple is the WBC region, while light and pinkish-purple area is the RBC area and background.
Figure 2. Sample of blood smeared image

2.1. Color space correction.
The color space correction function is to make the background color more significant, which will make it better for the background elimination process [19]. In this project, an approach of mapping the color distribution of an under the stained image to that well-stained target image was made. The input image underwent the process of color space correction, which made the input image color intensity changed with respect to the targeted variance and mean. An accurate representation of color needed to be achieved in order to make an ideal color correction. RGB color space does not meet the above criteria because it is not a uniform color space and some colors have negative coefficients. By using this method of color space correction, all the input images will follow the fixed target color space, which will give the uniform color space and get the ideal color correction. A system has been developed that automatically extracts the blood-smeared images, compute image color in color space, and corrects the color based on a standard calibration target. This method represents the matching color distribution of the source image and the target image by using the L*a*b* color space. L*a*b* color space is used in this step because it minimizes the correlation between channels for many natural scenes [18].

2.2. Extraction and segmentation of white blood cell (WBC).
After the color correction phase, the image needs to be segmented to localize the respective WBC region. The localization of WBC is accomplished by eliminating other regions and particles such as RBC, platelets and background. However, segmenting WBC region is tough as its morphological features are not consistent. This is because WBC consists of two parts which are nucleus and cytoplasm. Both parts can be differentiated by the color intensity level as the nucleus has higher color intensity than the cytoplasm and nucleus is the inner part of WBC while the cytoplasm is the outer part.

In our implementation, the segmentation process was done by applying Otsu thresholding on a color space analysis of RGB, CMYK and HSV as depicted in Figure 3. As can be seen, R, G, S, C, M and K has completely eliminated the background and RBC region. On top of that, we used two different kinds of approaches to segment the WBC region. The first approach was using the single band color analysis, and the second approach was using a combination of two chosen single bands to find the best result of WBC counting. The single color band was useful for locating the nucleus while the combination of the color band could be used to detect both the nucleus and cytoplasm area, which would be useful for WBC identification.
Figure 3. Result of color channel segmentation of green (G), red (R), blue (B), hue (H), saturation (S), value (V), cyan (C), magenta (M), yellow (Y) and key (K).

2.3. Morphological filter
The segmented images contain noises and unwanted regions that need to be diminished as it will eventually disturb and affect the system performance. In order to allocate this problem, a morphological filter was applied to extinguish the noise in the extracted image. This operation included the process of erosion and dilation. Erosion removed pixels on object boundaries which shrunken the object while dilation added pixels on object boundaries which grow the object. Both erosion and dilation were applied one time for all images to improve segmentation quality. Erosion was applied to the segmentation’s result images first followed by the dilation process. After that, Connected Component Labelling (CCL) was applied with the value of 150. As a result, the filtered image produced was better than the unfiltered image. The noises or small particles have been removed from the image, and the quality is improved. After the noise was eliminated, the white blood cell nucleus and cytoplasm were extracted based on its geometrical features.

2.4. Clumped area extraction
Next, after the image was free from noises, there was only a nucleus or cytoplasm region in the image. However, some of the regions were clumped and overlapped with each other. Feature extraction could not be done without extracting the clumped cell into an individual cell. This process was done by applying watershed segmentation to prune out the overlapping cell. After watershed determined the boundary of each cell, the cell area was pruned out by using a bounding box. This would make the clumped cell separated into the individual region for feature extraction purposes.
2.5. Feature extraction
After the nucleus and cytoplasm segmentation process, feature extraction was applied to the binary image. Feature extraction is a technique of redefining a large set of redundant data into a set of features of reduced dimension. In this project, shape features, texture features, and color feature were taken from segmented nucleus and cytoplasm. Features included are area, perimeter, compactness, convex area, solidity, ratio cytoplasm to nucleus area, form factor, energy, correlation, entropy, mean, standard deviation, entropy, RMS, variance, smoothness, kurtosis, skewness and Inverse Different Moment (IDM). Binary sub-image is used for shape features while for texture and color, grey level sub-image is used as depicted in Figure 12. These features extraction was very crucial as the classification process was highly dependent on the features extracted.

2.6. Lymphoblast classification
There are many classifiers that can be used to classify the lymphoblast. However, in this project, Support Vector Machine (SVM) was employed. It works by separating surface in the input surface of the data set. Basically, it creates hyperplane in the F space that has maximum separation. The linear hyperplane is the boundary in SVM. SVM is a more powerful tool than Neural Network in certain applications.

2.7. Assessment procedure
This section defines the quantitative assessment of WBC identification for nucleus and cytoplasm. The performance of the system was evaluated by using the statistical measurement of the accuracy, precision, recall and specificity. The original images were edited manually to get the ground through images which were used for this calculation. These parameters were calculated in the area. The area of automatic and manual segmentation of the white blood cell image was calculated.

3. Result and analysis

3.1. WBC identification performance
In this section, WBC detection performance is calculated based on its ability to correctly identify the area of the nucleus and cytoplasm. The filtered segmentation image is compared to the ground truth to obtain the accuracy, specificity and sensitivity of the system. Color analysis used for nucleus identification is S of HSV as it achieves highest counting accuracy while for cytoplasm identification, H-Y is used.

The value of TP, TN, FP and FN is taken for 30 images in the database. Table 1 summarizes the average value of accuracy, specificity and sensitivity for nucleus and cytoplasm identification. Based on the table, it can be seen that nucleus identification achieves higher accuracy compared to cytoplasm identification. Nucleus detection is able to obtain more than 95% for accuracy, specificity and sensitivity result.

| WBC Identification | Accuracy (%) | Specificity (%) | Sensitivity (%) |
|--------------------|--------------|-----------------|-----------------|
| Nucleus (S)        | 98.87        | 96.87           | 99.10           |
| Cytoplasm (H-Y)    | 74.12        | 65.32           | 99.87           |

Judging from the result of WBC identification, nucleus based segmentation gives better result compared to cytoplasm based segmentation. Hence, the next process, which is the feature extraction process, is proceeded by using nucleus based segmentation.

3.2. ALL classification
The last process and task for this project was image classification. Since the images consisted of lymphoblast and non-lymphoblast cell, the system required the classification for both types of cells. In
this system, the classification was done by using Support Vector Machine (SVM) method. The basic idea of SVM was to create linear hyperplane in the input space between classes. This method was chosen as it was more stable and superior than the Artificial Neural Network (ANN).

The classification was made to differentiate between classes which enabled the system to provide final results. There are 234 WBC sub-images in total and features matrix with a size of 16 x 234 and classification vector of a size 1 x 234 was created. The 16 x 234 file was the feature extraction and data file while 1 x 234 file was the labelled file that labels which image belonged to lymphoblast and vice versa.

Firstly, both files were loaded into the system, and the classification performance was evaluated using a 10-fold Cross-Validation. The data were trained by using kernel function, and after the process, the data were classified into their classes. Kernel Function of Polynomial and Rbf was compared to each other in order to choose the best kernel function to be used in the system. Both kernel functions were able to obtain 100% for training while classifier performance of testing was calculated for both functions for 10-fold and the results were as tabulated in Table 2. As can be clearly seen from the table, the polynomial kernel function provides the best result for classification compared to Rbf kernel function.

| Table 2. Classification accuracy of polynomial and Rbf kernel functions using 10-fold cross-validation |
|---|---|---|
| i | Polynomial | Rbf |
| 1 | 95.83 | 91.30 |
| 2 | 95.74 | 93.62 |
| 3 | 97.14 | 94.29 |
| 4 | 97.85 | 92.47 |
| 5 | 98.28 | 92.14 |
| 6 | 98.57 | 92.14 |
| 7 | 98.97 | 92.64 |
| 8 | 98.92 | 92.47 |
| 9 | 98.57 | 92.86 |
| 10 | 98.72 | 92.74 |

4. Conclusion and future works
In this paper, an empirical framework for automatically detecting and counting the number for WBC in blood smear image was proposed. This automatic system was developed under a computer vision system for image processing purposes. Several methods were compared to each other to find the best method to identify the nucleus and cytoplasm region in the image. However, for each method, the main blocks were the same, which consisted of color space correction, nucleus segmentation, cytoplasm segmentation, post-processing, feature extraction and lymphoblast identification. Firstly, the original blood smear image’s color intensity was corrected using L*a*b* color space correction based on the mean and standard deviation value of l, a and b. The mean values that were set for l, a and b were 81.9863, 4.5579 and 1.2176 while for standard deviation values were 4.7984, 4.7248 and 3.0947 respectively. The analysis was made to evaluate the presence of color space correction. Eventually, based on the evaluation made, color space correction was highly needed as it provided higher accuracy than images without color space correction. Next, the corrected images were segmented using the Otsu Thresholding method. In this process, various segmentations were done using color space analysis of RGB, CMYK and HSV. S showed the best result to segment nucleus region while for cytoplasm region, a combination of color space analysis was used. The segmented image contained noises and eliminated by applying a morphological filter. Morphological filter consisted of a few processes which were dilation, erosion and CCL. Dilation and erosion process was applied once for each sample image. Erosion process resulted in shrinkage of the white blood cell region and dilation increased the size of the region again. Pixels that were lower than 150 was removed by applying CCL.
The noises in the blood-smeread was completely removed, which also increased the quality of the segmented image. Two sub-images were created, which are grey-level and binary to extract the cell features. All the data for 16 feature extractions were taken to be fed into the classifier. In the feature extraction process, the cell’s nature was studied, and the data were collected. Lastly, the final step was the classification of lymphoblast cell. In this process, the Support Vector Machine (SVM) was used with a polynomial kernel function. There was a total of 242 sub-images that were fed in the classifier. At the end of this process, the ALL classification accuracy was calculated, and the system achieved 98.72% accuracy.

In future, improvements can be made by adding experts’ validation to be compared to the system performance. Next, feature extraction is expected to be widened and explore to have only the most suitable features included. Lastly, this system can be improvised by classifying other types of WBC in blood smear images.

5. Acknowledgement

The authors would like to thank the Ministry of Education (MOE) and Universiti Tun Hussein Onn Malaysia (UTHM) for supporting this research under Postgraduate Research Grant (GPPS) (Vot. no. H400).

References

[1] M. Tomari, M. Razali, W. Zakaria, and W. Nurshazwani, "An empirical framework for automatic red blood cell morphology identification and counting," 2015.
[2] C. Zhang, X. Xiao, X. Li, Y.-J. Chen, W. Zhen, J. Chang, et al., "White blood cell segmentation by color-space-based k-means clustering," Sensors, vol. 14, pp. 16128-16147, 2014.
[3] P. K. B. W. Stewart, "World Cancer Report", 200
[4] T. Allen and A. Ahmed, "Immunotherapy and Acute Lymphoblastic Leukemia," EC Cancer, vol. 2, pp. 141-147, 2016.
[5] L. Putzu and C. Di Ruberto, "White blood cells identification and classification from leukemic blood image," in Proceedings of the IWBBIO international work-conference on bioinformatics and biomedical engineering, 2013, pp. 99-106.
[6] S. Kothari, Q. Chaudry, and M. D. Wang,"Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques," in Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on,2009, pp. 795-798.
[7] D. Kolhatkar and N. Wankhade, "Detection and counting of blood cells using image segmentation: A review," in Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference on, 2016, pp. 1-5.
[8] M. Tomari, J. Lias, R. Musa, W. Zakaria, and W. Nurshazwani, "Development of red blood cell analysis system using NI vision builder AI," 2015.
[9] N. Ghane, A. Vard, A. Talebi, and P. Nematollahy, "Segmentation of White Blood Cells from Microscopic Images Using a Novel Combination of k-means Clustering and Modified Watershed Algorithm," Journal of medical signals and sensors, vol. 7, p. 92, 2017.
[10] Y. M. Alomari, S. N. H. Sheikh Abdullah, R. Zaharatul Azma, and K. Omar, "Automatic detection and quantification of WBCs and RBCs using iterative structured circle detection algorithm," Computational and mathematical methods in medicine, vol. 2014, 2014.
[11] S. Mohapatra and D. Patra, "Automated cell nucleus segmentation and acute leukemia detection in blood microscopic images," in Systems in Medicine and Biology (ICSMB), 2010 International Conference on, 2010, pp. 49-54.
[12] L. Putzu and C. Di Ruberto, "White blood cells identification and counting from microscopic blood image," World Academy of Science, Engineering and Technology, vol. 73, pp. 363-370, 2013.
[13] L. B. Dorini, R. Minetto, and N. J. Leite, "Semiautomatic white blood cell segmentation based on multiscale analysis," *IEEE journal of biomedical and health informatics*, vol. 17, pp. 250-256, 2013.

[14] S. L. Kumar, M. Swathy, S. Sathish, J. Sivaraman, and M. Rajasekar, "Identification of lung cancer cell using watershed segmentation on CT images," *Indian Journal of Science and Technology*, vol. 9, 2016.

[15] 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*, 2005.

[16] S. H. Rezatofghi and H. Soltanian-Zadeh, "Automatic recognition of five types of white blood cells in peripheral blood," *Computerized Medical Imaging and Graphics*, vol. 35, pp. 333-343, 2011.

[17] S. Mohapatra, S. S. Samanta, D. Patra, and S. Satpathi, "Fuzzy based blood image segmentation for automated leukemia detection," in *Devices and Communications (ICDeCom), 2011 International Conference on*, 2011, pp. 1-5.

[18] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, pp. 34-41, 2001.

[19] J.-L. Fan and B. Lei, "A modified valley-emphasis method for automatic thresholding," *Pattern Recognition Letters*, vol. 33, pp. 703-708, 2012.