Automatic Identification of Acute Lymphoblastic Leukemia on Blood Cell An image Using Geometric Features

Automatic Identification of Acute Lymphoblastic Leukemia on Blood Cell An image Using Geometric Features

Rahmi Rizkiana Putri[1], **Eka Prakarsa Mandyartha**[2], **and Annisaa Sri Indrawanti**[1]

[1]Institut Teknologi Adhi Tama Surabaya
[2]Universitas Pembangunan Nasional Veteran Jawa Timur

rahmirizkianaputri@gmail.com

Abstract. Acute Lymphoblastic Leukemia (ALL) is a human blood cancer which causes most deaths among other types of blood cancer. It commonly occurs in children and teenagers. ALL is recognised when there is a presence of lymphoblast cell in human blood. Haematologists diagnose ALL using visual microscopic observation from a human periphery blood sample. The inspection takes a long time and is limited to the subjective experience of the haematologist, and therefore, causes differences in the diagnostic results among haematologists. In this study, the researchers proposed automatic ALL Identification system to solve the problem. The method used an image processing technique that analysed a human blood cells morphology. The system identified the differences between normal white blood cells and lymphoblast cells by considering the shape and the size of the nucleus and also the cytoplasm. There are 33 microscopic an image data of a human blood cell that were used in this research. Geometric features used in the study were an area, perimeter, eccentricity, equivDiameter, solidity, roundness and circularity of a human blood the nucleus and cytoplasm and also the ratio of area and the perimeter of the cells and the nuclei. Firstly, the nucleus and cytoplasm were extracted from the other component of a human blood cell using Gram-Schmidt as an orthogonalisation-based segmentation method and then using Otsu threshold. Support Vector Machine was then used to classify whether the white blood cells were lymphoblasts or normal cells. Lastly, ALL identification was evaluated by the parameter of sensitivity and misclassification rate metrics using k-fold cross-validation.

1. Introduction

Acute lymphoblastic leukaemia (ALL) is a type of blood cancer caused by malignant blood cells [6,15,16,17] usually recognised by the presence of abnormal white blood cells called lymphocytes in the peripheral blood [3]. Those abnormal Lymphocyte cells are white blood cells which are imperfect (*immature*), also called leukemic cells or lymphoblast. ALL most commonly occurs in children aged 2 to 5 years old and adults aged 50 years old or older [13,12,9]. ALL diagnostic is challenging to do because ALL has quite unsuspicious symptoms such as fever, anaemia, lethargy, and pain in the joints [12]. Haematology specialist performs ALL diagnosis by examining a peripheral blood sample. The blood test includes complete blood count (with flow cytometry) and visual examination using a microscope [2]. Full blood count is performed to count the amount of each of the 5-type white blood cells. This automatic counting instrument uses flow cytometry to measure cytochemical contained in the blood and to observe the blood cells' physical characteristics using a light detector [9]. The light detector uses electric impedance or fluorescence to identify the white blood cells' type. Although the

results given by the counting instrument is very accurate, the instrument itself is unable to detect the morphology of the abnormal white blood cells, which is why the microscopic visual examination is also necessary. Visual examination is performed to identify lymphoblast based on the morphology. As a visual examination of the blood sample takes a long time, and the diagnosis results among haematologists cannot be used as a standard since the accuracy still depends on the experts' proficiency [9,6,12,5]. To solve the problem above, ALL automatic identification system is proposed. ALL automatic identification system analyses the morphology of blood cell an image using peripheral blood sample imaging; the result is not subjective; with standardised accuracy [9,12]. The difference between normal lymphocytes and abnormal lymphocytes is determined by the changes in the shapes and the sizes of the nucleus and cytoplasm. Normal lymphocytes have regular and smaller the nucleus shape and have the nucleus ratio with smaller cytoplasm if compared to lymphoblast cells. Therefore, the shape and the size of cytoplasm can be used to differentiate normal lymphocytes from lymphoblast (ALL cell). Geometric feature descriptors used in this study were the width, perimeter, eccentricity, equivalent diameter, solidity, roundness, and circularity of white blood cells' the nucleus and cytoplasm, as well as the ratio of area and the nucleus's perimeter to the cell. The nucleus and cytoplasm had to be first extracted from other blood component using Gram-Schmidt orthogonalisation-based segmentation method and Otsu threshold. Classifier Support Vector Machine (SVM) was also used to classify whether a white blood cell was considered normal or lymphoblastic.

2. Literature Review

2.1. Gram-Schmidt Orthogonalization.

An image segmentation using Gram-Schmidt orthogonalisation theory was implemented in the study proposed by [5] and [10]. Gram-Schmidt orthogonalisation is a method used for vector set orthogonalisation of the inner product space. Set of orthogonal vector *S* ' = {*u*1,..., *un* } generated from Gram-Schmidt orthogonalisation of vector set *S* ={*v*1,..., *vn* } has the same sub-space with Gram-Schmidt S. projection as shown in the equation (1).

$$proj_u v = \frac{(u,v)}{(u,u)} u = (u,v)\frac{u}{(u,u)} \quad (1)$$

With (u,v) as the multiplication in vector *u* and *v*. Vector *v* is orthogonal toward vector *u*. The Gram-Schmidt process is defined by equation (2) as follows:

$$u_k = v_k - \sum_{j=1}^{k=1} proj_{u_j} v_k, e_k = \frac{u_k}{\|u_k\|} \quad (2)$$

Vector $u_1,\ldots, u_k$ is an orthogonal vector, and vector $e_1,\ldots., e_k$ is a normalised vector that generates a set of orthonormal. Based on the method above, for vector S = $\{v_1,\ldots., v_n\}$, it can be found a vector with maximum orthogonality toward desirable vector, and also a vector with minimum orthogonality toward other vectors in N-dimension space. Vector w_k can be determined by using equation (3):

$$w_k = v_k - \sum_{j=1}^{k-1} proj_{v_j} v_k, k = 1,2,3 \quad (3)$$

Therefore, the result of the inner product of set S with w_k is shown by equation (4) below:

$$\begin{cases} \langle \boldsymbol{v_j}, \boldsymbol{w_k} \rangle = 0 \quad j \in 1, \ldots., n \; dan \; j \notin k \\ \langle \boldsymbol{v_k}, \boldsymbol{w_k} \rangle = K \quad K \neq 0 \end{cases} \quad (4)$$

2.2. Multilevel Thresholding Otsu

Otsu threshold method is used to find single threshold value which divides pixels into a bi-level value. The development of Otsu bilevel thresholding method for cases in which there are more than one threshold value in multilevel thresholding that has been proposed by [1], [8] and [5].

3. Research Methodology

The imaging data observed in the study were microscopic imaging of peripheral blood smear sample which contained lymphoblast (ALL) cell and normal cell. The microscopic imaging of peripheral blood smear sample itself included regions such as red blood cell, white blood cell, platelet, and background

(Figure 1). The white blood cell is comprised of the nucleus and cytoplasm. Error! Reference source not found. shows the structure of a white blood cell.

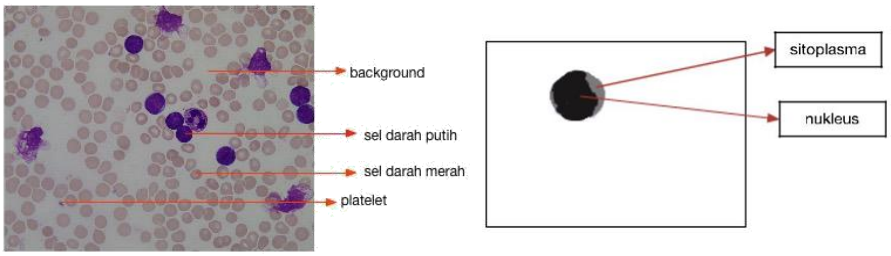

Figure 1. The physical structure of white blood cell

This study was divided into two parts. Those parts were segmentation process for extracting white blood cells' the nucleus and cytoplasm, which then followed by ALL identification. The white blood cells identification using Gram-Schmidt orthogonalisation was then followed by Otsu threshold. The identification was performed to divide the nucleus and cytoplasm. Geometric features used in the study were width, perimeter, eccentricity, equivDiameter, solidity, roundness, and circularity of the nucleus and cytoplasm, as well as the ratio of area and the edge of the cell and the nucleus. Those geometric features were used as attributes to classify white blood cells into lymphoblast or normal cells. Block diagram in Figure 2 represents the flow diagram of the complete process of automatic ALL identification.

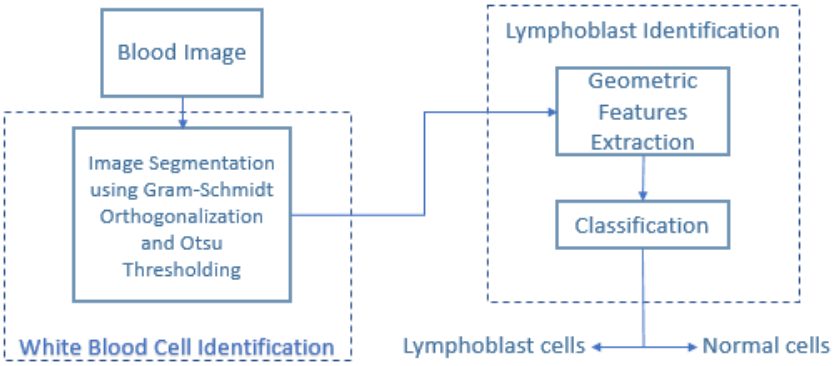

Figure 2. Flow diagram of the complete process of automatic ALL identification

3.1 White Blood Cell Segmentation

White blood cell segmentation was performed to extract white blood cells' nuclei and cytoplasm from the other blood components (red blood cell and platelet), including the background. In this study used segmentation method proposed by [5]. The algorithm steps of white blood cells segmentation are as the following:

1. Determine the feature vector for every pixel using imaging RGM component
2. Calculate weight vector *w* to strengthen the colour vector of the white blood cell *region* and to weaken unwanted colour vectors.
3. Calculate the inner product between weight vector *w* and the vector of input an images' pixels to produce composite an image with maximum intensity in the violet colour regions and minimum depth for the region of other colours, then continue with thresholding, and the result will be in binary an images

4. Cut input images into sub-an images, based on bounding box that surrounds connected component of regions in binary images.
5. For every sub-an image, do thresholding that finds out two threshold values using Otsu function: between class variance

3.2 Geometric Features Extractions

The ratio of area between the nucleus and cell (Figure 3) was determined by the surface width (the number of pixels) of the binary an image, Snu and Ssel, for the nucleus and cell area consecutively. The ratio of area between the nucleus and cell was defined by equation (5).

$$\mathrm{NCR}_{\mathrm{area}} = \frac{S_{nu}}{S_{sel}} \tag{5}$$

The perimeter ratio between the nucleus and cell was determined by the perimeter of the binary an image (the number of boundary pixel), Pnu dan Psel. The perimeter ratio between the nucleus and cell was defined by equation (6).

$$\mathrm{NCR}_{\mathrm{peri}} = \frac{P_{nu}}{P_{sel}} \tag{6}$$

Seven geometric feature values were counted. Those values were width, perimeter, eccentricity, equivDiameter, solidity, roundness and the nucleus circularity. Width value is the number of binary image pixels. Perimeter value is determined by the number of pixels surrounding (boundary) the binary an image (Figure 3(a)). Eccentricity value is determined by using the binary moment to count the ratio between major axis length and minor axis length from ellipse inertia as shown in equation (7).

$$\mathrm{Eccentricity} = \sqrt{\frac{\mathrm{major\_axis\_length}^2 - \mathrm{minor\_axis\_length}^2}{\mathrm{major\_axis\_length}}} \tag{7}$$

EquivDiameter is the diameter of a circle of the same width (region). EquivDiameter is determined using equation 8.

$$\mathrm{EquivDiameter} = \sqrt{\frac{4 \times \mathrm{width}}{\pi}} \tag{8}$$

Solidity is pixels proportion in the convex hull which was also in the region and was counted using equation (9). The convex hull is the smallest convex polygon which contains region. Convex area is the number of pixels on the convex an image (convex hull with all pixels in the hull filled), as shown by figure 3(b).

$$\mathrm{Solidity} = \frac{\mathrm{width}}{\mathrm{convex\_area}} \tag{9}$$

Roundness was calculated using equation (10). Significant axis length is the pixel length of major axis ellipse (Figure 3(c)) which is the same with the second central moment from the region.

$$\mathrm{Roundness} = \frac{4 \times \mathrm{width}}{\pi \times \mathrm{major\_axis\_length}^2} \tag{10}$$

Lastly, Circularity or the form factor is calculated using equation (11).

$$\mathrm{Circularity} = \frac{\mathrm{perimeter}^2}{4 \times \pi \times \mathrm{luas}} \tag{11}$$

Therefore, the number of the element from geometric feature vectors was fourteen, seven features each for the nucleus and cytoplasm.

(a) **(b)** **(c)** **(d)**

Figure 3. Geometric features [14]: (a) perimeter, (b) convex area, (c) major axis length, (d) minor axis length.

3.3 Classification

For the purpose of training and classifying, the algorithm of classifier Support Vector Machine (SVM) with linear kernel function was used. Training algorithm took training data to be trained by classifier, so that training model could be determined. Training model was then used in classifying new data into two classes: ALL cells and normal cells.

4. Results and Discussion

The *Dataset* of the microscopic peripheral blood an image (blood-smear images) used in the study was acute lymphoblastic leukaemia an image database ALL-IDB1 provided by Donida Labati, *et al*. [4]. ALL-IDB1 consists of 108 24-bit resolution JPG images. For system trial in this study, the researchers used 35 blood cell images comprised 269 white blood cells, taken by Olympus C2500L camera (1712 x 1368 pixel resolution) under the same lighting setting. The number of the successfully segmented white blood cell was 263 white blood cells with an average accuracy of 98%. Each segmented white blood cell had its 16 geometric features calculated. To evaluate the suggested method, k-fold cross-validation (k = 10) method was used. Since the problem of the study was the binary classification, which was the identification of ALL cells and normal cells, the classifier algorithm Support Vector Machine (SVM) with linear kernel was used. The classification result was analysed using a confusion matrix. This matrix, as shown in Table 1, contains information about the actual value and image prediction testing obtained from SVM. True Positive (TP) is the number of ALL cells correctly classified as ALL cells. False Negative (FN) is the number of ALL cells classified as normal cells. False Positive (FP) is the number of normal cells classified as ALL cell, and True Negative (TN) is the number of cells correctly classified as a normal cell.

Table 1. Confusion Matrix

	Actual Class	
Prediction Class	ALL Cell	Normal Cell
ALL Cell	True Positive (TP)	False Positive (FP)
Normal Cell	False Negative (FN)	True Negative (TN)

The result was then analyzed using two evaluation measures which were sensitivity and misclassification rate. Sensitivity is the probability of ALL cell being identified correctly defined by the equation (12) below:

$$\text{Sensitivity} = TP/(TP+FN) \quad (12)$$

Misclassification determines the number of cells that are classified incorrectly; Misclassification is defined by the equation (13) below:

$$\text{Misclassification rate} = (FP+FN)/(TP+FP+FN+TN) \quad (13)$$

The calculation results are shown in Table 2.

Table 2. The Evaluation of the Proposed Method Performance

Evaluation Matrix	Value
TP	247
TN	6
FP	6
FN	4
Sensitivity	0.984
Misclassification	0.061

Table 2 shows that most ALL cells were identified as normal cells which means that ALL cells were correctly classified, proven by 98.4% Sensitivity and 6.1% misclassification rate.

5. Conclusion

In this study proposed the use of automatic acute lymphoblastic leukaemia (ALL) cells identification method by processing and analysing peripheral blood-smear an images. The technique provided an automated procedure to support medical examination, especially to help acute lymphoblastic leukaemia screening. The process was initialized by identifying white blood cell using threshold technique, which then continued by geometric features extraction for classification using SVM Linear. The results of the experiment showed that the proposed method was able to classify all extracted white blood cells; both lymphoblastic cells (ALL cells) and normal cells, with 98.4% sensitivity and 6.1% misclassification rate. The focus for further development of the proposed method was on the phase of separating overlapped blood cells, which was very important in calculating all individual white blood cells on the microscopic peripheral blood an image. In addition, the multi-class classifier algorithm was also needed in identifying the types of acute lymphoblastic leukaemia.

6. References

[1] Huang, D C, Hung, K D, & Chan, Y K 2012. *A computer-assisted method for leukocyte the nucleus segmentation and recognition in blood smear images*. *Journal of Systems and Software*, *85*(9), 2104-2118.

[2] Inaba, H, Greaves, M, & Mullighan, C G 2013. *Acute lymphoblastic leukaemia*. *The Lancet*, *381*(9881), 1943-1955.

[3] Isnani, N, Perwitasari, D A, Andalusia, R, & Mahdi, H I 2014. *Evaluasi Toksisitas Hematologi Akibat Penggunaan 6-Merkaptopurin Dalam Fase Pemeliharaan Pada Pasien Pediatri Kanker Leukimia Limfoblastik Akut Di RS Kanker Dharmais Jakarta*. Media Farmasi, 11(1), 90-97.

[4] Labati, R D, Piuri, V, dan Scotti, F 2011, "*All-IDB: The Acute Lymphoblastic Leukemia An image Database for An image Processing*", *Proceedings of the 18th IEEE ICIP International Conference on An image Processing*, Eds: Macq, B., dan Schelkens, P., IEEE Signal Processing Society, Brussels, 2045-2048.

[5] Mandyartha, E P, & Fatichah, C 2016. *Three-level Local Thresholding Berbasis Metode Otsu untuk Segmentasi Leukosit pada Citra Leukemia Limfoblastik Akut*. *Jurnal Buana Informatika*, *7*(1).

[6] Mohapatra, S, Patra, D2, & Satpathy, S 2014. An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic an images. Neural Computing and Applications, 24(7-8), 1887-1904.

[7] Nanni, L, Lumini, A, & Brahnam, S 2010. *Local binary patterns variants as texture descriptors for medical an image analysis*. Artificial Intelligence in Medicine , 49, 117–125.

[8] Oliva, D, Cuevas, E, Pajares, G, Zaldivar, D, & Perez-Cisneros, M 2013. *Multilevel thresholding segmentation based on harmony search optimization*. *Journal of Applied Mathematics*, *2013*.

[9] Putzu, L, Caocci, G, & Di Ruberto, C 2014. *Leucocyte classification for leukaemia detection using an image processing techniques*. *Artificial intelligence in medicine*, *62*(3), 179-191.

[10] Rezatofighi, S H, & Soltanian-Zadeh, H 2011. *Automatic recognition of five types of white blood cells in peripheral blood*. *Computerized Medical Imaging and Graphics*, *35*(4), 333-343.

[11] Salsabila, K 2016. *Perbandingan Nilai Ureum Dan Kreatinin Sebelum Dan Sesudah Kemoterapi Fase Konsolidasi Pada Pasien Leukemia Limfoblastik Akut Anak Di RSUP H. Adam Malik Medan* Tahun 2011-2013.

[12] Singhal, V, & Singh, P 2016. *Texture Features for the Detection of Acute Lymphoblastic Leukemia*. In *Proceedings of International Conference on ICT for Sustainable Development* (pp. 535-543). Springer Singapore.

[13] Widjajanto, P H, Veerman, A J P, & Sutaryo, S 2016. *Apoptotic cell identification: An in-vivo study during induction treatment of childhood acute lymphoblastic leukemia*. Paediatrica Indonesiana, 46(5), 195-8.
[14] Singhal, V, & Singh, P 2014. *Local Binary Pattern for automatic detection of Acute Lymphoblastic Leukemia*. In 2014 Twentieth National Conference on Communications (NCC) (pp. 1–5). IEEE. doi:10.1109/NCC.2014.6811261
[15] Thanh, T.T.P, Vununu, C., Atoev, S., Lee, H-S., Kwong, K-R 2018. *Leukemia blood cell image classification using convolutional neural network*. International Jpurnal of Computer Theory and Engineering,
[16] Shafique, S & Tehsin S 2018.*Computer-aided diagnosis of acutelLymphoblastic leukemia*. Computational and Mathematical Methods in Medicine
[17] Taleb, A S T A & Atiya A F. 2017. *A new approach for leukemia identification based on cepstral analysis and wavelet transform*. International Journal of Advanced Computer Science and Applications.

Acknowledgement

The study was sponsored by DRPM DIKTI Ministry of Research, Technology, and Higher Education of The Republic of Indonesia.