Segmentation System of Acute Myeloid Leukemia (AML) Subtypes on Microscopic Blood Smear Image

Nur Khomairoh
Departement of Informatics and Computer Engineering
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
nuurkhomairoh@gmail.com

Riyanto Sigit
Departement of Informatics and Computer Engineering
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
riyanto@pens.ac.id

Tri Harsono
Departement of Informatics and Computer Engineering
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
trison@pens.ac.id

Yetti Hernaningsih
Departement of Clinical Pathology
Faculty of Medicine, Airlangga University
Dr. Soetomo General Academic Hospital
Surabaya, Indonesia
yetti-h@fk.unair.ac.id

Anwar Anwar
Departement of Informatics and Computer Engineering
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
anwar@pasca.student.pens.ac.id

Abstract— **Leukemia is a blood cancer that attacks human white blood cells. This disease is divided into four types, including Acute Myeloid Leukemia (AML). AML is the most common type of acute leukemia, and it has eight types of subtypes distinguished by the level of cell maturation. Medical personnel determines the type of AML based on microscopic images of blood cell smears that contain white blood cells, red blood cells, and pieces of blood. This research builds a segmentation system that can determine the boundary of an object with the surrounding area, where the object sought is white blood cells contained in microscopic images of blood cell smears. White blood cells are sought based on ROI using the Haar Cascade Classifier, and then segmentation is carried out on the nucleus and cytoplasm. AML sub-types used as objects in this study are M4, M5, and M7. Based on the results of experimental data on the segmentation system, the nucleus segmentation in each cell of M4, M5, and M7 with an accuracy of 87.5%, 90.4%, 84.6% in sequence, and the results of cytoplasm segmentation are 75%, 71.4%, and 80.76%, respectively.**

Keywords—Acute Myeloid Leukemia, Haar – Cascade Classifier, segmentation, blood smear image

I. Introduction

Leukemia is a cancer cell that attacks human white blood cells. This type of disease is generally characterized by an excessive amount of production in white blood cells that are not yet fully matured in the bone marrow. French-American-British (FAB) divides leukemia into four different types based on its formation cell [1]. Of the four types, one of them is Acute Myeloid Leukemia (AML) has eight different subtypes based on where leukemia develops, namely M0, M1, M2, M3, M4, M5, M6, and M7. Among the eight subtypes, AML M4, M5, and M7 are affected by almost the same level of cell maturation, so further analysis is needed to be able to be distinguished [2].

A standard procedure to identify leukemia is to perform a Complete Blood Count (CBC), which counts the number of white blood cells and red blood cells through microscopic images of blood cells based on cell morphology [3]. However, this procedure is relatively dependent on the length of time, the operator's ability, and the fatigue factor [4].

In addition to the CBC procedure, there are other ways to be able to assist the diagnosis of AML, by utilizing the image processing method [5]. The speed and accuracy in the process of diagnosing disease are needed in connection with the accuracy of the data to determine the policy to be taken in efforts to prevent and cure the disease [6]. So that the application of image processing can help carry out analysis through the steps that will be carried out on an image.

The image of a blood cell smear used in the CBC process to identify leukemia contains white blood cells and other components like red blood cells and pieces of blood. In addition, the presence of white blood cells that touching each other can also affect the detection process. The process of segmentation in blood smear images is important to be able to separate white blood cells as objects, and others as non-objects (background). After the object detection process is done, it will be easier to do the separation if there are cells that touch. Segmentation is an important stage in recognizing medical objects. Segmentation aims to separate cells from the parts containing the nucleus and cytoplasm, besides using thresholding and color to separate objects with K-Means [7].

This study discusses the use of the segmentation process in the image of blood smear cells. The results of this study are expected that the method applied can separate white blood cells and other cells, and minimize the presence of cells that touch.

II. Related research

In the process of recognizing a type of disease, a system that has received an image will do a preprocessing to perform processing with the median blur method and take advantage of changing the RGB image to the HSV model [6]. Utilization can help adjust the character of each image. In other learning cases, the use of CIELab colors is used because it reduces memory usage, reduces interpersonal perception, and has two color components (a and b) designed to approach human vision. While the L component in CIELab can be used to adjust the contrast and brightness so that it is under the human perception of the light seen [4]. Image improvement can also be used to improve the process of diagnosing acute myeloid leukemia [9].

Some studies apply ROI to determine the location of white blood cell objects [8]. The application of ROI is able to minimize errors during segmentation because the pixel size of the object has changed, and the most significant color variation can be seen on the object.

Segmentation is an important method in the identification stage of disease images [7]. Segmentation aims to separate objects from the background or other cells. Besides, this research needs to do a different segmentation of the cytoplasm and nucleus of blood cells. Using the K-Means Clustering method can be done to separate the color clusters of objects into K clusters [8].

III. METHODOLOGY

A. *Images Preparation*

The data of AML M4, M5, and M7 blood smear images in this study were obtained from Dr. Soetomo Surabaya. The preparations used for testing come from test samples with a range between 2015 – 2020. Blood cell images were obtained through a microscope that was equipped with a camera. Observation on a microscope is done with a magnification of 1000x. The results will be displayed on a computer screen and saved in *.jpg format with dimensions of 680 x 512 pixels.

B. *Region of Interest Images*

Before conducting data training on the Haar - Cascade Classifier method, it is necessary to know about the techniques. This method requires quite a lot of positive images (images containing objects) and negative images (images other than objects). After that, the feature is extracted from the data sample. This method is likened to a convolutional kernel. Each feature that has been extracted is compared to a single value obtained by measuring the number of pixels [10]. This amount of data and pixel size are very much the results of training. This study uses positive images with a size of 24x24 pixels and negative images with a size of 255x190 pixels with a number of 300 and 700 objects, respectively. Keep in mind if the minimum need between positive and negative files is 1: 2. Fig.1 shows the pattern of Haar – Cascade Classifier.

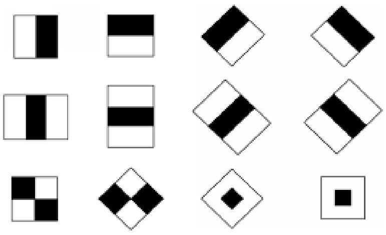


Fig. 1. The Haar – Cascade Pattern.

$$F\ (Haar) = \Sigma F\text{black} - \Sigma F\text{white} \quad (1)$$

Description :

$F\ (Haar)$ = total feature value.

ΣFblack = sum of dark pixel.

ΣFwhite = sum of white pixel.

The positive object image is obtained through the cropping results of red blood cell objects in the image to be tested and then rotated by 90°, 180°, and 270° to get variations of features. At the same time, negative objects come from cells other than objects, namely red blood cells or background. After the object is determined, a pos.dat file is then created containing positive images and neg.txt containing negative images. Next, a sample *.vec is made through a command prompt that contains positive images, and finally, a number of stages of training are required. More details can be seen in Fig. 2 below.

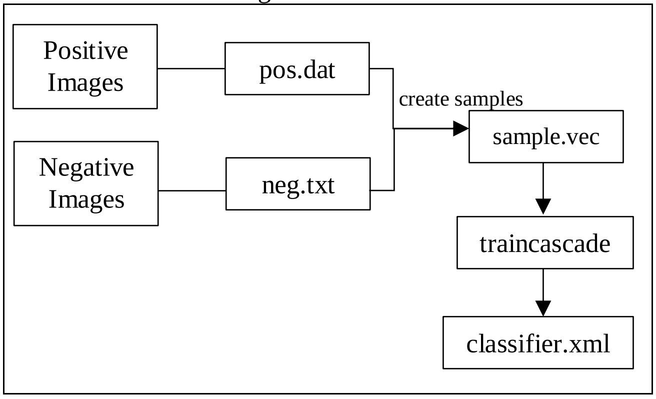


Fig. 2. The Haar – Cascade Classifier Training Process.

The result of the training process is the classifier.xml file. This file will then be used for the ROI search process in this method. The output of this stage is the appearance of a rectangular box on the object that has been detected, as shown in Fig. 2. The box used as a boundary for making bounding boxes is between 80 pixels and 120 pixels wide. If the object is between these ranges, the bounding box will be displayed.

C. *Preprocessing*

This study uses two different methods at the preprocessing stage. The use of this different method is because the detected object, the cytoplasm, and the cell nucleus have different color brightness variations. The image used at this stage comes from the results of cropping the previous process. Preprocessing in the nucleus begins with the color conversion from RGB to HSV, and the S channel is used. This color conversion can reduce the enchanting effect found in the nucleus, and the color in the S channel tends to be easily recognized by the human eye. The preprocessing stage in the nucleus can be seen in Fig. 3 and the results of this process can be seen in Fig. 4. To do this color conversion can be translated into equations :

$$V \leftarrow \max(R,G,B) \quad (2)$$

$$S \leftarrow \begin{cases} \dfrac{V - \min(R,G,B)}{v} & if\ V \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$H \leftarrow \begin{cases} 60(G-B)/(V-\min(R,G,B) & \text{if V = R} \\ 120 + 60(B-R)/(V-\min(R,G,B)) & if\ V = G \\ 240 + 60(R-G)/(V-\min(R,G,B)) & if\ V = B \end{cases} \quad (4)$$

if H < 0 then H ← H + 360. On output $0 \leq V \leq 1$, $0 \leq S \leq 1$, $0 \leq H \leq 360$. (5)

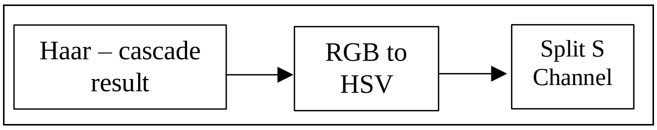


Fig. 3. Preprocessing of the nucleus.

Below is the image resulting from the conversion of RGB to HSV color with S Channel.

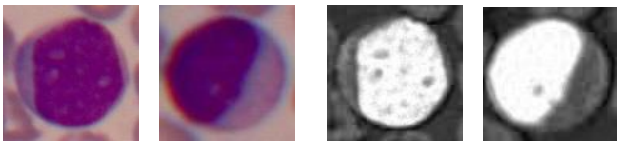

(a) Input Image (b) S Channel of HSV

Fig. 4. The HSV S Channel.

Whereas preprocessing in the cytoplasm begins with the color conversion from RGB to CIELab. The use of these different color elements is because the cytoplasm has a fainter color variation than the nucleus and tends to resemble red blood cells. The color conversion equation can be seen as follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \leftarrow \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} . \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (5)$$

$$X \leftarrow X / Xn, \text{ where } Xn = 0.950456 \quad (6)$$

$$Z \leftarrow Z/Zn, \text{ where } Zn = 1.088754 \quad (7)$$

$$L \leftarrow \begin{cases} 116 * Y^{1/3} - 16 & for\ Y > 0.008856 \\ 903.3 * Y & for\ Y \leq 0.008856 \end{cases} \quad (8)$$

$$a \leftarrow 500(f(X) - f(Y)) + delta \quad (9)$$

$$b \leftarrow 200(f(Y) - f(Z)) + delta \quad (10)$$

$$f(t) = \begin{cases} t^{1/3} & for\ t > 0.008856 \\ 7.787t + 16/116 & for\ t \leq 0.008856 \end{cases} \quad (11)$$

$$delta = \begin{cases} 128 & for\ 8 - bit\ images \\ 0 & for\ floating - point\ images \end{cases} \quad (12)$$

In this cytoplasmic preprocessing, channel * b is used in CIELab because this colour represents blue where the colour elements are similar to the cytoplasm. Image Fig. 5 below shows the stages of preprocessing in the cytoplasm.

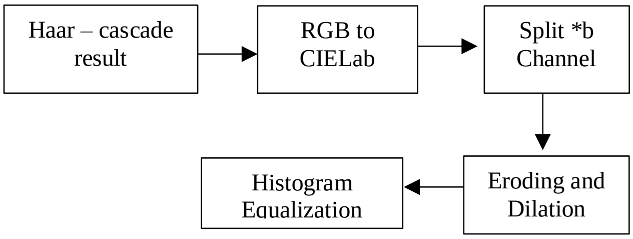


Fig. 5. Preprocessing of the cytoplasm.

Erosion and dilation morphology are needed to calculate the minimum location around the kernel area. The histogram equalization itself is used to increase contrast by expanding the intensity range of an image so that the lighting is evenly distributed on each object. Fig. 6 shows the result of histogram equalization.

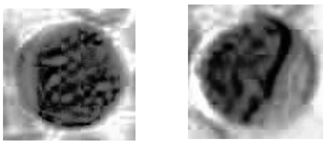


Fig. 6. Histogram Equalization Result.

D. Segmentation

Like the previous process, image processing between the nucleus and cytoplasm at the segmentation stage also uses a different method. At the stage of nucleation segmentation, the K – Means Clustering method distinguishes objects into three different clusters. In comparison, the segmentation of the cytoplasm uses otsu thresholding. For more details, can be seen in Fig. 7 of the following:

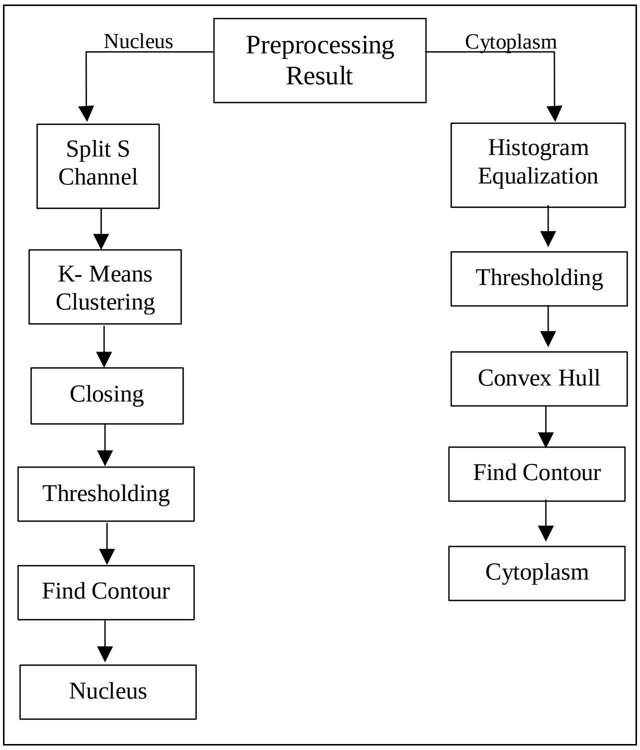


Fig. 7. Segmentation of Nucleus and Cytoplasm

The use of K-Means clustering on the object nucleus is to group object color data into three different classes. In this way, it can minimize objects that have a different color from the nucleus. Several distance spaces can be implemented to calculate the distance between centroids, one of which uses Euclidean Distance. The distance calculation on the Euclidean distance is often used because it can calculate the shortest distance between two points. The equation can be seen as:

$$\boldsymbol{dist(x, y)} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (13)$$

Where :
dist (x, y) = euclidean distance
n = amount of data
i = iterations

Whereas in the cytoplasm, the equalization histogram is used to increase the contrast of the preprocessing results, so the cytoplasmic intensity had the same amount. The use of closing morphology on the nucleus is used to close parts of the object that are still open. Thresholding is used to find the degree of gray and make the background and images that are not objects black. Convex Hull in the cytoplasm helps attach objects that are not completely closed. Next, find contour is used to find the contour of objects, both nucleus and cytoplasm. From this separate process, we get image segmentation results from the nucleus and cytoplasm in cropping images.

IV. RESULT AND DISCUSSION

A. *Image Preparation*

In this study, we obtained 300 positive data samples and 700 negative data from the results of cropping each trial image. Previously, there was an increase in the amount of data, but an error occurred in memory during the training process, so the amount of data had to be reduced. Positive sample data uses the minimum dimension value, which also influences the training process. If the dimensions of the image are too large, it can cause errors when creating vector samples. The image in Fig. 8 shows the sample data used in this study.

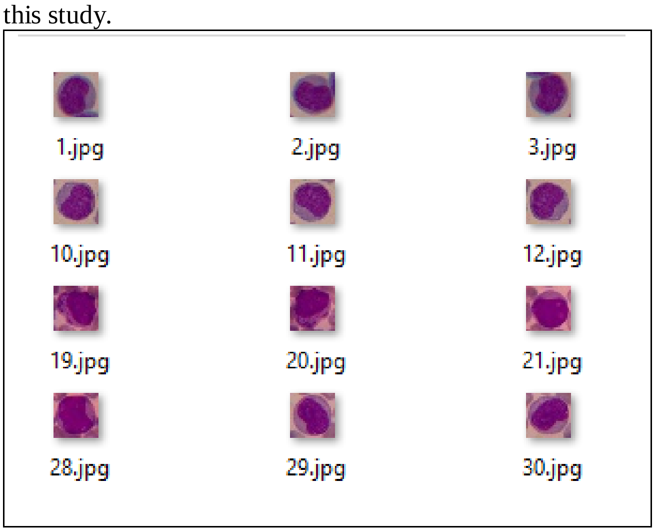
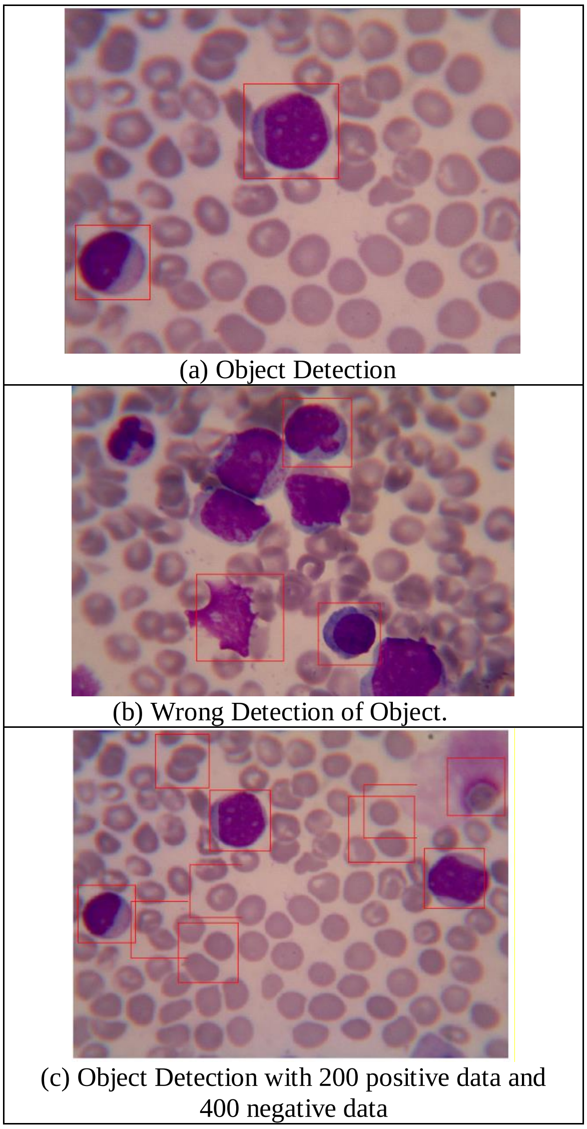


Fig. 8. Sample of Image Object.

B. *Cropping Region of Interest Images*

The training method for the Haar - Cascade Classifier is very dependent on the amount of data to be trained. The more training data, the better the results will be shown. In this study, 300 positive data and 700 negative data were used. Although the amount of data is not large enough, training data can still be done. Previously, training was also conducted with 200 positive data and 400 negative data. The test results from this training did not fully recognize the object. There are several images that experience object recognition errors during the testing process, as shown in Fig. 9.



(a) Object Detection

(b) Wrong Detection of Object.

(c) Object Detection with 200 positive data and 400 negative data

Fig. 9. Bounding Box of Object Detection.

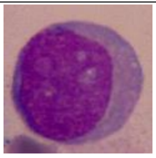
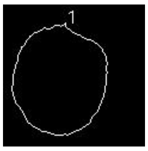
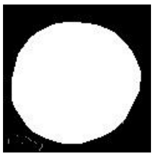
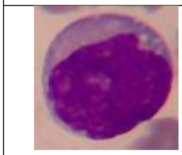
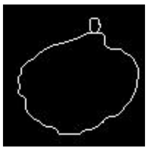
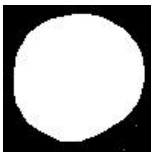
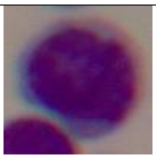
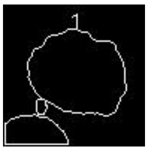
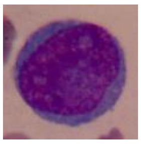
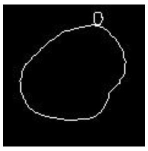
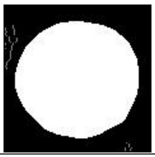
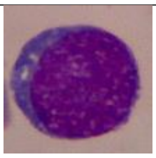
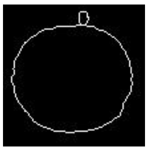
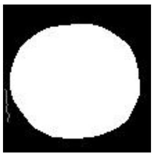
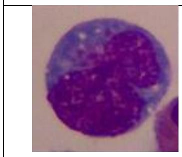
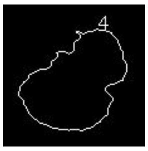
C. *Segmentation*

The calculation of the percentage of objects that have been segmented can be seen in Table I. While the results of the segmentation of each object can be seen in Table II. Table II. shows the segmentation process is quite reasonable according to the results of the input provided. There are even objects that touch each other and can be appropriately segmented.

TABLE I. DATA OF CORRECT AND INCORRECT SEGEMENTATION

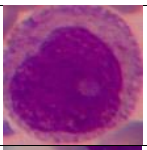
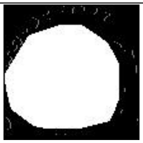
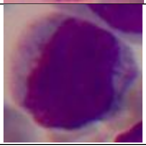
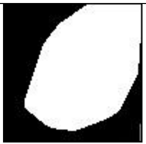
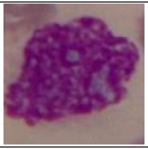
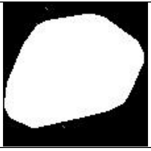
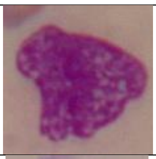
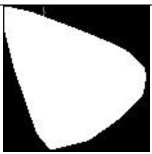
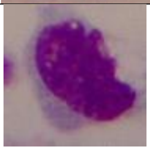
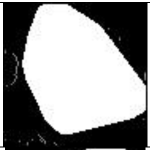
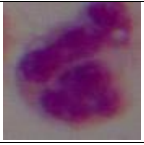
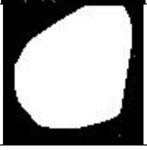
Object	Subtype Cells		
	M4	M5	M7
Nucleus Segmentation	21	19	22
Cytoplasm Segmentation	18	15	21
Total Image Cells	**24**	**21**	**26**
Percentage of Nucelus	**87.5%**	**90.4%**	**84.6%**
Percentage of Cytoplasm	**75%**	**71.4%**	**80.76%**

TABLE II. DETAILS OF OBEJECT SEGMENTATION

Type	Input and Output Image		
	Input	*Nucleus*	*Cytoplasm*
M4			
			
M5			
			
M7			
			

However, not all objects are segmented successfully. Because of the use of a convex hull that can make lines or connect objects nearby, the image that should be separated instead becomes one, as shown in Table III.

TABLE III. RESULT OF FALSE CYTOPLASM SEGMENTATION

Type	Input and Output Image	
	Input	*Segmentation*
M4		
		
M5		
		
M7		
		

CONCLUSION

The Haar - Cascade Classifier method can be used in the process of detecting an object, including leukemia blood smear images. The amount of data is very influential in the process as can be seen in this study which shows the results of the nucleus segmentation of M4, M5, and M7 respectively 87.5%, 90.4%, 84.6% and the results of cytoplasm segmentation 75%, 71.4%, 80.76% respectively. The more amount of data trained, the better the system can recognize objects.

Determination of object ROI can minimize errors during segmentation because the pixel size of the object has changed, and the largest color variation can be seen on the object. The use of two different methods in the learning segmentation process is needed because the nucleus and cytoplasm of cells have different levels of color brightness. Cytoplasm itself tends to have a faded color and almost resembles a background. The method proposed above can be used to minimize errors in distinguishing between cytoplasm and background.

REFERENCES

[1] American Cancer Society, “Colorectal Cancer Facts & Figures 20142016”. Atlanta, Ga: American Cancer Society, 2014

[2] Dacie and Lewis, “Practical Haematology, Eleventh Edition”, Elsevier Churcill Livingstone, London, ISBN-13: 9780702034084, 2011.

[3] E. Suryani, U. Salamah, and A. A. Wijaya, “Identifikasi Penyakit Acute Myeloid Leukemia (AML) Menggunakan ‘Rule Based System’ Berdasarkan Morfologi Sel Darah Putih Studi Kasus○6Π: AML2 dan AML4,” Seminar Nasional Teknologi Informasi, dan Komunikasi Terapan 2014, pp. 193–199, 2014.

[4] S. Agaian, M. Madhukar and A. T. Chronopoulos, "Automated Screening System for Acute Myelogenous Leukemia Detection in Blood Microscopic Images," in IEEE Systems Journal, vol. 8, no. 3, pp. 995-1004, Sept. 2014.

[5] E. S. Wiharto, S. Palgunadi and Y. R. Putra, "Cells identification of acute myeloid leukemia AML M0 and AML M1 using K-nearest neighbour based on morphological images," *2017 International Conference on Data and Software Engineering (ICoDSE)*, Palembang, 2017, pp. 1-6.

[6] R. Sigit, M. M. Bachtiar and M. I. Fikri, "Identification Of Leukemia Diseases Based On Microscopic Human Blood Cells Using Image Processing," 2018 International Conference on Applied Engineering (ICAE), Batam, 2018, pp. 1-5

[7] Setiawan, A. Harjoko, T. Ratnaningsih, E. Suryani, Wiharto and S. Palgunadi, "Classification of cell types in Acute Myeloid Leukemia (AML) of M4, M5 and M7 subtypes with support vector machine classifier," 2018.

[8] Harto, Aryo & Fatichah, Chastine. “Segmentasi Dan Pemisahan Sel Darah Putih Bersentuhan Menggunakan K-Means Dan Hierarchical Clustering Analysis Pada Citra Leukemia Myeloid Akut,” 2017, JUTI: Jurnal Ilmiah Teknologi Informasi. 15. 162

[9]] N. R. Mokhtar, N. H. Harun, M. Y. Mashor, H. Roseline, N. Mustafa, and R. Adollah, “Image Enhancement Techniques Using Local, Global, Bright, Dark, and Partial Contrast Stretching For Acute Leukemia Images,” Engineering, vol. I, pp. 3–8, 2009.

[10] P. Viola, M. Jones, D. Snow, “Detecting pedestrians using patterns of motion and appearance”, Int. J. Comput. Vis., 63 (2), pp. 153–161, 2005.