Cells Identification of Acute Myeloid Leukemia AML M0 and AML M1 using k-Nearest Neighbour Based on Morphological Images

Esti Suryani, Wiharto, Sarngadi Palgunadi, Yudha Rizki Putra
Informatics Department
Universitas Sebelas Maret
Surakarta, Indonesia
estisuryani@staff.uns.ac.id, wiharto@staff.uns.ac.id, palgunadi@staff.uns.ac.id, yr.putra@student.uns.ac.id

Abstract - Acute Myeloid Leukemia (AML) is a type of leukemia characterised by the occurrence of myeloid series cell differentiation that stops in the blast cells causing the accumulation of blast cells in the bone marrow. This study aims to determine leukemia typically in AML M0 and AML M1 based on the morphology of white blood cell image using image processing method. The steps performed are median filtering, YCbCr colour conversion, thresholding, and opening, and k-Nearest Neighbors classifier to classify cell types from feature extraction results. The result of characteristic extraction was done by mean difference test for each characteristic between cell type indicated that there was a significant difference in WBC diameter characteristic between cell type, while on a characteristic of nucleus ratio showed that there was no significant difference. Based on characteristic testing of each cell, a combination of a characteristic of WBC diameter and nucleus roundabout obtained the highest accuracy when k = 5 and k = 7 is 67,28%. Thus the characteristic of WBC diameter and the nuclear roundabout is the most influential data classification feature. Based on the test results of each cell, if the algorithm k = 6 k-Nearest Neighbors can classify the cell correctly 59.87% of the 162 data used based on the three characteristics each cell is the WBC diameter, the nucleus roundabout and the nucleus ratio.

Keywords-Acute Myeloid Leukemia, Image Processing, k-Nearest Neighbours.

I. INTRODUCTION

Leukemia is a disease of the bone marrow that causes the production of immature cells or excessive blast in normal bone marrow [1]. AML is a type of leukemia caused by maturity blockade that causes the process of differentiation of myeloid series cells stalled in blast cells that resulted in the accumulation of blast in the bone marrow resulting in bone marrow failure [1]. Based on the French-American-British (FAB) classification, AML is classified into eight types: M0, M1, M2, M3, M4, M5, M6, M7 [2].

Diagnosis of AML can be made by observing it by a microscope, or by a *flowcytometric* device, which is expensive and time-consuming [2] [3] [4]. Another way to help the process of AML diagnosis is to utilise image processing and data mining techniques.

A lot of research has been done so far on leukemia identification with image processing, among others, successfully identify ALL and AML M3 with Fuzzy Rule-Based System Sugeno method, with accuracy 73.68% [5]. Further research was done by identification of AML M2 and M4 with an accuracy of 81,67% by using Fuzzy Rule-Based System Sugeno [6]. Also, many studies have been conducted on the identification of leukemia with image processing and data mining. Among other things is the use of backpropagation momentum method to classify each cell type in leukemia type AML M2 or AMLM3 with the result of accuracy 94.298%, and can be used to classify each image into the type of AML M2 or AML M3 with 75% accuracy [4]. Another method used is the k-NN Method. This k-NN-method has been successfully used to classify blast cells in detecting acute leukemia, based on image processing [7][8]. The k-Nearest Neighbor (k-NN) algorithm is one of the classification techniques in data mining, where the training data set is stored to find new data classifications by comparing with the most similar data in training data [9]. The present study aims to classify the type of Acute Leukemia type AML M0 and AML M1 using k-NN, based on the morphology of white blood cell images.

II. MATERIALS AND METHODS

A. Materials

Data Image data was taken from white blood cell preparation from Bone Marrow Processing (BMP) using the microscope in Clinical Pathology Laboratory RSUD Dr Moewardi Surakarta. Each image used is observed by an expert using a digital microscope with 1000 times magnification. All images were saved in JPG format with size 1600 x 1200 pixels. The image data used are 50 white blood cell images consisted of 20 identified images as of AML M0

978-1-5386-1449-5/17/$31.00 ©2017 IEEE

and 30 identified as of AML M1 images. The AML M0 (Acute Myeloblastic Leukemia, minimally differentiated) is characterised by virtually no mature myeloid cells. In AML M0 is, however, dominated by myeloblast cells, very little promyelocyte [10]. Meanwhile, AML M1 (Acute Myeloblastic Leukemia with minimal maturation) is characterised by the existing of myeloblast of approximately 75%, and less than 15% consists of promyelocyte and myelocyte [11]. The characteristics of each cell can be shown in Table I.

TABLE I. CELL CHARACTERISTICS

Cell Types	Features		
	WBC Diameter (µm)	*Nucleus Ratio (%)*	*Nucleus Roundness*
Myeloblast	12-20	80-85	Round
Promyelocyte	15-25	75-85	Round /Oval
Myelocyte	10-20	50-65	Round/ Oval

B. Methods

The research steps are as follows. Image Acquisitions depicted in Fig.1 are stored in JPG format with size 1600 x 1200 pixels. The images then are preprocessing with a median filter, and its RGB features are converted to YCbCr. The next steps are thresholding and opening segmentation, feature extraction against WBC area, calculating the WBC diameter and Nucleus Ratio, then clarifying the features with k-NN. Furthermore, cell type identification is included in AML M0 or AML M1. Detailed research steps can be shown in Fig 2, and Fig. 3, then the identification rule used can be shown in Table II.

TABLE II. IDENTIFICATION RULES

No	Myeloblast	Promyelocyte	Myelocyte	Identification Result
1	Yes	Yes	Yes	AML M1
2	Yes	Yes	No	AML M0
3	Yes	No	Yes	Not Identified
4	Yes	No	No	Not Identified
5	No	Yes	Yes	Not Identified
6	No	Yes	No	Not Identified
7	No	No	Yes	Not Identified
8	No	No	No	Not Identified

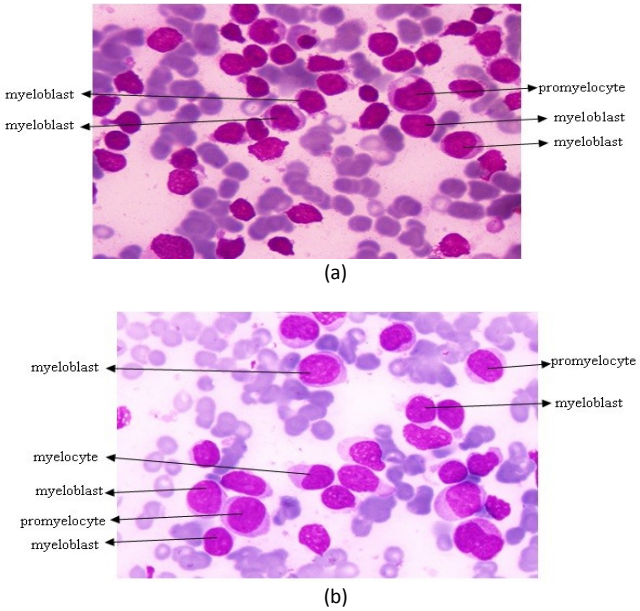


Fig. 1. (a) Image of WBC of AML M0 (b) Image of WBC of AML M1

Detailed research steps can be shown in Fig. 2, and Fig. 3:

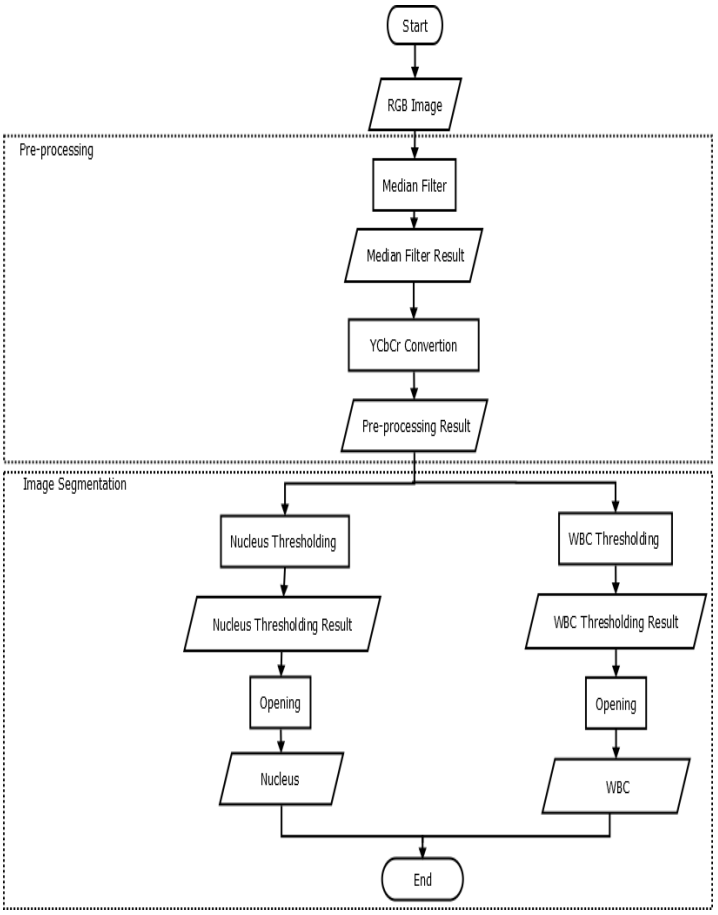


Fig. 2. Preprocessing and Image Segmentation

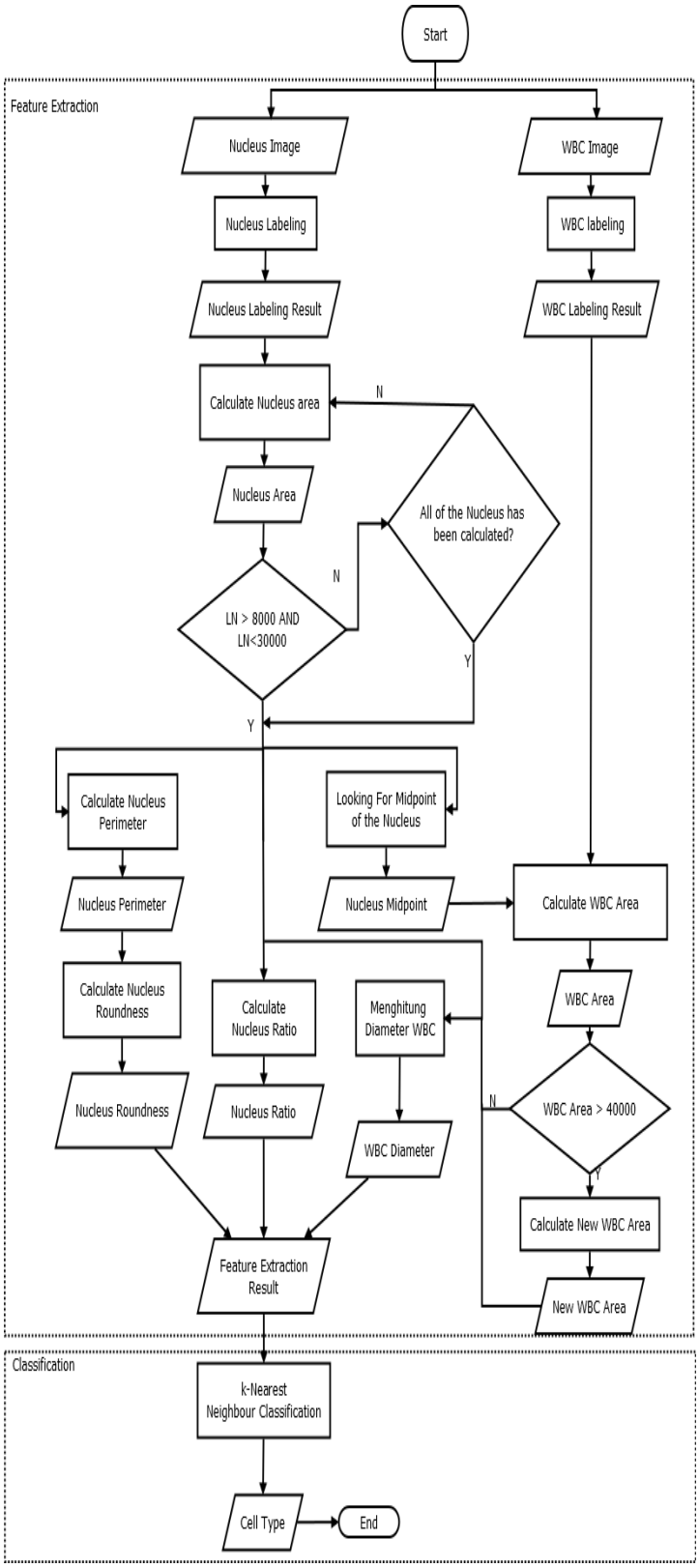


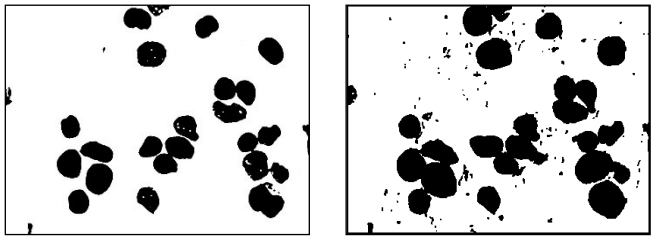
Fig. 3. Image Segmentation & Classification

Image acquisition is a shooting used in this study, observed by experts using a digital microscope with 1000 times magnification. All images are saved in JPG format with size 1600 x 1200 pixels shown Fig.1. Pre-processing is done to reduce noise with Median Filter without reducing important information from the image [12]. RGB colour conversion to YCbCr is then performed, use (1) [13]:

$$Y = 0.299\,R + 0.587\,G + 0.114\,B$$
$$Cb = -0.1687\,R - 0.3313\,G + 0.5\,B + 128$$
$$Cr = 0.5\,R - 0.4187\,G - 0.0813\,B + 128 \quad (1)$$

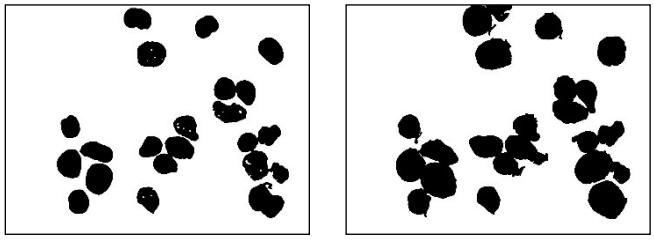
The colour representation of the YCbCr component can be used to overcome the lighting problem on the microscopic image of the blood so that it will provide the shape of a clear nucleus and WBC [12].

Image Segmentation, aiming to separate WBC from other blood components and background, is done by component-based thresholding on YCbCr to obtain the nucleus and WBC objects as shown in Fig. 4. The result of this thresholding produces a binary image with noise in the form of small black patches. To eliminate the black spots morphology opening operation is performed, the results as shown Fig.5.



(a) (b)

Fig. 4. Thresholding Result (a) nucleus, (b) WBC



(c) (d)

Fig. 5. Opening Result (a) nucleus, (b) WBC

Furthermore, the characteristic extraction process of the morphology of white blood cell image was performed. The characteristics to be extracted are WBC diameter, nuclear to cytoplasmic ratio, and nuclear roundabout. Each segmented cell will be searched for its characteristics, then matched to the cell type of the observed result with the expert. The result of characteristic extraction which can be known cell type according to the expert identification will be used as training data for cell classification. The results of the segmentation of the nucleus and WBC, as shown in Fig.6.

a) nucleus b) WBC c) Segmentasion result

Fig. 6. Nucleus dan WBC segmented

The next stage is classification, aiming to determine the cell type using the k-Nearest Neighbour algorithm. In this study, the cells will be classified into three types of cells, namely myeloblast, promyelocyte and myelocyte. Training data obtained from characteristic extraction results will be used for classification using k-NN. Attributes used are WBC diameter, nuclear ratio, and nuclear roundabout. The closest spacing or similarity of each attribute test are compared to the training attribute, at the classification stage of the k-NN algorithm is calculated using Euclidean Distance, use (2) [7].

.

$$d_{Euclidean}(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad (2)$$

To avoid outweighing of their components, in the WBC diameter attribute, min-max normalisation is performed (3), to obtain the value is in the range of 0 to 1 [7]. Normalisation is done to obtain the same range of values with the attributes of nucleus and nucleus ratios. Furthermore, it is tested by using leave-one-out cross-validation (LOOCV) to obtain the k value with the best accuracy. The experimental value of k starts from 1 to 10. The k values obtained will be used in the determination of cell types using k-NN, which will be searched for the majority of cell types from a number k which has the closest resemblance.

$$v' = \frac{v - min_A}{max_A - min_A} \quad (3)$$

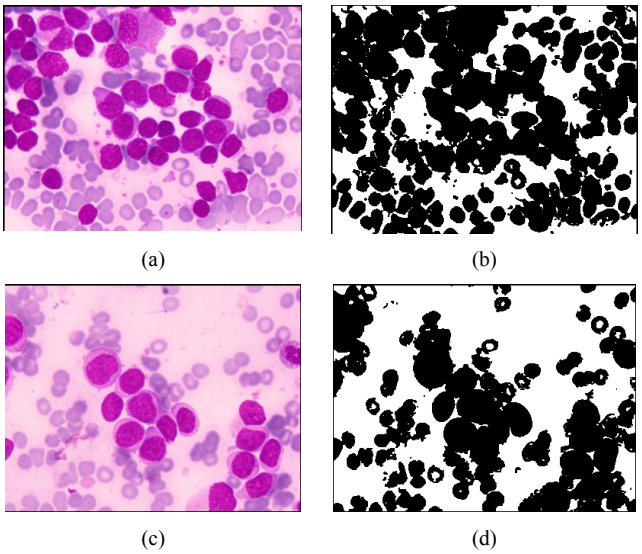
The next stage is the identification aims to determine the type of AML based on the cell content detected in one image. An image will be identified as AML M0 if there exist myeloblast cells and promyelocyte and will be identified as AML M1 if there exists myeloblast cells, promyelocyte and myelocyte. Rules for the process of identification of AML types have been shown in Table II.

The next stage is the analysis and testing of the results obtained. The analysis was performed based on the result of testing the extraction data characteristic and the image data used. At this stage of the analysis will be done three stages, namely the analysis of testing different means, testing each cell type, and testing each image. The different mean test is conducted to examine the level of difference of each characteristic between cell types. Testing each cell and testing each image are done to determine the level of system accuracy.

III. EXPERIMENTAL RESULT AND DISCUSSION

The output of the classification stage is three cell types, i.e. myeloblast, promyelocyte and myelocyte, while the output stage identification is type AML M0 or AML M1 based on detected cell content.

In the image segmentation stage used thresholding technique. The results obtained are not all WBC objects in the image well segmented, while the nucleus objects can be well segmented. WBC segmentation can not be done with the maximum because each image has a different colour intensity, the presence of cells that overlap. An example of a comparison of WBC segmentation results in images with different colour intensities can be shown in Fig. 7. The poor WBC segmentation results strongly impact the character extraction process as it will provide an excessive value of the WBC area. An excessive WBC range causes a change in the value of the WBC diameter and nuclear to cytoplasmic ratios. The feature extraction results yielded 165 data consisting of 97 myeloblasts, 31 promyelocytes, and 37 myelocytes. The test results of different mean and standard deviation can be shown in Table III, Table IV, and Table V.



(a) (b)

(c) (d)

Fig. 7. (a) Image AML_ M0_20 (b) WBC Images AML_M0_20 segmented (c) Image AML_M1_21 (d) WBC Image aml_M1_21 Segmented

A. *Different Mean Test*

The mean difference test is done by using T-Test method to get the value of significance (p-value) each characteristic between cell types. Different mean tests were performed to determine the relationship of each feature (WBC diameter, nuclear ratio, and nuclear roundabout) between cell types (myeloblast, promyelocyte, and myelocyte). The test was performed to find out the mean and standard deviation of WBC diameter, the ratio of the nucleus. The confidence level is 95% so that the error rate (α) is 5% (0.05). Table III shows the results of Different Mean Test between myeloblast and

promyelocyte. Table IV shows the mean difference test results between myeloblast and myelocyte. Table V shows the mean difference test results between promyelocyte and myelocyte.

TABLE III. MEAN DIFFERENCE TEST RESULT OF MYELOBLAST-PROMYELOCYTE

Cell Features	Mean ± Deviation Standart of		p-value
	Myeloblast	*Promyelocyte*	
WBC Diameter	16.700 ± 2.096	18.843 ± 3.161	0.000
Nucleus Ratio	0.677 ± 0.076	0.659 ± 0.075	0.248
Nucleus Roundness	0.579 ± 0.066	0.524 ± 0.108	0.001

TABLE IVV. MEAN DIFFERENCE TEST RESULT OF MYELOBLAST-MYELOCYTE

Cell Features	Mean ± Deviation Standart of		p-value
	Myeloblast	*Myelocyte*	
WBC Diameter	16.700 ± 2.096	15.660 ± 2.098	0.011
Nucleus Ratio	0.677 ± 0.076	0.677 ± 0.079	0.996
Nucleus Roundness	0.579 ± 0.066	0.575 ± 0.068	0.716

TABLE V. MEAN DIFFERENCE TEST RESULT OF PROMYELOCYTE-MYELOCYTE

Cell Features	Mean ± Deviation Standart of		p-value
	Promyelocyte	*Myelocyte*	
WBC Diameter	18.843 ± 3.161	15.660 ± 2.098	0.000
Nucleus Ratio	0.659 ± 0.075	0.677 ± 0.079	0.338
Nucleus Roundness	0.524 ± 0.108	0.575 ± 0.068	0.022

Based on Table III, it can be shown that the characteristics of WBC diameter and nuclear roundabout between myeloblast and promyelocyte are significant differences because the p-value is less than $\alpha = 0.05$. While the ratio of nuclei between a myeloblast and promyelocyte there is no significant difference because of the value of p-value more than 0.05.

Table IV shows that between myeloblast and myelocyte, only characteristic of WBC diameter has a significant difference with p-value value of 0.011. While the characteristic ratio of nucleus and nuclear roundabout there is no significant difference, seen from p-value 0.996 and 0.716 are both greater than 0.05

Based on Table V, it can be shown that Promyelocyte and Myelocyte have differences in WBC diameter and nuclear roundabout characteristics because the p-value is less than 0.05. The characteristics of the nucleus ratio do not have a significant difference since the p-value 0.388 is greater than 0.05.

B. Cross-Validation Test

The characteristic extraction data was used as training data by firstly normalising the values on the WBC diameter characteristics. Then tested by using LOOCV on a k-NN algorithm with experimental values k ranging from 1 to 10. The test results show the best accuracy is obtained when the value of k = 6 is 59.87%. Therefore the value of k = 6 is selected for the testing process on each image. The experimental results for determining the k value with cross-validation as shown in Fig. 8.

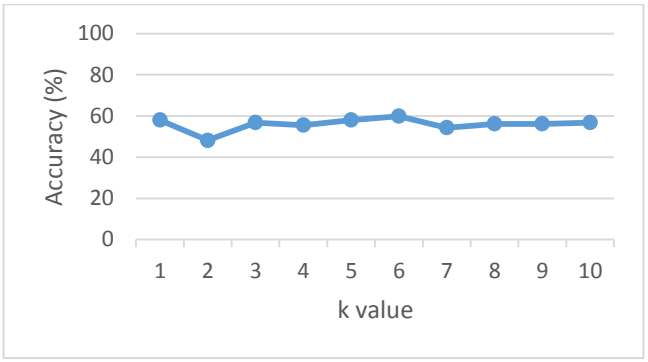


Fig. 8. Data Test Result with *Cross-Validation*

Out of the 162 data tested with k = 6, there are 97 data (59.87%) that can be correctly classified consisting of 76 myeloblast cells, eight promyelocytes and 13 myelocytes. The full results can be shown in Table VI.

TABLE VI. DATA TEST RESULT (K=6)

Actual Class	Predicted Class		
	Myeloblast	*Promyelocyte*	*Myelocyte*
Myeloblast	76	9	12
Promyelocyte	18	8	2
Myelocyte	22	2	13

Tests were also conducted on six other combinations of 3 features used (WBC Diameter, Nucleus Ratio, and Nucleus Roundness) to identify characteristics that greatly affect the accuracy of the data classification test. The test was performed using LOOCV with experimental values of k = 1 to k = 10. Full test results are shown in Table VII.

TABLE VII. TEST RESULT OF FEATURES COMBINATION

k	Accuration of feature combination (%)						
	A	B	C	A – B	A – C	B – C	A - B – C
1	43,209	47,530	51,851	58,024	63,580	53,703	58,024
2	41,358	39,506	43,209	51,234	54,321	37,037	48,148
3	51,234	51,851	54,321	54,321	64,814	46,913	56,790
4	53,703	51,234	50,617	51,851	64,814	45,679	55,555
5	52,469	48,765	48,148	51,851	67,284	41,358	58,024
6	54,938	48,765	48,148	53,086	64,197	47,530	59,876
7	54,321	50,617	51,234	53,707	67,284	50,617	54,321
8	49,382	51,851	52,469	54,321	62,963	47,530	56,172
9	53,703	51,234	57,407	52,469	65,432	50,617	56,172
10	55,555	52,469	54,938	50	64,197	54,321	56,790

Note on Table VII:
A: WBC Diameter, B: Nucleus Ratio. C: Nucleus Roundness

Tests were also conducted on six other combinations of 3 features used (WBC Diameter, Nucleus Ratio, and Nucleus Roundness) to identify characteristics that greatly affect the accuracy of the data classification test. The test was performed using LOOCV with experimental values of k = 1 to k = 10. The complete test results are shown in Table VII. Based on the results of the tests in Table VII, it can be seen that the combination of features between the WBC diameter and the nuclear roundabout has higher accuracy than other combinations of traits. The highest accuracy results of the combination test of WBC diameter and nuclear roundabout characteristics occur when the k = 5 and k = 7 values are 67.284%. While the combination of characteristics between WBC diameter, nucleus ratio and roundabout k = 6 is 59.27%. Thus the characteristics of WBC diameter and nuclear roundabout are the most influential features of data classification.

IV. CONCLUSION

The result of characteristic extraction was done by mean difference test for each characteristic between cell types indicated that there was a significant difference in WBC diameter characteristic between cell types, whereas on a characteristic of nucleus ratio showed that there was no significant difference. Based on characteristic testing of each cell, a combination of a characteristic of WBC diameter and nucleus roundabout obtained the highest accuracy when k = 5 and k = 7 is 67,28%. Thus the characteristics of WBC diameter and nuclear roundabout are the most influential features of data classification. Based on the test results of each cell, if the algorithm k = 6 k-Nearest Neighbors can classify the cell correctly 59.87% of the 162 data used based on the three characteristics of each cell. I.e. the WBC diameter, nucleus roundabout, and the ratio of the nucleus.

ACKNOWLEDGMENT

This research was supported by Grant PNBP 2017 Universitas Sebelas Maret.

REFERENCES

[1] Ciesla, B., *Hematology in Practice*. 2007, Philadelphia: F.A. Davis.

[2] Theml, H., Diem, H. and Haferlach, T., *Color Atlas of Hematology*. Second Revised ed. 2004, New York: Thieme.

[3] Nazlibilek, S., et al., *Automatic Segmentation, Counting, Size Determination and Classification of White Blood Cells.* Measurement, 2014(55): p. 58-65.

[4] Suryani, E., Wiharto, Palgunadi, S., Pradana T.P.N., 2016. *Classification of Acute Myelogenous Leukemia (AML M2 and AML M3) using Momentum Back Propagation from Watershed Distance Transform Segmented Images.* International Conference on Computing and Applied Informatics 2016. IOP Conf. Series: Journal of Physics: Conf. Series 801 (2017) 012044 doi:10.1088/1742-6596/801/1/012044.

[5] Suryani, E., Wiharto, and Polvonov, N., (2013). Identifikasi Penyakit Acute Lymphoblastic Leukemia (ALL) Menggunakan Fuzzy Rule Based System Berdasarkan Morfologi Sel Darah Putih. Prosiding Seminar Nasional Teknologi Informasi dan Komunikasi Terapan (SEMANTIK), Udinus, Semarang, ISBN :979-26-0266-6.

[6] Suryani, E., Salamah, U and Wijaya,A.A., *Identifikasi Penyakit Acute Myeloid Leukemia (AML) Menggunakan Rule Based System Berdasarkan Morfologi Sel Darah Putih*, in *Seminar Nasional Teknologi Informasi & Komunikasi Terapan* 2014: Semarang. p. 193-199.

[7] Larose, D.T., *Discovering Knowledge in Data: An Introduction to Data Mining*. 2005, New Jersey: Wiley.

[8] Joshi, M.D., Karode, A.H. and.Suralkar, S.R *White Blood Cell Segmentation and Classification to Detect Acute Leukemia.* International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2013. **2**(3): p. 147-151.

[9] Supardi, N.Z., et al., *Classification of Blasts in Acute Leukemia Blood Sample Using k-Nearest Neighbour.* International Colloquium on Signal Processing and its Applications, 2012: p. 461-465.

[10] Antica, M., *Acute Leukemia – The Scientist's Perspective and Challenge*. 2011, Rijeka: InTech.

[11] Bain, B.J., *A Beginner's Guide to Blood Cells*. Second ed. 2004, Oxford: Blackwell Publishing.

[12] Sharif, J.M., et al., Red Blood Cell Segmentation Using Masking and Watershed Algorithm: A Preliminary Study, in International Conference on Biomedical Engineering (ICoBE). 2012, IEEE: Penang, Malaysia.

[13] Jack, K. 2005, Video Demystified: A Handbook for the Digital Engineer, Fourth edn, Newnes, Oxford.