Blood Cells Classification for Identification of Acute Lymphoblastic Leukemia on Microscopic Images Using Image Processing

1st Shelly Oktia Heriawati
Departement of Informatics and Computer Engineering
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
shelly.oktia@gmail.com

2nd Tri Harsono
Departement of Informatics and Computer Engineering
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
trison@eepis-its.edu

3rd Mochamad Mobed Bachtiar
Departement of Informatics and Computer Engineering
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
mochamadmobed@gmail.com

4th Yetti Hernaningsih
line 2: *Departement of Clinical Pathology Faculty of Medicine, Airlangga University*
Dr. Soetomo General Academic Hospital
Surabaya, Indonesia
yetti-h@fk.unair.ac.id

2021 International Electronics Symposium (IES) | 978-1-6654-4346-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IES53407.2021.9593939

***Abstract*—Acute lymphoblastic leukemia (ALL) is a type of leukemia (cancer of the white blood cells) that generally occurs in children. ALL have 3 sub-types, namely L1, L2, and L3. Microscopic examination to classify ALL subtypes are still done manually by hematologists through visual identification under a microscope, it is difficult to classify ALL subtypes because the characteristics of each subtype are almost the same. This paper proposes a system that is able to detect and classify subtypes of Acute Lymphoblastic Leukemia blood cells using Image Processing. The classification method using K-Nearest Neighbor (K-NN) algorithm based on geometrical and statistical features. In cell object detection, the pre-processing step is used to improve the image quality before going further to the segmentation step using threshold and watershed algorithms. 73 K-NN Dataset from all subtypes of ALL image features were generated to calculate the similarity between new unseen data. In testing results, our proposed classification system achieves 80 % overall accuracy. Each subtype's accuracy was 75 %, 73.33 %, and 93.33 % for the L1 subtype, L2 subtype, and L3 subtype.**

Keywords—Acute Lymphoblastic Leukemia, ALL Subtypes, Image Processing, Classification, K-Nearest Neighbor.

I. INTRODUCTION

Leukemia is a blood cancer caused by the body producing too many abnormal white blood cells. White blood cells are part of the immune system which are produced in the bone marrow. When the function of the bone marrow is disturbed, the white blood cells produced will change and no longer perform their role effectively. Leukemia is divided into four main types, namely Acute Lymphoblastic Leukemia (ALL), Chronic Lymphocytic leukemia (CLL), Acute Myeloblastic Leukemia (AML), and Chronic Myelocytic Leukemia (CML). Acute Lymphoblastic Leukemia occurs when the bone marrow produces too many white blood cells, a type of immature lymphocytes or lymphoblasts. Acute lymphoblastic leukemia often occurs in children, although adults can also get this diseases. According to the French-American-British (FAB) classification, Acute Lymphoblastic Leukemia (ALL) has several sub-types, namely L1, L2, and L3. L1 has a small and homogeneous cell size, the nuclei are round and regular with little clefting and inconspicuous nucleoli. The cytoplasm is scanty and usually without vacuoles. L2 has a large and heterogeneous cell size, the nuclei are irregular and often clefted. One or more, usually large nucleoli are present. The volume of cytoplasm is variable, but often abundant and may contain vacuoles. L3 has a large and heterogeneous cell size, the nuclei are regular and round-oval in shape. One or more prominent nucleoli are present. The volume of the cytoplasm is moderate and contains prominent vacuoles [1].

Acute Lymphoblastic leukemia was diagnosed with a microscopic testing method for blood smear. Microscope investigations are done manually by a haematologist through visual identification under a microscope, this method requires more effort and time [2]. The identification process of Acute Lymphoblastic Leukemia will be more effective if done using image processing, it can make it easier for haematologists to identify Acute Lymphoblastic Leukemia correctly.

A lot of researches have been conducted to identify white blood cells using image processing. Khosrosereshki et al [3] have conducted research using the binary thresholding method for the segmentation process and the fuzzy algorithm method for the classification process. This study produces an output in the form of subtype diagnosis of Acute Lymphoblastic Leukemia with an accuracy rate of 93.70%. Purwanti et al [4] have conducted research using the thresholding method in the segmentation process and the K-Nearest Neighbor method in the classification process. This study produces an output in the form of a diagnosis of normal or abnormal blood cells with an accuracy rate of 90.00%, but this study has not been able to classify subtypes of ALL.

In this research, a system that is able to detect and classify subtypes of Acute Lymphoblastic Leukemia (ALL) will be built using image processing. We us median filter method and RGB to HSV color conversion in pre-processing step before going further into main segmentation. The methods at the segmentation step used are thresholding, morphological operations, and watershed. The feature extraction used as

input for the classification process is the area, perimeter, roundness, mean, and standard deviation. Furthermore, at the classification step, the K-Nearest Neighbor method is used. The K-NN algorithm classification is calculated using Euclidean Distance. From this classification process, the output will be obtained in the form of identification of Acute Lymphoblastic Leukemia subtypes L1, L2, and L3.

II. Design and Implementation of The System

The design of the system used in this study is explained in Fig. 1.

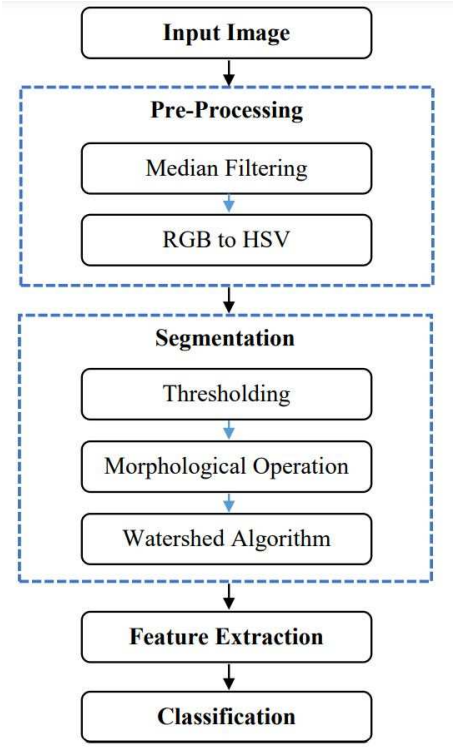

Fig. 1. Block Diagram of the System

A. Images Preparation

Microscopic image of peripheral blood smear of Acute Lymphoblastic Leukemia obtained from Dr. Soetomo Hospital Surabaya with a size of 1360 × 1024 pixels. Data were collected using a microscope with a magnification of 1000x. An example of the input image used in this study is shown in Fig. 2.

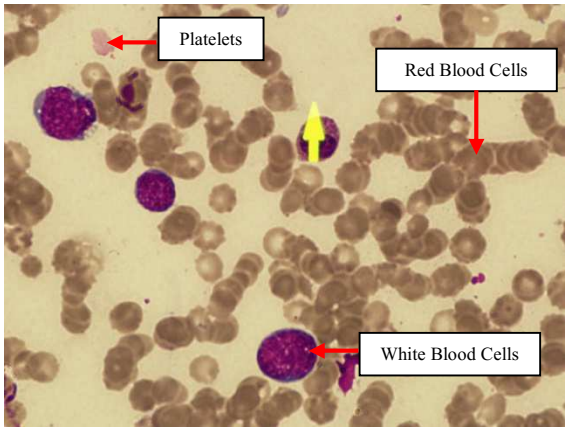

Fig. 2. Input Image

B. Preprocessing

The pre-processing step is used to refine and improve image quality to facilitate the segmentation process. The method used in the pre-processing stage is median filtering and RGB to HSV color conversion.

Median Filtering can be used to reduce noise in images in the spatial domain so that it can make the image smoother. Median filtering is done by sorting the pixel values and their neighbors from the smallest and then getting the middle value. The average pixel value in the image is obtained by using equation (1) [5].

$$F(x,y) = \frac{1}{mn}\sum\binom{m}{x-1}\sum\binom{n}{y-1}G(x,y) \quad (1)$$

Where:

F(x,y) = The result of the filtered pixel value.
G(x,y) = Neighbor matrix pixel value.
m,n = Average value of matrix size.

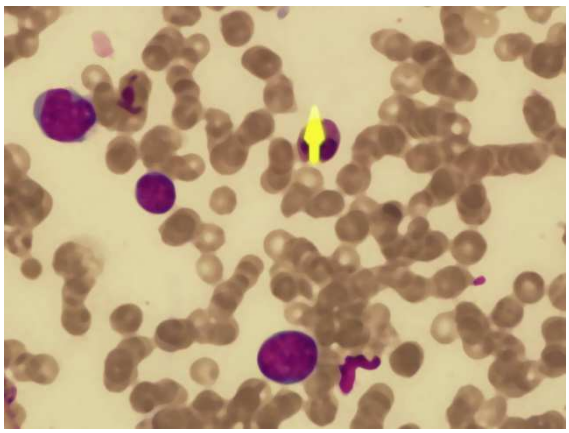

Fig. 3. Median Filter Result

From Fig. 3 it can be seen that the results of the median filtering process are able to make the image smoother.

After the median filtering process, the image will be converted from the RGB to HSV colour space. Unlike the RGB color space which is the result of a mixture of primary colors, HSV has the same colors as the colors perceived by the human senses. HSV is a color model that remaps RGB primary colors into dimensions that are easier for humans to understand.

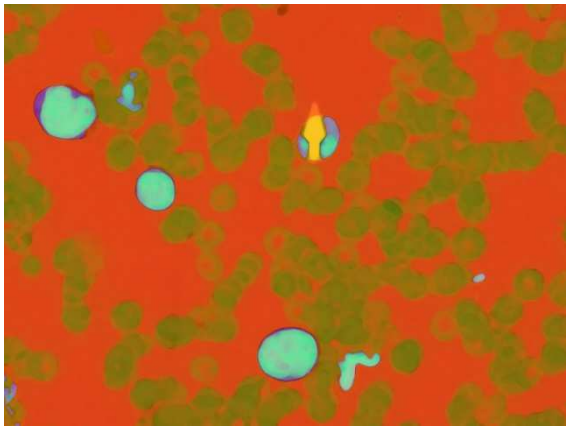

Fig. 4. RGB to HSV Conversion Result

It can be seen from Fig. 4, the results of the RGB to HSV color conversion produce different colors from the previous color.

C. Segmentation

The segmentation stage is used to separate white blood cell objects from other objects. In this step, image threshold is used to convert the grayscale to binary image. The threshold method is done by determining the threshold value, if the pixel value in the image is greater than the threshold value, then the pixel value will be changed to 1. Otherwise, the pixel value will be changed to 0. The threshold equation is shown in equation (2).

$$g(x,y) = \begin{cases} 1, & if\ f(x,y) \geq T \\ 0, & if\ f(x,y) < T \end{cases} \quad (2)$$

Where:

f(x,y) = Grayscale image.
g(x,y) = Binary image.
T = Threshold value.

The results of the thresholding process can be seen in Fig. 5 below.

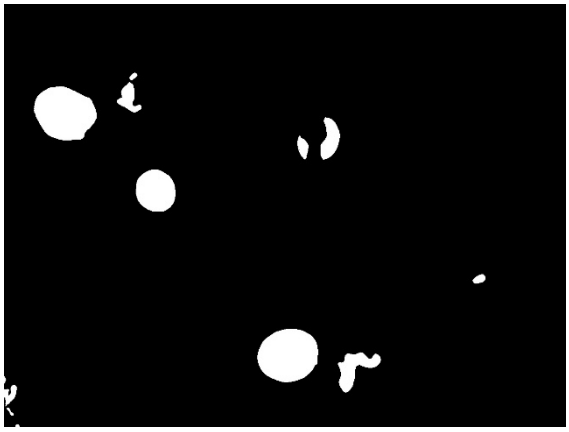

Fig. 5. Thresholding Result

The noise from thresholding operations before could be removed using morphological operations. Several white blood cell have the vacuoles which will cause the result of thresholding white blood cell objects to become hollow, therefore it is necessary to perform a morphology closing operation to fill the hollow part in white blood cells in order to be perfectly segmented. The vacuole is an organelle that is enclosed by the largest cell membrane, which looks like a white hole inside the cell. Closing is the morphological transformations of dilation followed by Erosion. It is useful in closing small holes inside the foreground objects, or small black points on the object. The dilation process is morphological operation to enlarge the object segment by adding layers around the object. While the erosion process is the morphological operation to reduce objects by eroding the edges of the object.

Opening morphology transformations also needs to be done in order to remove noise from the threshold process. The process is used to remove small objects and make the edges of white blood cell objects smoother. Opening is the reverse of closing, just another name of erosion followed by dilation. It is useful in removing noise, as we explained before. The results of the morphological operations can be seen in Fig. 6 below.

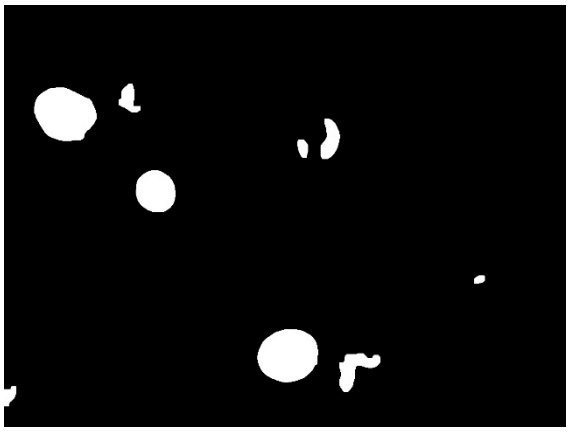

Fig. 6. Morphological Operation Result

The image resulting from the previous morphological transformation will be used in watershed process. The watershed method is used to separate white blood cell objects that are stacked or overlapped. The watershed process is done automatically. This method uses the distance transform function which is used to calculate the distance from each pixel of the image to the nearest zero pixel. The normalization process must be done before using threshold on the image. From resulted image, finding contours and give marker to each of them is necessary in order to do watershed process. To indicate that the object has been separated, each white blood cell object is labeled with a different color. The results of the watershed process can be seen in Fig. 7 below.

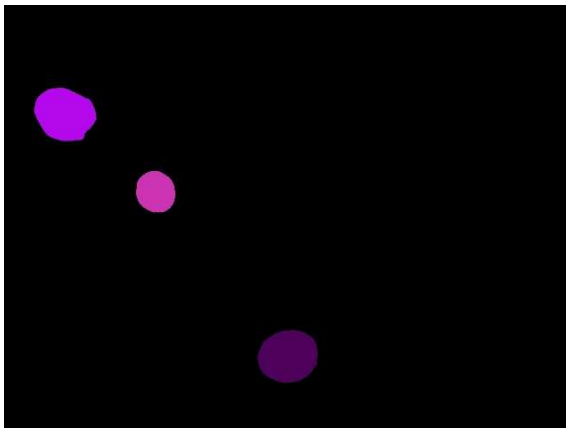

Fig. 7. Watershed Algorithm Result

From the segmentation result as shown in Fig. 7 above, then we find the contour for every white blood cell object. To get the contour, it can be done by looking for a white object from a black background. Therefore, the white blood cell image that has been watershed will be converted into a binary image first. This finding contour process is necessary to check the segmentation process result, whether is it good or not. The contour also used to object extraction features.

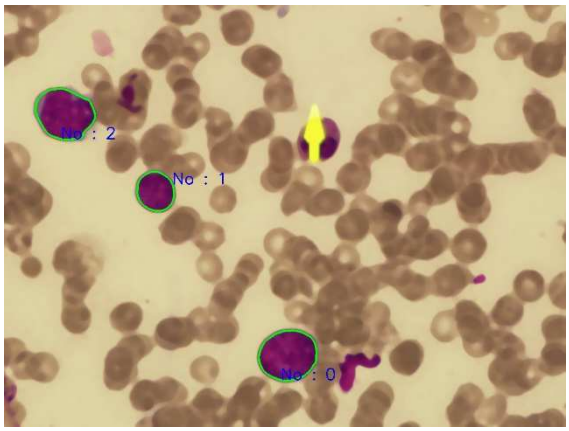

Fig. 8. Segmentation Result

The results of the segmentation that have obtained the contours are given a green marker line as shown in Fig. 8.

D. Feature Extraction

The feature extraction step is performed to estimate the characteristics of objects that distinguish them from one another. In this study, there are several parameters to be extracted, namely geometric features and statistical features. For geometric features, the white blood cell area, white blood cell perimeter, and white blood cell roundness are calculated. Meanwhile, statistical feature extraction will calculate the mean and standard deviation.

Here the equation used to calculate the values of the feature extraction results:

- Area

  The area is the number of total pixels from object of an image that forms an area. The area can show the size of the actual object [2]. The equation for the area is defined by equation (3) [16].

$$A = \frac{1}{n.m} \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} [i(x,y) | i(x,y) = 1, f(x,y)] \quad (3)$$

  Where the variable $i(x,y)$ is the pixel value of the image $f(x,y)$. Variable A defines the area. Variable n is row length and variable m is column length.

- Perimeter

  The edge of the area is the outermost part of an image object that is right next to the image background. The edge of the area can be found by counting the number of pixels that are on the border of the object [2]. The equation for the perimeter is defined by equation (4) [16].

$$P = A - (A \ominus B) \quad (4)$$

  Where P is the perimeter. Variable B is the morphological strand of the image. Variable A is the area of the object and P is the perimeter of the object.

- Roundness

  Used to see the shape of the cell nucleus that is irregular or varied. Equation (5) shows the formula for calculating form factor or also known as roundness (roundness factor) [2].

$$Roundness = \frac{4.\pi.area}{perimeter^2} \quad (5)$$

- Mean

  Used to determine the average color intensity of white blood cell objects, defined by the equation (6) [16].

$$\tilde{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad (6)$$

  Where n is the number of pixels in the image and x_i is the color intensity value for each pixel

- Standard Deviation

  The standard deviation is a measurement of the spread of a data set from its mean. The standard deviation defines the magnitude of the difference in the value of the sample data with the mean value defined by equation (7) [16].

$$\sigma = \frac{1}{n.m} (\sum_{x=0}^{m-1} \sum_{y=0}^{n-1} i(x,y) - \bar{x})^2 \quad (7)$$

  $i(x,y)$ Variable is pixel value. n variable is the total length of the image, while m is the length of the image column. $\overline{x}$ Variable define the average pixel intensity value and is the standard deviation of the white blood cell object.

E. Classification

The classification stage is carried out to identify the type of cell using the K-Nearest Neighbor (K-NN) method. White blood cells will be classified into 3 types, namely L1, L2, and L3. The K-Nearest Neighbor algorithm is calculated using the Euclidean Distance equation (8) [7].

$$d_{Euclidean}(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad (8)$$

K-Nearest Neighbor algorithm is done by determining the value of K first, then calculate the distance of the test data to each training data using Euclidean Distance. After getting the value from the distance calculation, the value will be sorted from the smallest value to the largest value (the closest distance to the farthest distance). Then we take K number from sorted data. By using the most majority nearest neighbor category, the object category can be identified. The results of the identification of Acute Lymphoblastic Leukemia subtypes can be seen in Fig. 9 below.

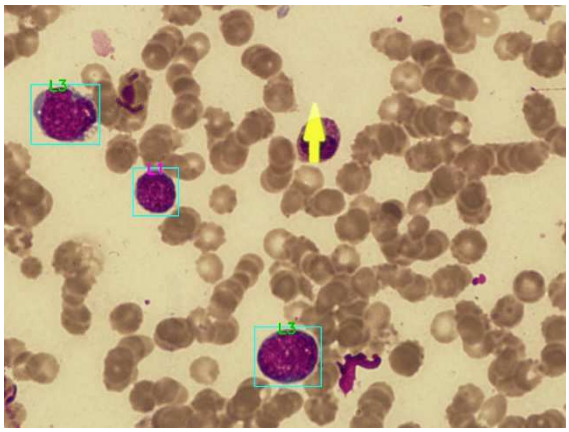

Fig. 9. Identification Result

III. Result and Discussion

In this study, we use 20 test data for the L1 subtype, 15 test data for the L2 subtype, and 15 test data for the L3 subtype. Classification is done using the K-Nearest Neighbor method, which is a method for classifying objects based on learning data that is closest to the object. In this study, an experiment was conducted to determine the best K value. The experiment by changing the parameter values from K=1 to K=20 shows that the best K value is 5. A comparison of the accuracy levels from K = 1 to K = 20 can be seen in Fig. 10.

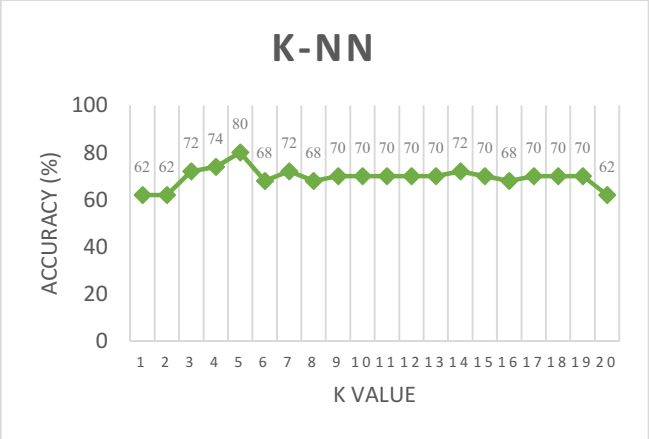

Fig. 10. K Value of K-NN

The results of system testing for each subtype of Acute Lymphoblastic Leukemia can be seen in tables I, II, and III.

TABLE I. Identification Result of L1

No.	Actual Class	System Result	Validation
1	ALL L1	ALL L1	True
2	ALL L1	ALL L1	True
3	ALL L1	ALL L1	True
4	ALL L1	ALL L1	True
5	ALL L1	ALL L1	True
6	ALL L1	ALL L2	False
7	ALL L1	ALL L1	True
8	ALL L1	ALL L1	True
9	ALL L1	ALL L2	False
10	ALL L1	ALL L1	True
11	ALL L1	ALL L3	False
12	ALL L1	Undefined	False
13	ALL L1	ALL L3	False
14	ALL L1	ALL L1	True
15	ALL L1	ALL L1	True
16	ALL L1	ALL L1	True
17	ALL L1	ALL L1	True
18	ALL L1	ALL L1	True
19	ALL L1	ALL L1	True
20	ALL L1	ALL L1	True

From Table I it can be seen that the average success of the system using the proposed method is 75% for the L1 subtype.

TABLE II. Identification Result of L2

No.	Actual Class	System Result	Validation
1	ALL L2	ALL L3	False
2	ALL L2	ALL L2	True
3	ALL L2	ALL L2	True
4	ALL L2	ALL L1	False
5	ALL L2	ALL L2	True
6	ALL L2	ALL L2	True
7	ALL L2	ALL L2	True
8	ALL L2	ALL L2	True
9	ALL L2	ALL L2	True
10	ALL L2	ALL L2	True
11	ALL L2	ALL L2	True
12	ALL L2	ALL L1	False
13	ALL L2	ALL L3	False
14	ALL L2	ALL L2	True
15	ALL L2	ALL L2	True

From Table II it can be seen that the average success of the system using the proposed method is 73.33% for the L2 subtype.

TABLE III. Identification Result of L3

No.	Actual Class	System Result	Validation
1	ALL L3	ALL L3	True
2	ALL L3	ALL L3	True
3	ALL L3	ALL L3	True
4	ALL L3	ALL L3	True
5	ALL L3	ALL L3	True
6	ALL L3	ALL L3	True
7	ALL L3	ALL L3	True
8	ALL L3	ALL L3	True
9	ALL L3	ALL L3	True
10	ALL L3	ALL L3	True
11	ALL L3	ALL L3	True
12	ALL L3	ALL L3	True
13	ALL L3	ALL L3	True
14	ALL L3	ALL L3	True
15	ALL L3	ALL L1	False

From Table III it can be seen that the average success of the system using the proposed method is 93.33% for the L3 subtype.

In this study, the L3 subtype had a higher average success rate than the L1 subtype and L2 subtype because the characteristics of the L1 subtype and L2 subtype were almost the same, while the L3 subtype has very different characteristics because of its larger cell size. Undefined results occur because there is new data that has the closest distance equal to more than 1 class type. Therefore, for further research, several things can be done, such as improve the pre-processing and segmentation processes so that cytoplasm objects in white blood cells can be segmented perfectly, adding features to the feature extraction process and adding training data will strengthen the classification results because Acute Lymphoblastic Leukemia subtype L1 and subtype L2 have almost the same characteristics, and trying to use other classification methods to produce more accurate identification results.

With a total of 73 training data and 50 test data, this study was able to classify with a success rate of 80%, therefore this system can be used as a doctor's companion to diagnose subtypes of Acute Lymphoblastic Leukemia.

IV. CONCLUSSION

From this study, it can be concluded that the classification of blood cells to identify subtypes of acute lymphoblastic leukemia using the k-nearest neighbor method with a value of K = 5 obtained an overall accuracy rate of 80%. Accuracy results for each subtype are 75% for the L1 subtype, 73.33% for the L2 subtype, and 93.33% for the L3 subtype. The K-Nearest Neighbour method in this system will achieve more accuracy if it uses more training data.

REFERENCES

[1] R. D. Labati, V. Piuri, and F. Scotti, "ALL-IDB: THE ACUTE LYMPHOBLASTIC LEUKEMIA IMAGE DATABASE FOR IMAGE PROCESSING", 2011 18th IEEE International Conferences on Image Processing, 2011.

[2] M. D. Suratin, Rahmadwati and A. Muslim, "Identifikasi Sel *Acute Lymphoblastic Leukemia* (ALL) pada Citra *Peripheral Blood Smear* Berdasarkan Morfologi Sel Darah Putih", elektronik Jurnal Arus Elektro Indonesia (eJAEI), 2015.

[3] M.A. Khosrosereskhi and M.B. Menhaj, "A Fuzzy Based Classifier for Diagnosis of Acute Lumphoblastic Leukemia using Blood Smear Image Processing", 2017 5th Irian Joint Congress on Fuzzy and Intelligent System (CFIS), 2017.

[4] E. Purwanti and E. Calista, "Detection od Acute Lymphocyte Leukemia using K-Nearest Neighbor Algorithm Based on Shape and Histogram Features", International Conference on Physical Instrumentation and Advance Materials, 2017.

[5] 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), 2018.

[6] M. Santoso, T. Indriyani, and R. E. Putra. "Deteksi *Microaneurysms* pada Citra Retina Mata Menggunakan *Matched Filter*", Journal of Information Technology, 2017.

[7] E. Suryani, Wiharto, S. Palgunadi and Yudha Rizki 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), 2017.

[8] S. Shafique, S. Tehsin, S. Anas and F. Masud, "Computer-assisted Acute Lymphoblastic Leukemia Detection and Diagnosis", 2019 2nd International Conference on Communication, Computing and Digital System (C-CODE), 2019.

[9] N. Z. Supardi, M.Y. Mashor, N.H Harun, F.A. Bakri and R. Hassan, "Classification of Blast in Acute Leukemia Blood Sample Using *k*-Nearets Neighbour", 2012 IEEE 8th International Colooquium on Signal Processing and its Application, 2012.

[10] Cancer Treatment Centers of America. Types of Leukemia. https://www.cancercenter.com/cancer-types/leukemia/types. Accessed on July 1, 2020.

[11] American Cancer Society, "Cancer Facts & Figures 2020", Atlanta, Ga: American Cancer Society, 2020.

[12] J. Dong, X. Qu, and H. Li, "Color tattoo segmentation based on Skin color space and K-mean clustering", 2017 4th International Conference on Information, Cybernetics and Computational Social System (ICCSS), 2017.

[13] G. Cheng, and J. Wei, "Color Quantization Application Based on K-Means in Remote Sensing Image Processing". ICA ACE 2019, 2019.

[14] A. M. Raid, W. M Khedr, M.A. EL-dosuky, and Mona Aoud, "Image Restoration Based on Morphological Operations", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2014.

[15] M. M. Amin, S. Kermani, A. Talebi, and M. G. Oghli, "Recognition of Acute Lymphoblastic Leukemia Cells in Microscopic Images Using K-Means Clustering and Support Vector Machine Classifier", Journal of Medical Signal & Sensors, 2015.

[16] A. 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 International Conference on Information and Communications Technolgy (ICOIACT), 2018.