



# Classification of White Blood Cell Abnormalities for Early Detection of Myeloproliferative Neoplasms Syndrome Based on K-Nearest Neighbor

Zilvanhisna Emka Fitri<sup>1</sup>, Lindri Nalentine Yolanda Syahputri<sup>2</sup>, Arizal Mujibtamala Nanda Imron<sup>3</sup>

<sup>1,2</sup>Informatics Engineering, Technology Information Departement, Politeknik Negeri Jember, Indonesia

<sup>3</sup>Electrical Engineering Department, Faculty of Engineering, Universitas Jember, Indonesia

Email: <sup>1</sup>zilvanhisnaef@polije.ac.id, <sup>2</sup>nalentine23@gmail.com, <sup>3</sup>arizal.tamala@unej.ac.id

## Abstract

The myeloproliferative neoplasms (MPNs) are clonal hematopoietic stem cell disorders characterized by dysregulated proliferation and expansion of one or more of the myeloid lineages. The initial symptoms of MPN is a bone marrow abnormalities when producing red blood cells, white blood cells and platelets in large numbers and uncontrolled. An automatic and accurate white blood cell abnormality classification system is needed. This research uses digital image processing techniques such as conversion to the modified CIELab color space, segmentation techniques based on threshold values and feature extraction processes that produce four morphological features consisting of area, perimeter, metric and compactness. then the four features become input to the K-Nearest Neighbor (KNN) method. The testing process is based on variations in the value of K to get the best accuracy percentage of 94.3% tested on 159 test data.

**Keywords:** MPNs, WBC Abnormalities, CIELab, Thresholding, KNN

## 1. INTRODUCTION

The myeloproliferative neoplasms (MPNs) are clonal hematopoietic stem cell disorders characterized by dysregulated proliferation and expansion of one or more of the myeloid lineages (erythroid, granulocytic, megakaryocytic, monocytic/macrophage, or mast cell) [1]. In the WHO classification, the spectrum of myeloproliferative disorders includes: Chronic Myelogenous Leukemia (CML) Ph positive, t(9;22) (q34;q11), Chronic Neutrophilic Leukemia (CNL), Chronic Eosinophilic Leukemia (CEL)/hypereosinophilic syndrome (HES), Chronic Idiopathic Myelofibrosis (CIMF) also known as *myelofibrosis with myeloid metaplasia* (MMM), Polycythemia vera (PV), Essential thrombocythemia (ET) and Myeloproliferative disease, unclassifiable (MPD-u) [2]. The initial symptoms of MPN is a bone marrow abnormalities when producing red blood cells, white blood cells and platelets in large numbers and uncontrolled. Peripheral blood smear examination is the first step to detect MPN early. On this examination, the medical analyst will count and classify the type of white blood cells in the peripheral blood smear. Generally, leukocyte cells are divided into 5 classes consisting of five types, namely basophils, eosinophils, neutrophils, lymphocytes and monocytes [3]. Classification of white blood cells will be done if there is an abnormality in the calculation of the number of white blood cells. The problem is the classification of white blood cell types is still manual so the results are not accurate enough and

depend on the experience of medical analysis. Therefore, we need a white blood cell classification system automatically and accurately.

Several research related to the identification of types of white blood cells carried out. In 2010, research on the recognition of leukocyte image patterns using image feature extraction methods obtained error predictions of 30% [4]. In 2012, a method for introducing leukocyte segmentation was developed by converting the HSV color space and multilevel otsu segmentation methods then detecting the region of interest (ROI) so that the shape and texture features were obtained. Then the feature selection process was carried out using the principle component analysis (PCA) method. ). the pattern recognition process uses a comparison between the KNN and SVM methods. [5]. The next development is the classification of white blood cells based on color and shape characteristics with the KNN classification method, however the resulting accuracy is 64% [6]. The KNN classification method is reused by adding contrast stretching as a step to improve image quality. The best value of k constant used is 5 so that an increase in accuracy is 94% [7]. In 2017, researchers have examined one of the myeloproliferative syndrome diseases, namely Essentia Thrombocytemia (ET). In the case of ET, there is a similarity in morphological form between giant platelets and young lymphocyte cells, so it is necessary to classify the type of platelets compared to the type of white blood cells. In that research, researchers used a modified CIELab color space conversion, then the feature used was the Gray Level Co-occurrence Matrix (GLCM) feature which became an input to the classification system. The research also compared two classification methods namely Learning Vector Quantization (LVQ) and K-Nearest Neighbor (KNN) so that it was found that KNN was able to classify better than LVQ with an accuracy of 83.67%. [8]. Based on the above research description, the researchers propose a research that classifies white blood cell types as early detection of myeloproliferative syndrome based on K-Nearest Neighbor.

## 2. METHODS

This research method adopts previous research from researchers entitled "A comparison of platelets classification from digitalization microscopic peripheral blood smears" [8]. The data used also came from the study which consisted of white blood cell image data shown in Figure 1. There are 5 types of white blood cells consisting of Polymorphonuclear (PMN), Neutrophil Bands, Lymphocytes, Monocytes and Eosinophils. Basophile cell image data does not exist because based on the results of examination of peripheral blood smear patients these cells were not found.

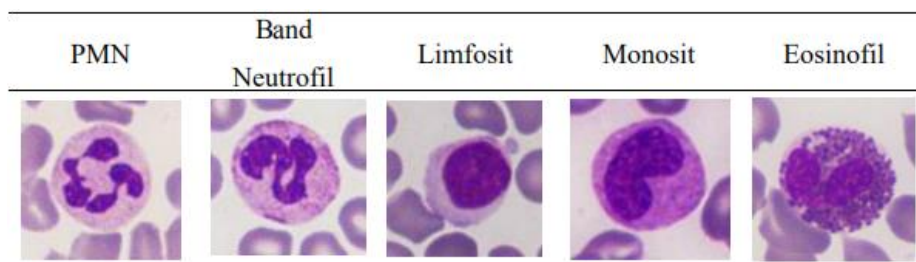


Figure 1. Variations in types of white blood cells

The first step is to convert the white cell image from the RGB image into the modified CIELab color space, then do the segmentation process based on the threshold value. This process will produce a binary image that will be carried out from the morphological feature extraction process. The result of these feature values is input from K-Nearest Neighbor which will produce an output of normal and abnormal white blood cell types as shown in the block diagram of the research in Figure 2.

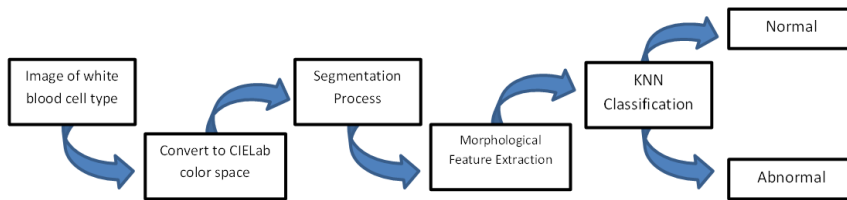


Figure 2. The research block diagram

### 2.1. Color conversion in the CIELab color space

The initial step taken in this study is to convert colors from the RGB color space to the CIELab color space. The color conversion cannot be done directly, the RGB image must first be converted to the XYZ color space using equation 1. After obtaining the XYZ value, then the conversion in the CIELab color space uses equations 2, 3, and 4

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

$$L = \begin{cases} 116 \times f\left(\frac{Y}{Y_n}\right) - 16 & \text{for } \frac{Y}{Y_n} \leq 0.008856 \\ 903.4 \times f\left(\frac{Y}{Y_n}\right) & \end{cases} \quad (2)$$

$$a = 500 \times \left[ f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right) \right] \quad (3)$$

$$b = 200 \times \left[ f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right) \right] \quad (4)$$

Where the value of  $X_n = 1$ ,  $Y_n = 0.98072$  and  $Z_n = 1.18225$ . After getting the value of each component of the CIELab color space, do the reduction process on each color component so that the AL color space is obtained. AL color space is obtained from the process of subtracting color components A from color components L [8]. The modified CIELab color space image results are shown in Figure 3.

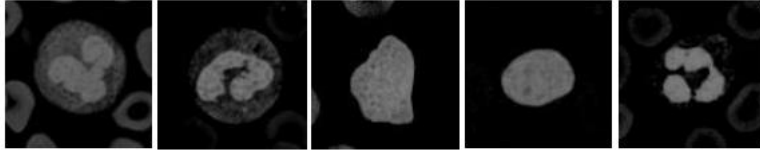


Figure 3. Converted image in the AL color space

## 2.2. Segmentation process

Segmentation is a process that aims to separate objects and backgrounds using the thresholding process so that a binary image is produced using equation 5. T value is the threshold value used in the thresholding process.

$$g(x, y) = \begin{cases} 1, & g(x, y) \geq T \\ 0, & g(x, y) < T \end{cases} \quad (5)$$

## 2.3. Feature extraction

Feature extraction is a process taking the characteristics of the object studied. In this research, morphological feature extraction which consists of area, perimeter, metric and compactness is used. Metric is usually called the form factor or backwardness of an object obtained from equation 6, while there is also compactness identifying shapes with equation 7.

$$\text{Metric} = \frac{4\pi A}{P^2} \quad (6)$$

$$\text{Compactness} = \frac{P^2}{A} \quad (7)$$

## 2.4. K-Nearest Neighbor classification

K-Nearest Neighbor (KNN) is a classification method based on learning data that is the closest distance to the object. KNN is also one of the instance based learning which algorithm is categorized as lazy learning. The basic principle of KNN is to find the K value, where the K value is the nearest amount of data that will determine the results of the classification [8]. The proximity of a data with other data is the key of the KNN algorithm which is usually calculated using the Euclidean Distance calculation with equation 8.

$$d(x_i, x_j) = \sum_{r=1}^n (x_{ir} - x_{jr})^2 \quad (8)$$

where  $x_i$  is the sample data and  $x_j$  is the test data, while  $r$  is the data variable,  $d$  is the Euclidean distance and  $n$  is the data dimension. The number of training data used was 516 images and the test data used were 159 images.

## 3. RESULT AND DISCUSSION

Researchers used several tests on digital image processing techniques, especially in the thresholding process. Some T values used are 50, 90 and 100 so the binary image results are shown in Table 1.

Table 1. Binary imagery with variations in threshold value (T)

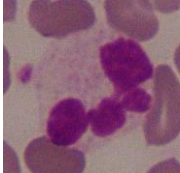



Citra PMNA 88	Threshold value (T)		
	50	90	100
			

Table 1 shows that the best threshold value is 90. This is because images with a threshold value of 90 are able to represent the best objects and separating objects from the background are better when compared to the threshold values of 50 and 100. After the thresholding process, binary image features are retrieved using the morphological feature extraction process. Morphological feature extraction is used to distinguish the size of the shape of the abnormalities of white blood cells, where the abnormal cells experience enlargement of cells from normal limits so that morphological feature extraction is used by determining the area, perimeter, metric, and compactness features. This was chosen because there was a difference from the normal white blood cell form and the abnormal white blood cell as shown in Figure 4 while the results of cell morphological feature extraction in Figure 4 are shown in Table 2.

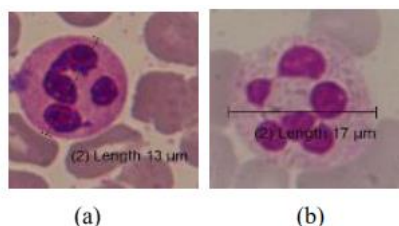


Figure 4. Image (a) normal PMN cells and (b) abnormal PMN cells

Figure 4 shows that there is a difference between normal white blood cells and abnormal white blood cells that can be seen from the cell size and the number of nucleus lobes so that to distinguish the two cells morphological feature extraction is needed.

Table 2. The difference between normal and abnormal white blood cells is based on feature values

The features	Normal white blood cells	Abnormal white blood cells
Area	8087	9905
Perimeter	626	832
Metric	0.25932754	0.179811594
Compactness	48.45752442	69.88632004

Table 2 shows that the comparison of feature extraction values from normal white blood cells and abnormal white blood cells is seen from the fourth value of the morphological feature values of each image. The area, perimeter and compactness feature values of normal white blood cell images are smaller than abnormal white blood cells, while the normal white blood cell metric feature values are greater than abnormal white blood cells. This is proven in Figure 4, normal white blood cells only consist of 2-3 lobes while abnormal white blood cells have an enlarged cell size and the number of lobes more than normal cells.

The next process is the classification process using the K-Nearest Neighbor method which is tested on 159 test data. In this study, researchers compared several K values consisting of 9, 17, 23 and 25. The percentage accuracy of the KNN method in classifying white blood cell abnormalities based on variations in K values is shown in Table 3.

Table 3. Percentage of accuracy in the classification of abnormalities of white blood cells based on variations in the value of k

K Value	Amount of data	Amount of data	Percentage of accuracy (%)
	Classification Results Correct	Classification Results Incorrect	
9	144	15	90.5
17	146	13	91.8
23	150	9	94.3
25	148	11	93.08

Based on data from Table 3, it was found that the value of  $k = 23$  shows the percentage of the best accuracy of the KNN method in classifying white blood cell abnormalities by 94.3%. Because of the 159 number of images that have been tested there are 9 test images whose classification results are not in accordance with the target. Some examples of data on the results of the KNN method classification test are shown in Table 4.

Table 4. The example of the results of the classification of the KNN method on test data

Data Name	Feature Value				KNN Classification	
	Area	Perimeter	Metric	Compactness	Target	Result
LA54	12846	527	0.581	21.62	Abnormal	Abnormal
LA55	11779	399	0.93	13.52	Abnormal	Normal
LA56	13046	450	0.81	15.52	Abnormal	Abnormal
LA57	12402	429	0.847	14.84	Abnormal	Abnormal
LA58	11846	358	1.161	10.82	Abnormal	Normal
LN80	9341	362	0.896	14.03	Normal	Normal

Table 4 shows that the results of the classification of the LA55 image and LA58 image do not match the target. LA55 image has an abnormal classification target, but after classification the result is a normal value. However, when compared with the results of

the classification of LN80 images whose classification results are in accordance with the target, which is normal, the value of the four features of the LA55 and LA58 images approaches the value of the four features in the LN80 image so that the classification results become normal.

#### 4. CONCLUSION

The KNN classification method is able to classify white blood cell abnormalities with the best accuracy of 94.3%. From a total of 159 test data, there were 150 successfully classified data according to the target, while 9 test data classification results did not match the target. This is because some features have the same value as other classes. For example the area, perimeter and compactness features indicate the range of values that point to the normal blood cell class while the metric features show the range of values that point to abnormal blood cells, so the classification results will lead to the normal blood cell class even though the target is abnormal blood cells, and vice versa. To overcome this, it is possible to make improvements in terms of the addition of white blood cell image data, improvement of digital image processing techniques or replace other classification methods so as to be able to classify white blood cell abnormalities for the better.

#### 5. REFERENCES

- [1] Verstovsek, S., & Tefferi, A. (2011). *Myeloproliferative neoplasms*. Totowa, NJ: Humana Press.
- [2] McKenzie, S. B. (2014). *Clinical laboratory hematology*. Harlow: Pearson.
- [3] Bain, B. J. (2015). *Blood Cells - A Practical Guide*. Fifth. United Kingdom: John Wiley & Sons Ltd.
- [4] Fifin, D. R. (2010). Pengenalan pola citra leukosit dengan metode ekstraksi fitur citra. *Jurnal Pendidikan Fisika Indonesia*, 6(2), 133-137.
- [5] Huang, D.C., Hung, K.D., & Chan, Y.K. (2012). A computer assisted method for leukocyte nucleus segmentation and recognition in blood smear images. *Journal of Systems and Software*, 85(9), 2104–2118.
- [6] Khasanah, M. N., Harjoko, A., & Candradewi, I. (2016). Klasifikasi Sel Darah Putih Berdasarkan Ciri Warna dan Bentuk dengan Metode K-Nearest Neighbor (K-NN). *IJEIS*, 6(2), 151.
- [7] Susilowati, I. (2016). Identifikasi penyakit leukimia akut pada citra darah mikroskopis. *Orbith: Majalah Ilmiah Pengembangan Rekayasa dan Sosial*, 12(1).
- [8] Fitri, Z. E., Purnama, I. K. E., Pramunanto, E., & Pumomo, M. H. (2017, August). A comparison of platelets classification from digitalization microscopic peripheral blood smear. In *2017 International Seminar on Intelligent Technology and Its Applications (ISITIA)* (pp. 356-361). IEEE.