Identification of Bacilli Bacteria in Acute Respiratory Infection (ARI) using Learning Vector Quantization

Zilvanhisna Emka Fitri[1,*] Lalitya Nindita Sahenda[2,] Pramuditha Shinta Dewi Puspitasari[3,] Arizal Mujibtamala Nanda Imron[4]

[1,2,3]*Department of Information Technology, Politeknik Negeri Jember, Indonesia*
[4] *Department of Electrical Engineering, Universitas Jember, Indonesia*
[*]*Corresponding author. Email:* zilvanhisnaef@polije.ac.id

ABSTRACT
Two diseases that include Acute Respiratory Infections (ARI) are diphtheria and tuberculosis. Both diseases have a large number of sufferers and can cause extraordinary events (KLB). One of the achievement indicators of infectious disease control and management programs is discovery. However, the limited number of medical analysts causes the discovery process (examination) long and subjective. To help with this problem, a bacillus identification system was created for early detection of Acute Respiratory Infections (ARI). This system is an implementation of computer vision. The data used are preparations of the bacteria Mycobacterium tuberculosis and Corynebacterium diphtheriae obtained at Besar Laboratorium Kesehatan (BBLK) Surabaya. The parameters used are the area, perimeter and shape factor. The Learning Vector Quantization (LVQ) method can classify and identify bacillus bacteria that cause acute respiratory infections with a training accuracy of 97% and a test accuracy of 86% with a learning rate of 0.01 and a reduced learning rate of 0.25.

Keywords: *acute respiratory infections, bacilli bacteria, computer vision, learning vector quantization*

1. INTRODUCTION

Acute Respiratory Infections (ARI) including the incidence of infectious diseases (disease prevalence) and morality (a measure of the number of deaths in a population) are the highest in the world [1]. ARI is divided into two, namely lower respiratory tract infections (LRTIs) and upper respiratory tract infections (URTIs). ARI is caused by bacteria, fungi and viruses [2]. Two examples of ARI diseases caused by bacteria are tuberculosis and diphtheria. The similarity between these two diseases is that the bacteria are rod-shaped or bacilli.

Tuberculosis (TB) is one of the lower respiratory tract infections caused by the bacterium *Mycobacterium tuberculosis* [3]. In 2019, the discovery and treatment rate of all TB cases in East Java ranked second in Indonesia as many as 64,311 cases with a Case Detection Rate (CDR) of 66% while the CDR target set was at least 70% [4]. While diphtheria is an upper respiratory tract infection, which is usually found in the larynx and is caused by the bacterium *Corynebacterium diphtheriae* [3]. From year to year in East Java, the number of diphtheria sufferers is reported to continue to increase until in 2019 there were 358 cases [4].

One of the indicator achievements of the infectious disease control and handling program is disease discovery [4]. Generally, the discovery of the disease is done by sputum examination or sputum specimen from suspected then examined microscopically. The problem that often occurs is a large number of patients and the limited number of medical analysts causing the microscopic examination process to take a long time. In addition, medical analysts need high experience in identifying bacteria so that the identification process is subjective. Based on the description of the problem, the researchers created a bacilli identification system for early detection of acute respiratory infection (ARI) which can assist medical analysts in microscopic examination.

2. RELATED WORKS

Some of the references we use are the application of the Learning Vector Quantization (LVQ) method in identifying *Mycobacterium tuberculosis* and non-

bacterial bacteria with an accuracy rate of 91.33% [5]. K-means method is used to classify color image segmentation on tuberculosis identification with an accuracy rate of 97.68% [6]. The Channel Area Thresholding (CAT) segmentation algorithm can identify bacillus bacteria in sputum images with an accuracy of 97.58% [7]. The LVQ method is used again in classifying acute respiratory infections (ARI) with the highest accuracy of 100% [1]. In other cases, the LVQ method is also able to classify Neisseria gonorrhoeae bacteria as early detection of gonorrhoea with an accuracy rate of 91.67% [8].

Based on the reference explanation above, the researcher used the Learning Vector Quantization (LVQ) method to identify bacilli in acute respiratory infections (ARI).

3. METHODS

The stages of research is data collection, process cropping, color conversion, segmentation, feature extraction and identification of bacteria using the Learning Vector Quantization method as shown in Figure 1.

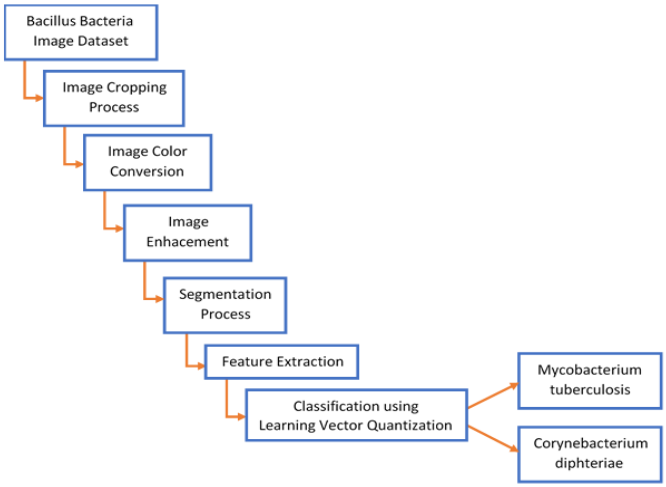

Figure 1 Block Diagram of System Identification of Bacilli Bacteria in Acute Respiratory Infection (ARI)

3.1. Bacillus Bacteria Image Dataset

The data used are preparations of the bacteria Mycobacterium tuberculosis and Corynebacterium diphtheriae obtained at Besar Laboratorium Kesehatan (BBLK) Surabaya. Then the data is taken for the image of the bacteria using a digital microscope. Generally, bacteria have three forms, namely cocci, bacilli and spirochetes and the size of the bacteria is 0.4 to 2 μm [9]. Based on the atlas of medical microbiology, Corynebacterium diphtheria includes gram-positive bacteria, clubshape and pleomorphic rods [10], while Mycobacterium tuberculosis is an aerobic acid-fast rods bacteria [11] as shown in Table 1.

3.2. Image Cropping Process

The purpose of the image cropping process is to reduce the computational load because the original size of the bacteria image is 4000x3000 pixels. In addition, the cropping process is also part of the normalization of image size so that the image used becomes 161x241 pixels as shown in Figure 2.

Table 1. Bacilli Bacteria in Acute Respiratory Infection

Disease	Bacteria Name	Bacteria Image	
Tuberculosis	*Mycobacterium tuberculosis*		
Diphtheria	*Corynebacterium diphtheria*		

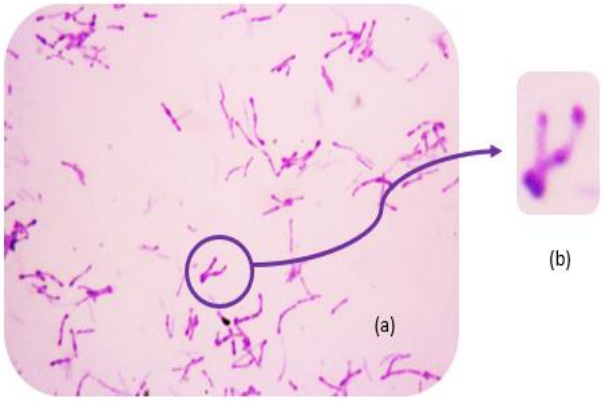

Figure 2 Image cropping results from (a) 4000x3000 pixels to (b) 161x241 pixels

3.3. Image Color Conversion

Generally, the original image is an RGB color space image or consists of 3 components, namely a red component, a green component and a blue component. The RGB color space is also considered to have a large size so that the color space is difficult to segment [12] then each color component must be separated [13] or conversion to another color space. In this research, the researcher separates each RGB color component and converts the RGB color to a grayscale color space in each bacterial image as shown in Figure 3.

3.4. Image Enhancement

The purpose of this process is to improve the quality of the image, for example, increasing the intensity of the light and darkness of the RGB component image and the

grayscale image. In this research, we used a contrast stretching process that serves to distribute the intensity of the light and dark on the whole scale of intensity, so that the resulting image with high contrast value [14].

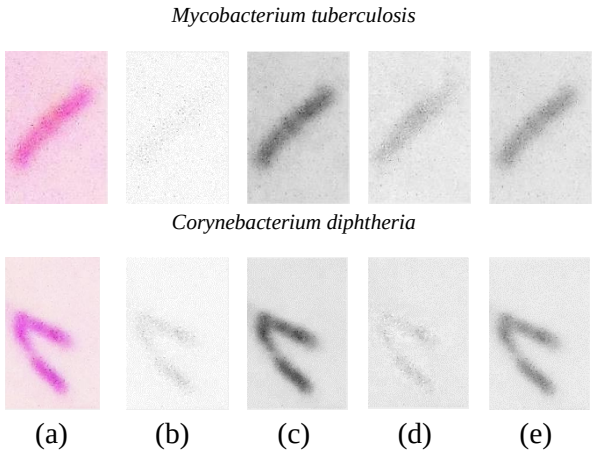

Figure 3 (a) The original image of bacteria, (b) the image of the red component, (c) the image of the green component, (d) the image of the blue component and (e) the image of the grayscale.

3.5. Segmentation Process

This stage aims to separate the object and background. This process is divided into two stages, namely, segmentation based on the grey threshold value (T) and segmentation based on Channel Area Thresholding (S.Area). The formula equation for the segmentation process based on the grey threshold value (T) is as follows [15]:

$$segmentation\ (a,b) = \begin{cases} 1, if\ grayscale(a,b) \leq T \\ 0, if\ grayscale(a,b) > T \end{cases} \quad (1)$$

The segmentation based on Channel Area Thresholding (CAT) uses a labeling process with neighboring techniques in 4 or 8 directions. Then we look for the area on each labeling and segmented based on the area.

3.6. Feature Extraction

The next step is feature extraction where at this stage the function is to take features or parameters from the shape of the bacillus bacteria. Some of the morphological features used in this research are area, perimeter and shape with the formula equation [14]:

$$Area = Number\ of\ pixels\ in\ row - 1 + row\ to - 2 + \cdots + row\ to - 8 \quad (2)$$

$$Perimeter = \sum Even\ code + \sqrt{2}\,x \sum odd\ code \quad (3)$$

$$Shape = \frac{P^2}{A} \quad (4)$$

Where A is the Area, P is the perimeter and S is the shape factor. These three parameters are obtained from the chain code calculation process using labelling based on neighboring relationships.

3.7. Learning Vector Quantization

Learning Vector Quantization (LVQ) is one of the artificial neural network methods with supervised learning methods to classify input vectors into certain classes. The classification results depend on the distance between the input vectors. If there are two input vectors that have similar distances, the competitive layer will classify the two input vectors into the same class [13]. The LVQ network architecture is shown in Figure 4. The algorithm of the Learning Vector Quantization (LVQ) method is as follows :

1. Set the input parameters: weight (W), learning rate (α), reduction of learning rate (Dec α), minimum learning rate (Min α) and Maximum Iteration (MaxEpoch),
2. Enter the input matrix: B (x, y) and target matrix: T (1, y). where x = number of inputs and n = amount of data.
3. Determine the initial condition :
   a. Error = 1
   b. Epoch = 0
4. If (epoch <MaxEpoch) or (α> Minα) then
   a. An update of epoh value, Epoch = epoch +1
   b. Perform the iteration in the input until some data is trained, for x = 1 to n
5. Find J so that | B - Wy | is a minimum
6. Updating weight value Wy.
   a. IF T = J Then,
      Wj (new) = Wj (old) + α (B- Wj (old))
   b. IF T ≠ J Then
      Wj (new) = Wj (old) + α (B - Wj (old))
7. Reduce the value of learning rate α.

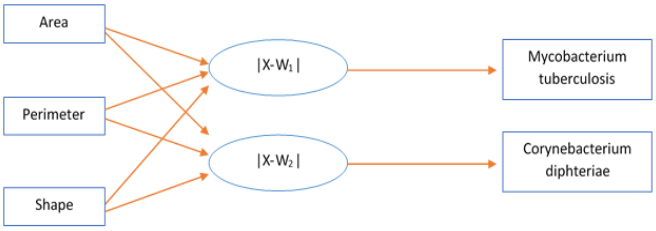

Figure 4 Network architecture on the Learning Vector Quantization method

The total number of data in this research is 150 data consisting of 100 training data and 50 testing data.

3.8. The Calculation of Performance Method

One technique to calculate the ability of the classification method is to use the Receiver Operating Characteristic (ROC) technique. This technique produces four characteristic values, namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) which are shown in the illustration Table 2. Based on this value, the sensitivity or True Positive Rate (TPR) value, specificity or False Positive Rate (FPR) and accuracy values are obtained using the formula equation [16] :

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

$$FPR = \frac{FP}{FP + TN} \quad (6)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} x100\% \quad (7)$$

Table 2. Confusion Matrix

Output Class		Target Class
A	B	
TP	FP	A
FN	TN	B

4. EXPERIMENTAL RESULT

4.1. Image Color Conversion

Figure 3 shows that the image that best represents the shape of bacteria is (c) the image of the green color component for both images of bacteria. In the green component image, the difference between the object and the background is clearly visible when compared to (e) grayscale images and (d) blue component images, so to improve image quality, an image enhancement process is carried out.

4.2. Contrast Stretching

In the stages of image enhancement, the image of the green component is added with contrast or the contrast stretching process. The addition process also affects the histogram of the bacterial image as shown in Figure 5. The picture shows that there is a change in the histogram of the green component of *Mycobacterium tuberculosis*. In Figure 5 (a) the gray value has a dark-light intensity range of 70 to 240, while after adding contrast stretching, the gray value in the green component image histogram becomes evenly distributed according to the intensity scale or has a gray value range of 0 to 255.

4.3. Segmentation Process

The next step is the segmentation process based on the gray threshold value using the equation (1). In this research, several threshold values were used, namely 10, 50, 100, 150. This value determines the segmentation results as shown in Figure 6. That figure shows that the gray threshold value (T) that best represents the shape of bacteria is T = 150. if using a value of T < 100, then the shape of the bacteria *Corynebacterium diphtheria* (bottom) is separate, this is different when using T = 150. However, in the *Mycobacterium tuberculosis* bacterial segmentation image there are still other objects or noise (Figure 7) when using T=150, so further segmentation is needed using Channel Area Thresholding (CAT).

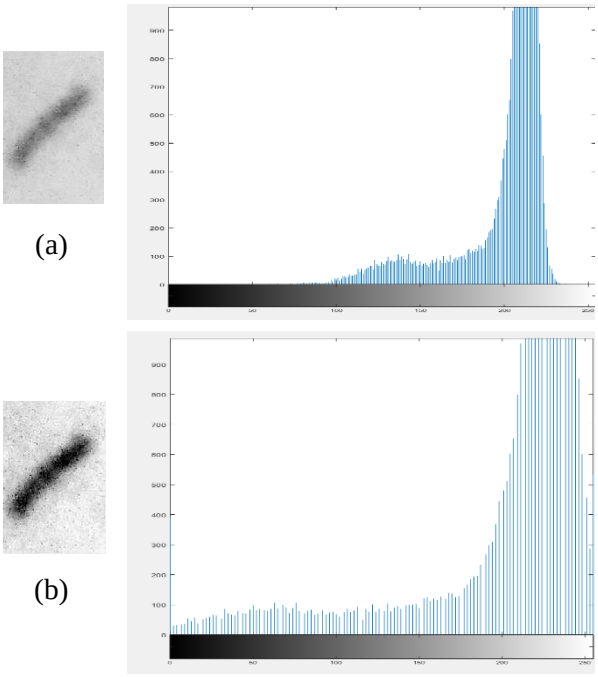

Figure 5 Histogram of the green component image (a) before and (b) after the addition of contrast stretching on Mycobacterium tuberculosis.

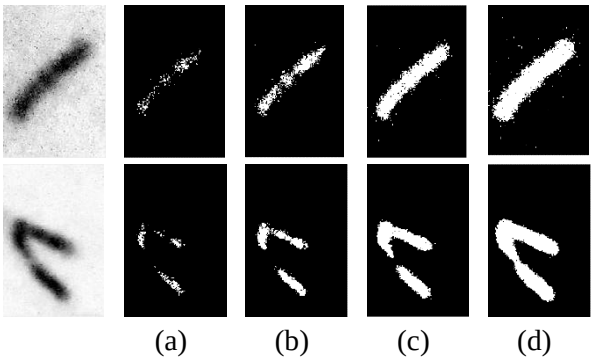

Figure 6 The results of image segmentation based on a threshold value of gray (a) T = 10, (b) T = 50, (c) T = 100 and (d) T = 150

The segmentation image used a labeling process with a chain code based on the 8-direction neighbor relationship. After the objects are labeled, then we look for the threshold value for the area [S.Area]. In this research, one threshold value was used, namely [S. Area] ≤ 100, so that the segmentation results are obtained as shown in Figure 7. The image shows that there are white dots that have an area of fewer than 100 pixels. When the CAT process is carried out, the white dots will disappear and leave a large object that has an area of more than 100 pixels or the bacteria itself.

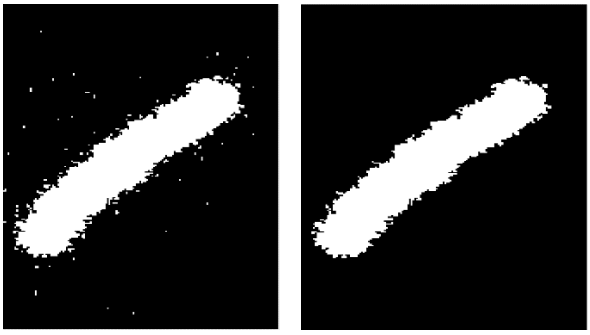

Figure 7 The results of Image Segmentation based on Channel Area Thresholding (CAT).

4.4. Feature Extraction

After getting the segmentation image based on the CAT technique, the next step is the feature extraction technique. The features used in this research are 3, namely area, perimeter and shape obtained from the equations of formulas 2 to 4. Table 3 shows the results of feature extraction for each bacterium.

Table 3. The Value of Features on Each Class of Bacteria

Class		Area	Perimeter	Shape
Mycobacterium tuberculosis	Minimum	1075	115	12.30
	Maximum	6295	505	63.27
	Average	3276	295	26.89
Corynebacterium diphtheriae	Minimum	1022	149	14.83
	Maximum	6528	556	75.82
	Average	3720	368	39.44

Based on the results in Table 3, the average area of *Corynebacterium diphtheriae* is 3276 while the average area of *Mycobacterium tuberculosis* is 3720. On average *Corynebacterium diphtheriae* perimeter is 368. This feature is larger than the average perimeter of *Mycobacterium tuberculosis* which is 295. The average form factor (shape) of *Mycobacterium tuberculosis* was 26.89 while the average shape of *Corynebacterium diphtheriae* was 39.44. It can be concluded that *Corynebacterium diphtheriae* has a larger area, perimeter and shape than *Mycobacterium tuberculosis*.

4.5. Classification using Learning Vector Quantization

In this research, 100 training data and 50 testing data were used. The learning rate (α) used in this research were 0.1; 0.01; 0.001 with a reduction of the learning rate (dec α) were 0.01; 0.1; 0.25. The maximum epoch that is used is the 5000 iteration with an error rate 0.000001. Based on these input parameters, the training accuracy is shown in Table 4.

Table 4. Accuracy Results of The Learning Vector Quantization Method

Learning rate (α)	Reduction of learning rate (Dec α)	Accuracy (%)
0.1	0.01	96
	0.1	96
	0.25	96
0.01	0.01	96
	0.1	96
	0.25	97
0.001	0.01	96
	0.1	95
	0.25	96

Table 4 shows that the highest accuracy on the training system is 97% with a learning rate (α) of 0.01 and a reduction of learning rate (dec α) of 0.25. The calculation of accuracy is obtained from the confusion matrix of the training process which is shown in Table 5.

Table 5. The Confusion Matrix of Training LVQ

Output Class		Target Class
A	B	
47	3	A : *Corynebacterium diphtheriae*
0	50	B : *Mycobacterium tuberculosis*

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} x100\% = 97\%$$

The next step is the testing process with 50 data with a learning rate (α) of 0.01 and a reduction of the learning rate (dec α) of 0.25. The testing results were obtained using the calculations in the confusion matrix shown in Table 6.

Table 6. The Confusion Matrix of Testing LVQ

Output Class		Target Class
A	B	
21	4	A : *Corynebacterium diphtheriae*
3	22	B : *Mycobacterium tuberculosis*

Table 6 shows that the Learning Vector Quantization classification method can correctly classify 21 data on *Corynebacterium diphtheriae* bacteria and 22 data on *Mycobacterium tuberculosis* bacteria. While the four data should be *Corynebacterium diphtheriae* bacteria, but the classification results show that the four data are classified as *Mycobacterium tuberculosis* bacteria. On the other hand, three data were also misclassified into *Corynebacterium diphtheriae* bacteria.

4.6. The Calculation of Performance Method

Based on the confusion matrix in Table 6, the sensitivity or True Positive Rate (TPR), specificity or False Positive Rate (FPR) and accuracy values are calculated using the equations 5 to 7.

$$Sensitivity\ (TPR) = \frac{TP}{TP + FN} = \frac{21}{21 + 3} = 0.875$$

$$Specificity\ (FPR) = \frac{FP}{FP + TN} = \frac{4}{4 + 22} = 0.154$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} x100\% = 86\%$$

Sensitivity is a value that indicates that the method can identify the image correctly, while specificity indicates the error rate of the classification method in identifying bacterial images. The value of sensitivity (TPR) is 0.875, the value of specificity (FPR) is 0.154 and the level of accuracy in the testing process of the system is 86%.

5. CONCLUSION

The results of the research showed that the Learning Vector Quantization (LVQ) method was able to identify bacillus bacteria that cause Acute Respiratory Infection (ARI) with the best training accuracy rate of 97%. The learning rate used is 0.01 and the reduction of learning rate is 0.25, so that the testing accuracy rate on 50 data is 86%.

ACKNOWLEDGMENTS

The authors would like to thank the Balai Besar Laboratorium Kesehatan (BBLK) Surabaya for their willingness to provide data on TB and diphtheria bacteria. We also thank the Politeknik Negeri Jember for funding our PNBP research.

REFERENCES

[1] E. Setyowati and S. Mariani, "Penerapan Jaringan Syaraf Tiruan dengan Metode Learning Vector Quantization (LVQ) untuk Klasifikasi Penyakit Infeksi Saluran Pernapasan Akut (ISPA)," in *PRISMA*, Semarang, 2021, vol. 4, p. 10.

[2] J. K. Struthers, "Clinical Microbiology, Second Edition," *CRC Press*, p. 299, 2017.

[3] S. J. Pitt, *Clinical Microbiology for Diagnostic Laboratory Scientists*. Chichester, UK: John Wiley & Sons, Ltd, 2017. doi: 10.1002/9781118745847.

[4] Dinas Kesehatan Provinsi Jawa Timur, *Profil Kesehatan Provinsi Jawa Timur 2019*. Surabaya: Dinas Kesehatan Provinsi Jawa Timur, 2020.

[5] E. Purwanti and P. Widiyanti, "Using Learning Vector Quantization Method For Automated Identification of Mycobacterium Tuberculosis," *Indonesian Journal of Tropical Infectious Disease*, vol. 3, no. 1, p. 26, Jul. 2015, doi: 10.20473/ijtid.v3i1.198.

[6] R. Rulaningtyas, Andriyan Bayu Suksmono, T. Mengko, and P. Saptawati, "Multi patch approach in K-means clustering method for color image segmentation in pulmonary tuberculosis identification," in *2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, Bandung, Indonesia, Nov. 2015, pp. 75–78. doi: 10.1109/ICICI-BME.2015.7401338.

[7] K. S. Mithra and W. R. S. Emmanuel, "Segmentation of Mycobacterium Tuberculosis Bacterium From ZN Stained Microscopic Sputum Images," in *2018 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, Dec. 2018, pp. 150–154. doi: 10.1109/ICSSIT.2018.8748294.

[8] Z. E. Fitri, "Implementasi Jaringan Syaraf Tiruan untuk Identifikasi Bakteri Neisseria gonorrhoeae Sebagai Deteksi Dini Gonore," Universitas Airlangga, Surabaya, 2015.

[9] C. R. Mahon and D. C. Lehman, Eds., *Textbook of diagnostic microbiology*, Sixth edition. St. Louis, Missouri: Elsevier Saunders, 2019.

[10] F. H. Kayser, Ed., *Medical microbiology*. Stuttgart ; New York, NY: Georg Thieme Verlag, 2005.

[11] P. R. Murray, *Basic Medical Microbiology*. Philadelphia: Elsevier, 2018.

[12] A. M. Nanda Imron and Z. E. Fitri, "A Classification of Platelets in Peripheral Blood Smear Image as an Early Detection of Myeloproliferative Syndrome Using Gray Level Co-Occurence Matrix," *Journal of Physics: Conference Series*, vol. 1201, p. 012049, May 2019, doi: 10.1088/1742-6596/1201/1/012049.

[13] Z. E. Fitri, I. K. E. Purnama, E. Pramunanto, and M. H. Pumomo, "A comparison of platelets classification from digitalization microscopic peripheral blood smear," in *2017 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Surabaya, Aug. 2017, pp. 356–361. doi: 10.1109/ISITIA.2017.8124109.

[14] Z. E. Fitri, L. N. Sahenda, P. S. D. Puspitasari, P. Destarianto, D. L. Rukmi, and A. M. N. Imron, "The Classification of Acute Respiratory Infection (ARI) Bacteria Based on K-Nearest Neighbor," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 12, no. 2, p. 11, 2021.

[15] Z. E. Fitri, L. N. Y. Syahputri, and M. N. Imron, "Classification of White Blood Cell Abnormalities for Early Detection of Myeloproliferative Neoplasms Syndrome Based on K-Nearest Neighbor," *Scientific Journal of Informatics*, vol. 7, no. 1, p. 7, 2020.

[16] Z. E. Fitri, "Klasifikasi Trombosit pada Citra Hapusan Darah Tepi Berdasarkan Gray Level Co-Occurrence Matrix Menggunakan Backpropagation," thesis, Institut Teknologi Sepuluh Nopember, Surabaya, 2017.