Detection of acute lymphocyte leukemia using k-nearest neighbor algorithm based on shape and histogram features

Endah Purwanti and Evelyn Calista
Biomedical Engineering, Faculty of Science and Technology, Universitas Airlangga, Kampus C, Mulyorejo, Surabaya 60115, Indonesia
Corresponding email address: endah-p-1@fst.unair.ac.id

Abstract. Leukemia is a type of cancer which is caused by malignant neoplasms in leukocyte cells. Leukemia disease which can cause death quickly enough for the sufferer is a type of acute lymphocyte leukemia (ALL). In this study, we propose automatic detection of lymphocyte leukemia through classification of lymphocyte cell images obtained from peripheral blood smear single cell. There are two main objectives in this study. The first is to extract featuring cells. The second objective is to classify the lymphocyte cells into two classes, namely normal and abnormal lymphocytes. In conducting this study, we use combination of shape feature and histogram feature, and the classification algorithm is k-nearest Neighbour with k variation is 1, 3, 5, 7, 9, 11, 13, and 15. The best level of accuracy, sensitivity, and specificity in this study are 90%, 90%, and 90%, and they were obtained from combined features of area-perimeter-mean-standard deviation with k=7.

1. Introduction
The leukemia type will be used in this study is ALL. ALL is leukocyte cancer characterized by overproduction and continuous multiplication of malignant and immature leukocyte (lymphoblast or blast cell). One of the processes carried out in this study is obtaining the pattern image of normal leukocyte cells and blast cells. Therefore, the selection of features is essential in order to obtain information which distinguishes normal leukocyte cells from blast cells. Features used in this study were shape feature of the nucleus and histogram features. Shape features of the nucleus were adapted from those used by Kulkarni and Bhosale (2014), and the types of shape features that used were area, perimeter, eccentricity, form factor, and solidity [1]. Meanwhile, the histogram feature was adapted from that used by Scotti (2005), and it included mean deviation and standard deviation. The result of extraction features would be used as the input data in the next process, which was classification with k-Nearest neighbour. Euclidean distance was used to calculate the distance between testing data and reference data.

2. Methodology / experimental

2.1. Image dataset
Image dataset in this study used a public dataset ALL-IDB2 provided by Dr Fabio Scotti. Used imagery is a single image of peripheral blood cell with the size of 257 x 257 pixels.
2.2. **Proposed system**

Proposed system included pre-processing, nucleus segmentation, feature extraction, and classification. Figure 1 shows the proposed system algorithm in this study.

![Proposed algorithm system](image)

**Figure 1.** Proposed algorithm system

2.2.1. **Gray scaling.**

Gray scaling is an RGB conversion stage to a gray scale image by eliminating the hue and saturation information to maintain luminance [6]. This step was conducted to change RGB image with three-dimensional matrix into a gray scale image with two-dimensional matrix to make it easier to be processed.

2.2.2. **Pre-processing.**

In some cases, the preparation image encountered blurring, low contrast, and unwanted noise [7]. Therefore, pre-processing step was conducted to enhance the quality of the image. Median filtering was used in this step to remove unwanted noise.

2.2.3. **Nucleus segmentation.**

In this step, the segmentation used Otsu’s thresholding method to take the nucleus of lymphocyte. The shape of nucleus is important to describe a blast cells. Lymphocyte cell has blue and regular shape of
nucleus, whereas lymphoblast cell has irregular shape and spherical particles in nucleus [5]. To remove unwanted area which was still segmented with Otsu’s thresholding method, clearing border and component labelling were used. Clearing border is one of morphological image processing as reconstruction application. It is used to remove objects touching the edge of image [8]. Component labelling is an image examination and classification of each pixel into a connected component according to the rules of connectivity [10].

2.2.4. Feature Extraction.

The extracted features were shape feature and histogram feature. Histogram feature was extracted using the first level of the nucleus image which had been created with preprocessing. Meanwhile, shape feature was extracted using an image which had been created with nucleus segmentation.

- Area: total number of nonzero pixel [5]
- Perimeter: total pixels of boundary image [5]
- Eccentricity: object’s roundness with value 0 to 1. Perfect round object has 0 eccentricity and line segment has 1 eccentricity [5]
- Form factor: function of an area and perimeter of an object [5]
- Solidity: ratio of actual and convex hull area [5]
- Mean: the average intensity of the image
- Standard deviation: the deviation of mean

2.2.5. Classification.

Reference data used was 50 images for each class of normal and abnormal. Meanwhile, testing data used was 10 images for each class of normal and abnormal. This study was conducted by varying the parameter values of $k=1$ to $k=15$ [11] and also varying combinations of the image features. K-NN classifiers were used to look for a group of $k$ objects on the training data set which was the closest to the test data and base the labelling of the dominant class on neighbouring regions [12]. According to current knowledge on the field, the selection of $k$ optimal value depends on the amount of data that it requires different $k$ values for different applications. This method is simple, but it is very effective in some cases [4]. Euclidean distance is used as a proximal distance calculation among neighbours. Equation 1 shows the Euclidean formula.

$$d_2(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^{r}(x_i - y_i)^2}$$

in which $x, y \in X$, and $x_i, y_i$ are featuring values of -I from $x$ and $y$, and whereas $r$ is the number of features in vector [9].

Determination of accuracy, sensitivity, and specificity is based on a comparison of the results of the test data output from the k-NN classification with the classification results of medical experts. Calculation used to get accuracy used ratio of correctly classified testing data and the number of testing data. Sensitivity was calculated from ratio of between true positive and the sum result of true positive and false negative, whereas specificity was calculated from ratio of true negative and the sum result of true negative and false positive.

TP (true positive) is the number of images correctly classified as positive on the test. TN (true negative) is the number of images correctly classified as negative in the test. FP (false positive) is the number of images classified as positive on the test while actually they are not. Meanwhile, FN (false negative) is the number of images that classified as negative image on the test while actually they are not [3].

3. Results and discussion

The image results of each step in the proposed system to segmenting nucleus are shown in Figure 2.
There are four combination features used in this study, such as (mean-standard deviation-area-perimeter-eccentricity-form factor-solidity), (area-perimeter-eccentricity-form factor-solidity), (mean-standard deviation), (area-perimeter), and (Mean-Standard Deviation-Area-Perimeter). Table 1 shows the highest accuracy, sensitivity, and specificity of each combination feature obtained with k-NN.

| Combination Features | $k$ | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|----------------------|-----|--------------|-----------------|-----------------|
| All features         | 7   | 90           | 90              | 90              |
| Shape features       | 7   | 85           | 80              | 90              |
|                      | 5   | 75           | 80              | 70              |
|                      | 9   | 75           | 80              | 70              |
|                      | 11  | 75           | 80              | 70              |
|                      | 13  | 75           | 80              | 70              |
|                      | 15  | 75           | 80              | 70              |
| Area-Perimeter       | 7   | 85           | 80              | 90              |
| Mean-Standard Deviation-Area-Perimeter | 7 | 90 | 90 | 90 |

Based on the information shown in Table 1, the highest accuracy, sensitivity, and specificity are obtained in combination (mean-standard deviation-area-perimeter-eccentricity-form factor-solidity) and (mean-standard deviation-area-perimeter) with $k=7$. However, the most effective combination features are mean-standard deviation-area-perimeter since the number of combinations is fewer than that of the combination features with similar accuracy, sensitivity, and specificity. Therefore, the computational load is also fewer than the others. In terms of application, it is advantageous for accelerating the working time of the program.

As a classification tool for medical experts to detect any abnormality in leukocytes led to acute leukemia, these applications still need to be developed, such as adding class classification. For example, the class is divided into more specific abnormalities namely acute lymphocytic Leukemia-L1 (ALL-L1), acute lymphocytic Leukemia-L2 (ALL-L2), and acute lymphocytic Leukemia-L3 (ALL-
Furthermore, it also needs to be added with a feature which is able to count the number of abnormal leukocyte cells in the visual field preparations. Since to determine whether a person suffers from Leukemia and the type of Leukemia they suffer, it is required to calculate the number of abnormal leukocytes and any abnormal cell types found in the visual field preparations.

4. Conclusion
This study proposes a strategy and methodology to detect Acute Lymphocyte Leukemia (ALL). We have used combination made of shape and histogram features. The purpose of using $k$-NN algorithm is to classify the lymphocyte cells into two classes, namely normal and abnormal lymphocytes. The results show that the best combination features in this study are the combination of area-perimeter-mean-standard deviation with $k=7$. The values of accuracy, sensitivity, and specificity obtained are 90%, 90%, and 90%.

References
[1] Kulkarni T A dan Dilip S Bhosale 2014 A Robust Morphological Analysis of Normal and Abnormal Leukemic Cells Populations in Acute Lymphoblastic Leukemia, National Conf. on Recent Innovations in Science Engineering & Technology India
[2] American Cancer Society 2014 Leukemia-Acute Lymphocytic (Adults) www.cancer.org/Leukemia-acute-lymphocytic-pd 10 September 2014
[3] Bain B J 2003 Leukemia Diagnosis (United Kingdom: Blackwell Publishing)
[4] Deore S G and Neeta Nemade 2013 Image Analysis Framework for Automatic Extraction of the Progress of an Infection International Journal of Advanced Research in Computer Science and Software Engineering Volume 3 Issue 6 pp 703-6
[5] Joshi M D, Atul H Karode, S R Suralkar, 2013 White Blood Cells Segmentation and Classification to Detect Acute Leukemia International Journal of Emerging Trends & Technology in Computer Science Volume 2 Issue 3 May-June 2013: 147-50
[6] Mathworks 2008 Image Processing Toolbox\textsuperscript{TM} User’s Guide (United States: The MathWorks, Inc.)
[7] Mokhtar N R, Nor Hazlyna Harun, M Y Mashor, H.Roseline, Nazahah Mustafa, R Adollah, H Adillah, N F Moh Nasir 2009 Image Enhancement Techniques Using Local, Global, Bright, Dark, and Partial Contrast Stretching for Acute Leukemia Images (London: WCE) July 1-3
[8] Prasetyo E 2011 Pengolahan Citra Digital dan Aplikasinya Menggunakan MATLAB (Yogyakarta: Penerbit Andi)
[9] Prasetyo E 2014 Data Mining: Mengolah Data Menjadi Informasi Menggunakan MATLAB (Yogyakarta: Penerbit Andi)
[10] Putra D 2010 Pengolahan Citra Digital (Yogyakarta: Penerbit Andi)
[11] Scotti F 2005 Automatic Morphological Analysis for Acute Leukemia Identification in Peripheral Blood Microscope Images IEEE International Conference on Computationa; Intelligence for Measurement Systems and Applications.
[12] Wu Xindong and Vipin Kumar 2009 The Top Ten Algorithms in Data Mining (United States: Taylor & Francis Group)