Automated Intelligent Diagnostic Procedure for Chronic Leukemia

Chan Lok Pooi1, Mohd Yusoff Mashor1*, Robiyanti Adollah1 and Rosline Hassan2

1 School of Mechatronics Engineering, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia.
2 Department of Haematology, School of Medical Science, Univeristi Sains Malaysia, Kubang Kerian, Kelantan, Malaysia.

*yusoff@unimap.edu.my

Abstract. Due to growing statistics and the important role of early diagnosis for chronic Leukemia, an automated intelligent diagnostic procedure for chronic Leukemia is needed. This paper presents an automated procedure for this diagnostic system which consisted of four main stages; namely Image Segmentation, Feature Extraction, Feature Selection and Classification. Colour Thresholding and Gradient Edge Detection were applied to segment the nucleus of white blood cells from the image. A total of 548 cells nuclei were extracted from 100 images and used to develop and validate the proposed procedure. Three feature selection algorithms and three classifiers were analyzed and evaluated in order to obtain a reliable and robust automated diagnostic procedure. Based on the results, the ReliefF algorithm was selected and applied for feature selection, the method yields the best accuracy of 98.1% for testing accuracy. Besides, the selected classifier was Multilayer Perceptron (MLP) network trained by Levenberg Marquardt (LM) algorithm.

1. Introduction

Leukocytes (or White Blood Cells) is the cell of the immune system that protects the human body by fighting infection or foreign invaders. However, Leukemia disrupts the function of blood. Generally, there are four types of Leukemia, which are Chronic Lymphocytic Leukemia (CLL), Acute Lymphocytic Leukemia (ALL), Chronic Myelogenous Leukemia (CML), and Acute Myelogenous Leukemia (AML). Generally, the family history and physical checking are the starting point in the diagnosis of Leukemia, but this is not enough to determine the presence of Leukemia for a person. Tissue biopsies taken from the bone marrow (bone marrow biopsy) or lymph nodes (lymph node biopsy) is done to look for the evidence of Leukemia.

According to statistics of the Leukemia & Lymphoma Society (LLS) [1] and MediMedia Asia (MIMS) Malaysia [2], the problem of Leukemia in Malaysia is growing, but the survivor rate for chronic leukemia is relatively high, provided it can be diagnosed and treated at the early stage [3]. In addition, World Life Expectancy [4] claimed that the death rate of leukemia in Malaysia is 4.18 that corresponding to 75 ranks over 183 countries in the world. Due to growing statistics and the important role of early diagnosis for chronic leukemia, an intelligent screening system is proposed to fulfill the demand.
2. Methodology

In general, there are four main stages involved in this study, namely Image Segmentation, Feature Extraction, Feature Selection and Classification. The general methodology was shown in Figure 1. In this section, the development process of the intelligent screening system was presented.

![General methodology flowchart of an intelligent screening system](image)

2.1. Image Segmentation

Image segmentation technique is important to segment the irrelevance and unwanted information with object of interest. The purpose of image segmentation is to provide the meaningful representation of the image to the human [5]. The techniques applied in this study were Colour Thresholding [6] and Gradient Edge Detection [7].

In Colour Thresholding, determination of the threshold value was the main task in this study. The threshold value was determined by the pixel value of RGB components for different regions, red blood cells (RBCs), nucleus of white blood cells (N) and the background (BG) with 50 arbitrarily selected points, thus a total of 2700 points will be analyzed in this section (50 points of pixel value were selected for each regions for six samples in which the pixel value consisted of RBG information).

After segmenting the nucleus from an image, Gradient Edge Detection was applied to separate the cell in the image as individual object. Selection of the Gradient Edge Detection technique was done by feeding the segmented image into different operators, included Sobel, Prewitt and Roberts. The operator which is able to provide the clearer boundary of nucleus between each other would be selected and applied in this study.

Besides, Filtering technique [8] was applied to remove the artefact and prepared the image for the next stage. In this study, the filtering range was determined by the convex area of each nucleus (for both CML and CLL) which only focused on the maximum and minimum of convex area.

2.2. Feature Extraction

Feature extraction is a stage that transform the large input data into a reduced representation set of features [9]. Different type of input will result in different type of feature. In this study, geometrical, colour and textural features were used to distinguish the CML and CLL. A total of 28 features were extracted from these categories, included area, major axis length, minor axis length, eccentricity, convex area, solidity, extent, perimeter, compactness, mean intensity, minimum intensity, mean of R, G and B components, variance of R, G and B components, standard deviation of R, G and B components, Hue (H), Saturation (S), Intensity (I), entropy, contrast, correlation, energy and homogeneity.

2.3. Feature Selection

Feature selection algorithm was used to reduce the dimension of the feature space and improve the performance of classification [10]. The feature selection algorithms that have been applied in this study were Genetic Algorithm (GA), ReliefF algorithm (RfF) and Neighbourhood Component Analysis (NCA) algorithm.
Optimization of feature selection algorithm was conducted to ensure the algorithm was fit with the data pattern. After adjusting the optimal arguments and parameters, a total of 28 features were fed into the related algorithm and recorded the results accordingly. However, with this selected features alone will never showed any effect toward the performance of the system, therefore evaluation of these selected features with a classifier was needed in which the effect can be seen by examining the accuracy of the classifier, given that the classifiers used for the evaluation are same for all.

2.4. Classification
Classification is a process that classifier is implemented to divide the feature vector into various classes based on feature similarity. In order to ensure the performance of the classifier in this system, three types of classifiers were chosen to be evaluated, which are k-Nearest Neighbour (kNN), Support Vector Support (SVM) and Multilayer Perceptron (MLP) network trained by Levenberg Marquardt (LM) algorithm.

Before the selection of a classifier, each classifier was optimized by the related optimization procedure. Since $k$-value is the factor that affect the performance of classifier, hence, $k$-value analysis was the optimization of $k$NN, the $k$-value involved included 3, 5, 7, 9, 11, 13 and 15. For SVM, the optimization was more challenge than $k$NN as the optimization procedure is longer than $k$NN. Kernel function, box constraint was the basic parameters to be optimize in SVM, if the selected SVM was under the polynomial family, the polynomial order analysis needed to organize. On the other hand, the optimization that was applied in MLP included number of hidden nodes analysis, learning rate analysis and number of epochs. The range of these optimization parameters for MLP was [1, 30], [0.1, 0.5] and [1, 20], respectively.

After optimizing the classifiers, evaluation the classifier was conducted to select the best classifier in which the evaluation parameter involved testing accuracy for both CML and CLL, the difference of testing accuracy between CML and CLL, and the performance error. Each classifier would run the relevant algorithm with the optimal setting and the evaluation parameters, as mentioned above, were recorded for analysis. A weightage scoring method was used for the evaluation of classifier in order to overcome the uneven contribution of parameter in which the mark was given by the ranking method, for instance, 3 marks was given to the classifier with the best performance and vice versa. Then, the classifier with the highest final score would be selected and implemented in this study.

3. Result and discussion
To develop an accurate and reliable diagnostic procedure for chronic Leukemia, several parameters needed to be investigated and analyzed. Hence, this chapter was discussed and presented the results of the development process. The results were presented based on the development stage in which the first section was about the determination of threshold values for Colour Thresholding and Convex Area Filtering, and the selection of Gradient Edge operators. The second section was discussed about the evaluation and selection of feature selection algorithm. The last section of this chapter presented about the optimization of each classifier, and selection of the best classifier for the screening system.

3.1. Image Segmentation
From the pixel analysis, the threshold value can be set from the G component as it showed a gap between nucleus of WBCs, RBCs and background regions, while other component values (R and B components) were overlapped between each other. After the analysis, it was found that the nucleus of WBCs candidate threshold value for CML was 115 and 117 while CLL was within the range of 118 and 120 in which the common point of threshold value did not exist. Although the threshold value of Colour Thresholding for CML and CLL was different, the average between the ranges was able to be determined. Hence, 117 was selected for Colour Thresholding.
Another threshold value was determined by the mean of total value of RGB components for a pixel. This was used to improve the performance of segmentation. Based on the result, the threshold value will be set between 300 and 450 as the mean value for total RGB components was always maintain in this range.

Selection of Gradient Edge operator was conducted after the determination of these threshold values. Based on result, the Roberts operator was selected and implemented in this system as it able to produce low noise image.

After that, the determination of filtering range for Convex Area Filtering was conducted which only focused on the minimum and maximum value of convex area for both CML and CLL. Based on the collected data, the minimum and maximum value was 2520 and 8803, respectively, hence, the filtering range was set between 2500 and 9000.

3.2. Feature Extraction
A total of 28 features was extracted successfully that included three categories: geometrical, colour and also textural. The geometrical category were area, major axis length, minor axis length, eccentricity, convex area, solidity, extent, perimeter and compactness, while the colour feature were mean intensity, minimum intensity, mean of R, G, B component, variance of R, G, B component, standard deviation of R, G, B components and HSI. Entropy, contrast, correlation, energy and homogeneity were extracted for the textural category.

3.3. Feature Selection
Although only three feature selection algorithms were involved, five set of features have been analyzed as RfF produced three different set of features at three different cut off weights, namely RfF group 1, RfF group 2 and RfF group 3 with the cut off weights of 0.080, 0.035 and 0.028, respectively.

Based on Figure 2, RfF group 1 exhibited the lowest average accuracies, this is because the number of features involved in RfF group 1 was only two features that caused the underfitting phenomena to occur. On the other hand, the RfF group 3 exhibited the highest average accuracies, therefore, the feature set of RfF group 3 was selected and applied for the following analysis, included area, minor axis length, convex area, perimeter, compactness, mean intensity, minimum intensity, mean of R component, variance of R, G, B components, standard deviation of R, G, B components and energy and correlation.

3.4. Classification
In kNN, the optimization parameter was only focused on the k-value which vary from 1 to 15 with the increment of 2 and testing accuracy was the optimization target. Based on the result, the k-value of 1 was selected as it achieved the highest testing accuracy that was 91.5%.

For SVM, the optimization involved were kernel function and box constraint. The result showed that the quadratic SVM, which was under the polynomial family of SVM, was selected as it exhibited the highest accuracy (98.2%). Based on the selected kernel function, box constraint analysis was conducted. The result (as shown in Figure 3) showed that the box constraint of 0.001 achieved the highest testing...
accuracy (100%), thus 0.001 was applied in SVM. Since the selected SVM was under the SVM family of polynomial, hence optimization of polynomial order was also conducted. From the Figure 4, the polynomial order of 3 was selected in this study as it achieved the highest testing accuracy (100%) and less computational complexity.

Number of hidden node analysis, learning rate analysis and number of epochs analysis were the optimization procedure for MLP network classifier. The evaluation target for these three analysis were testing accuracy and Mean Square Error (MSE) that expressed in 10^−6. The optimization was started with number of hidden nodes analysis and followed by the learning rate analysis and then number of epochs analysis. Figure 5, Figure 6 and Figure 7 were the results for the optimization, respectively.

Based on Figure 5, the highest testing accuracy were achieved by 1, 15 and 20 hidden nodes, however 20 hidden nodes was selected as the MSE was the lowest. According to the results in Figure 6, it can be concluded that the learning rate of 0.1 was selected since it yields the highest accuracy and the lowest MSE. As this analysis was used to avoid the overfitting of a classifier, hence the starting point of the constant trend (as shown in Figure 7) would be selected. Based on this criteria, the 10 epochs was selected and implemented in this system.

Evaluation and selection of classifier was conducted after the optimization of classifier. This weightage scoring method was used to overcome the problem of uneven contribution of evaluation parameter. The result of scores and the weightage distribution were shown in Table 1. Table 1 presented the highest score in this analysis was MLP with score of 3.0 while ANN only score 1.0 and SVM was score of 2.0. Hence, MLP trained by LM algorithm was the selected classifier.
Table 1. The summary of scores and weightage distribution

| Evaluation Parameter                                      | Weightage | ANN Value | ANN Score | SVM Value | SVM Score | MLP Value | MLP Score |
|-----------------------------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Testing Accuracy (both CML and CLL)                      | 0.5       | 0.5       | 91.5%     | 1         | 98.8%     | 2         | 100.0%    | 3         |
| Difference Testing Accuracy (between CML and CLL)        | 0.3       | 0.3       | 18.9%     | 1         | 2.7%      | 2         | 0%        | 3         |
| Error (10^-6)                                            | 0.2       | 0.2       | 1.0891    | 1         | 1.0123    | 2         | 1.0000    | 3         |
| Total                                                    | 1.0       | 1.0       | 1.0       | 2.0       | 2.0       | 3.0       |

4. Conclusion
In conclusion, the study was able to propose a diagnostic system to classify the chronic leukemia slide images by analyzing the stage of image segmentation, feature extraction, feature selection and classification by LM algorithm. The proposed method was able to produce a testing accuracy of 100% with 1.0 x 10^-6 error which able to help the hematologist eliminate the step of screening of microscopic slide in the conventional diagnosis system. However, this diagnostic system is only software part of the overall system, therefore, a further study could focus on the development a fully automated diagnostic system that enable the system to be interfaced and controlled the microscope and digital camera.

References
[1] LaRussaA, “Facts and Statistics,” Leukemia and Lymphoma Society, 03-Mar-2015. [Online]. Available: http://www.lls.org/http://llsorg.prod.acquia-sites.com/facts-and-statistics/facts-and-statistics-overview/facts-and-statistics. [Accessed: 06-Oct-2018].
[2] V. J. Khoo, “Health DG: Malaysian National Cancer Registry Report 2012 – 2014 in progress,” MIMS News. [Online]. Available: https://today.mims.com/health-dg--malaysian-national-cancer-registry-report-2012---2014-in-progress. [Accessed: 06-Oct-2018].
[3] Leukemia: Risk Factors, Symptoms, Types, and Treatment,” Healthline. [Online]. Available: https://www.healthline.com/health/leukemia#. [Accessed: 06-Oct-2018].
[4] “Leukemia death rate by country,” World Life Expectancy. [Online]. Available: https://www.worldlifeexpectancy.com/cause-of-death/leukemia/by-country/. [Accessed: 06-Oct-2018].
[5] A. Garg 2016 A Review on Image Segmentation Techniques International Journal of Recent Research Aspects (Conscientious and Unimpeachable Technologies 2016) 53–55
[6] R. Adollah, M. Y. Mashor, H. Rosline, and N. H. Harun 2012 Multilevel Thresholding as a Simple Segmentation Technique in Acute Leukemia Images Journal of Medical Imaging and Health Informatics 2 285–288
[7] S. Shubhashree 2014 A Review on Edge Detection Techniques for Image Segmentation International Journal of Computer Science and Information Technologies 5 5898–5900
[8] S. Eddins, “Binary image area filtering,” Steve on Image Processing and MATLAB, 22-Aug-2016. [Online]. Available: https://blogs.mathworks.com/steve/2016/08/22/binary-image-area-filtering/. [Accessed: 10-Oct-2018].
[9] K. Chinnathambi, A. Ramasamy, and P. Rajendran 2014 Modified segmentation algorithm and its feature extraction of cancer affected white blood cells 2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014] 1202–1210
[10] Z. Wang, Y. Zhang, Z. Chen, H. Yang, Y. Sun, J. Kang, Y. Yang, and X. Liang 2016 Application of ReliefF algorithm to selecting feature sets for classification of high resolution remote sensing image IEEE International Geoscience and Remote Sensing Symposium 755–758