The Detection and Classification of blast cell in Leukaemia Acute Promyelocytic Leukaemia (AML M3) blood using Simulated Annealing and Neural Networks

Waidah Ismail[1,3], Rosline Hassan[2], Annette Payne[1], Stephen Swift[1]

[1] Department of Information System and Computing, Brunel University, West London, UB8 3PH, United Kingdom
{waidah.ismail, annette.payne, stephen.swift}@brunel.ac.uk
[2] Haematology Department, Universiti Sains Malaysia, Kelantan, Malaysia
roslin@kb.usm.my
[3] Faculty Science and Technology, Universiti Sains Islam Malaysia, Malaysia
waidah@usim.edu.my

Abstract. This paper presents a method for the detection and classification of blast cells in M3 with others sub-types using simulated annealing and neural networks. In this paper, we increased our test result from 10 images to 20 images. We performed Hill Climbing, Simulated Annealing and Genetic Algorithms for detecting the blast cells. As a result, simulated annealing is the "best" heuristic search for detecting the leukaemia cells. From the detection, we performed features extraction on the blast cells and we classifying based on M3 and other sub-types using neural networks. We received convincing result which has targeting around 97% in classifying of M3 with other sub-types. Our results are based on real world image data from a Haematology Department.

Keywords: Heuristic Search, Simulated Annealing, Classification, Leukaemia cells,

1 Introduction

In 1895, Wilhelm Roentgen discovered that X-rays tubes were used extensively for imaging bones and then for treating a variety of conditions. The technicians who ran the X-ray machines as well as many of the exposed patients contracted skin cancer tumours and leukaemia. The changes such as lifestyle, diet and environments are the main contributions of increasing the number of people are acquiring by cancers [1,2]. Haematologists have concluded that there are series of investigations performed to diagnose acute leukaemia such as microscope colour image, segmentation and classification or clustering by count of leukaemia cells by using microscopes analysis of human blood smears. [2,3]. Recently, there are several image processing application that has been developed in collaboration between research groups and clinicians to extract the useful and fast information for the diagnosis especially on the cancer areas. [4,5]. At the same time, contribution of computer software is not limited to image processing but it is extends to analysing and tracking on the drugs recovery or chemotherapy. It has been proven successful in treating cancer [6].

This paper aims to help the medical doctors to detect and classify AML M3. We employ new techniques based on the "best" heuristic method using similarity matrix and neural network which can reduce the time taken by a haematologist in detecting and classified of leukaemia cells.

This paper is organised as follows: in the rest of this section we detail the motivation behind our paper; in section 2 we describe previous work in the area, section 3 details our proposed methodology for detecting and classifying leukaemia cells and section 4 explains in detail our methods. The data sets and experiments are presented in section 5 and are followed by section 6 in which we discuss the results. Lastly, in section 7, we draw conclusions and discuss future research.

1.1 Leukaemia cells

Blood cancer or clinical called as Leukaemia is a disease which has unknown cause where the bone marrow produces large numbers of abnormal cells [7]. According to a paper published [8], the analysis of counting of white blood cells via microscope image can provide very useful information of the patient related to the health of the patient. When viewing the image of white blood especially for the diagnosis of leukaemia, there will be problems to identify the blast cells if the blast is minimal in number or the staining of cells are poor in addition it is viewed by an inexperienced morphologist, the diagnosis might be delay or incorrect. From this problem, it can also lead to time consuming and tiredness in viewing the blood cells.

Acute myelogenous leukaemia (AML) is a serious illness caused by the abnormal growth and development of early nongranular white blood cells. AML begins with abnormalities in the bone marrow blast cells that develop to form granulocytes, the white blood cells that contain small particles, or granules. The AML blasts do not mature, and they become too numerous in the blood and bone marrow. As the cells build up, they hamper the body's ability to fight infection and prevent bleeding. Therefore, it is necessary to treat this disease within a short time after making a diagnosis. The recognition of the blast cells in the bone marrow of the patients suffering from myelogenous leukaemia is a very important step in the recognition of the development stage of the illness and proper treatment of the patients. The clinicians have to identify these abnormal cells under microscope in order to suspect that patient is having leukaemia which they need patient's bone marrow to count the leukaemia cells(blast cell) to confirm the diagnosis [9,10].

There are few subtypes of AML which are classified as M0 to M7 according to French-American-British (FAB) classification. The blast cells of these subtypes are different in size, shape, amount of cytoplasm, shape and amount of nucleus and the constituent in the cytoplasm. It is important to identify certain subtype such as M3 (Acute Promyelocytic Leukaemia) because the treatment is different from other subtypes and it is good prognostic index for the patient. Early diagnosis is also required as this will improve patient's survival. Currently in addition to blast cell morphology, the confirmation of the diagnosis is by genetic testing which are time consuming and expensive. However, sometimes morphologist also have problem to classify this subtype especially the hypogranular variant of M3 and resulted in misclassification. Inter-observer variation in classification can lead to improper

treatment for the patient. The counting procedure should cover 100 immature cells in bone marrow order to diagnosis if someone has leukaemia. [8]. As shown in the figure 1 are the photographs picture of M1 to M7 sub-types of AML.

2 Previous Work

In this paper, we have extended our research [12, 13] by finding the final coordinate are located at the same location which we called as similarity matrix. This method we used to choose the "best" heuristic search and multilayer perceptron which is one of the feed-forward artificial neural networks used in classification

There are a few papers on detecting the white blood cells but not blast cells in leukaemia using image processing. Most of the papers are based on morphological techniques. Previous work in [14] has concentrated on using the nucleus for classification through image segmentation. In [15, 16] a colour gradient method was used before the application of a GVF (Gradient Vector Flow) snake and scale-space filtering with watershed algorithm on colour images for detecting nuclei. In [17] Genetic Algorithms are used to optimise the parameters of an elastic contour model using edge information for automating the segmentation of the prostate. Another technique using eigen cells for detecting white blood cells was introduced in [18] but this study had limited success in correctly classify all of the white blood cells. However, in the case of overlapping cell there can be some problems. In [19] a histogram of pixel counts focusing on the touching cells was created and an edge cutting algorithm was then applied to separate the cells. This technique can be used for touching cells but not for overlapping cells. The technique is not possible to use in this project because cutting the cells will create different morphological features which might lead to incorrect classification.

From our previous research which published in [12,13] we have stated two work process which are random and seeded in finding the "best" Heuristic Search method using Hill Climbing, Simulated Annealing and Genetic Algorithm in detecting blast cells in leukaemia. In the work process for random experiment, we only have three steps which are Otsu, random coordinate as starting point and heuristic search.

But in work process seeded experiments, the process involves Otsu's method, running the Cellular Automata (CA), detecting the starting coordinates, colour image clustering and coordinate overlap checking. From the previous result, we using 10 images but there are drawback in detecting the coordinate and clustering the images which only target 7 out of 10 images. We found that Simulated Annealing Seeded have highest fitness functions. In this research, we have enhanced our drawbacks in detecting the coordinate and clustering the images and increased our images.

We performed image classifications in this paper to classify blast cells in leukaemia for M3 with other subtypes AML. In the paper [18], they use three combinations which are shape and appearance, automatic selections and random forests using grid in the images as region of interest. We cannot use grid because the leukaemia cells might divided into two sections when we grid it. This can lead to wrong diagnosis. In our research, we need to the whole of leukaemia cells for better diagnosis.

There are also techniques using Fuzzy Relative Information Measure [19], they using discretisation algorithms which to discretise numeric features using K-Mean, Fuzzy C Means and Median as the initial centriod of K-Mean. In the paper [20], using the difference between the transform coefficient histogram of compressed and uncompressed images which the difference of the peaks of histogram characterize using the distance metric. We have difficulty in performing compressed and uncompressed which the information may lose during the uncompressed and can lead to different sub-types of leukaemia.

3 Work Process

In this paper, we extended our research into in choosing the "best" heuristic search and classification of leukaemia cells. We have explained our previous steps in [12,13] which we used Otsu, cellular automata, finding coordinate, colour image clustering, . The figure 2 shows the flowchart of our research methodology in similarity matrix and classification.

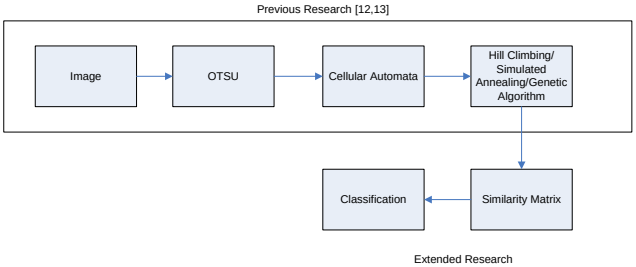

Fig. 2. Research Methodology Flowchart

4 Method

In the following sections we present the methods that we use which are: coordinate detection, image clustering, coordinate's similarity and the classification of blast cells in leukaemia between M3 and other sub-types.

4.1 Detecting Coordinates

After we performed Otsu and Cellular Automata which have been explained in [12,13], then we performed detecting coordinates. In the previous research, we just draw a rectangle at the seed from cellular automata result. It may have extra a few pixels which create an extra unnecessary coordinate. In this research, we did an enhancement by performing shrink rectangle to ensure that they are no extra pixel at the seed from cellular automata result. In the shrink rectangle, it performs iteration until they meet the 1st pixel of seed area.

4.2 Colour Image Clustering

We performed the colour image clustering by creating a simple partition of the RGB data based on the image detection. For pink which are plasma and red blood cells are R = 132, G = 81 and B = 79. Purple which define as leukaemia cell has R = 103, G = 15 and B = 81. Then we find based on the nearest RGB to purple, then it classify as blast cells.

4.3 Heuristic Search

We executed the program for Simulated Annealing, Hill Climbing and Genetic Algorithms after enhancing the detecting coordinate and colour image clustering. We performed the heuristic search to locate the location of leukaemia cells before we proceed to classification.

4.3.1 Fitness Function for Simulated Annealing and Hill Climbing

A fitness function is particular type of objective function that prescribes the optimaly solution in heuristic search. In the equation 1, we define C as our list of circles, C_i as the ith circle and $|C|$ as the number of circles within C. $R(i)$ is the radius of circle i, $B(i)$ is the number of black points within circle i for a given image, $W(i)$ is the number of white points for a given image. The equation 2 stated in [12]:

$$F(C) = \frac{R(C_i)\left(\sum_{i=1}^{|C|} B(C_i)+1\right)}{\sum_{i=1}^{|C|} W(C_i)+1} \quad (1)$$

4.3.2 Ratio/Similarity Matrix

We performed ratio to know how much two circles are overlap based on the Figure 3. We also used Figure 3 method as similarity matrix to find out the heuristic search location at the same position all the times as final results.

a) 1st scenario Circles overlap totally

b) 2nd scenario Circles do not overlap at all

c) 3rd scenario Circles union

d) 4th scenario Circles intersect

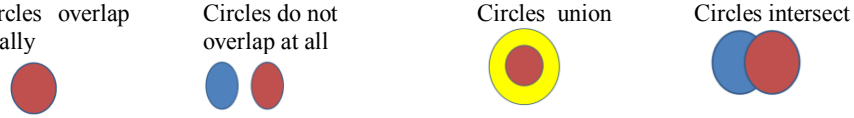

Fig. 3. Overlapping Circles [13]

4.3.3 Fitness Function for Genetic Algorithms

In the Genetic Algorithms, we introduced a measure of how much two circles overlap. We used this information to ensure that a small change (an integral part of In the Genetic Algorithm we introduced ratio to ensure that a small change (an integral part of hill climbing and simulated annealing) did not create any overlap. However, this was computationally expensive process and affected the performance of the algorithms. We integrate this calculation into the fitness function, which means that the Genetic Algorithm should produce non-overlapping solution. [13]

$$F(C) = \frac{R(C_i)\left(\sum_{i=1}^{|C|} B(C_i)+1\right)}{\sum_{i=1}^{|C|} W(C_i)+1} /(1+\text{ratio}) \quad (2)$$

4.5 Classification

We have detected the location of blast cells then we perform the feature extraction and classification by using WEKA application [21]. Weka is a collection of machine learning algorithms for data mining tasks which contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. From the feature extraction, we partitional by creating a simple partition of the data separated by RGB to find the mean, median, variance and add in the extra requirement by separated the individual RGB into the average high and low.

5 Data Sets and Experiments

In this research we have increased our testing from 10 to 20 images of leukaemia cells (AML). The images were 1280 by 960 in size as in Figure 4. We executed each of the seeded for 20 times for 10000 iterations for Hill Climbing and Simulated Annealing and 500 iteration and 100 population for Genetic Algorithms; the repeats were so that any sampling bias is removed from the stochastic nature of the algorithms for test data. After we have chosen the “best” heuristic search, then we executed for 322 images of real images data sets from Hospital University Science Malaysia (USM), Kota Bahru, Kelantan, Malaysia.

6 Results

6.1 Detecting Coordinates

From the previous result [12,13], we only performed 7 out of 10 images. In this paper, it targeting 20 images which included 3 images is not targeting from our previous research which is 100%.

6.2 Colour Image Clustering

We clustered the pixels using Euclidean distance. By using Euclidean distance we can determine whether a candidate cell is purple (a leukaemia cell (AML)) or pink (either plasma or red blood cells). We have increased of 5% from our previous research which are 91.5% stated in [12] and new target is 96.27%.

6.3 Hill Climbing, Simulated Annealing and Genetic Algorithms

From the below result, Simulated Annealing has targeting the highest average of fitness function and Hill Climbing is the second best in the fitness function Surprisingly, Genetic Algorithm still did not targeting the best average fitness function as show in Table 1.

Table. 1. Result fitness function for Hill Climbing, Simulated Annealing and Genetic Algorithms

Image	HCSeed	SASeed	GASeed
Average	1885876.873	**1891374.73**	1561045.7

6.4 Similarity Matrix

We executing the similarity matrix based on final coordinate result for 20 images for 20 times execution in heuristic search. We only performed for Simulated Annealing and Hill Climbing based on the average in the Heuristic Search. It shows that Simulated Annealing has highest average and standard deviation and decided that Simulated Annealing is the "best" heuristic search in located blast cells in. Then, we performed 322 of real images using Simulated Annealing.

Table. 2. Result from Similarity Matrix for Simulated Annealing and Hill Climbing

Method	Simulated Annealing		Hill Climbing	
	Average	Standard Deviation	Average	Standard Deviation
	0.964	0.772	0.963	0.758

6.5 Classification

After we have detecting as show in the Figure 4, then we extracted the leukaemia cell as in the Figure 5. We performed feature extraction based on Mean, Median, Variance and High and Low.

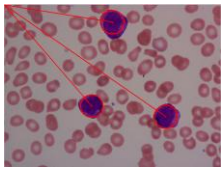

Fig. 4. Real Image of M3

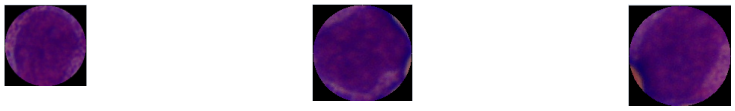

Fig. 5. Feature extraction of M3

From the feature extraction, we have tested using a few neural network such as Bayesian Logistic Regression (78.43%), Mutilayer Perceptron (92.15%), DecisionTable (62.74%), Random Tree (76.47%) and KStar (33.33%). The above results are based on the 20 images in clustering of M1, M2, M3 and M5.

From the result above, we choose Multilayer Perceptron and performing 322 images with 648 feature extraction of images. We have performed clustering for M3 and others because the treatments for M3 are different from other subtypes and it is good prognostic index for the patient. We received convincing results of 97.22% for 630 images and 2.77% for 18 images for misclassification. Below Table 5 show the summary of the result on Multilayer Perceptron with 10 cross validation

Table. 3. Summary of the result of Multilayer Perceptron

Classify	Test Options		
	Multilayer Perceptron		
Correctly Classified Instances		630	97.2222%
Incorrectly Classifier Instances		18	2.7778%
Kappa Statistic		0.9233	
Total Number of Instances		648	
Detailed Accuracy by class		Roc Area	Class
		0.99	0
		0.99	3
Weighted Average		0.99	
Confusion matrix	a	b	Classified as
	485	8	a = 0
	10	145	b=3

7 Conclusions

In this paper, we have extended our research by using more images, enhance previous research drawback, performing similarity matrix and classification. We have increasing our test images from ten to twenty images for finding the "best" heuristic search. We have performing the seeded method which consists of Otsu's, Cellular Automata seed generation, coordinate location, colour image clustering, and heuristic search for twenty images.

Then we extract the leukaemia location and performed feature extraction of the leukaemia located. We performed tested for twenty images in a few methods in WEKA as Bayesian Logistic Regression , DecisionTable, Multilayer Perceptron, etc. We found out that Multilayer Perceptron has the highest classification for twenty images. The we finalised using Simulated Annealing seeded for detection leukaemia cells and Multilayer Perceptron as classification for 322 images of real data. We received around 97.22% for real data image to classify between M3 and other sub-types.

As for the main contribution of this paper, we provide the first steps towards helping the haematologist's detection and classification of Leukaemia cells and using new techniques in Simulated Annealing and Neural Network. However, our future work, we still need to improve our research by classifying M1 to M7 of leukaemia cells.

Acknowledgments. We would like to thank the Universiti Sains Islam Malaysia and the Ministry of Higher Education, Malaysia for supporting this research and providing the scholarship.

References

1. King, R. J. B. & Robins, M. W. (2006) *Cancer Biology*, Pearson Prentice Hall.
2. Weinberg, R. (1998) *One Renegade Cell, The Quest for the Origins of Cancer,* London Weidenfeld & Nicolson.
3. Uthman, E,: Blood Cells and the CBC. http://web2.airmail.net/uthman/blood_cells.html(2000)
4. Poulsen, R.S. & Pedron, I (1995). Region of Interest Finding in Reduced Resolution Color Imagery – Application To Cancer Cell Detection in Cell Overlaps and Clusters. *IEEE-EMBC and CMBEC*.
5. Nilsson, B. & Heyden, A. (2002) Model-based Segmentation of Leukocytes Clusters. *IEEE*
6. Chen, X., Zhou, X. & Wong, S. T. C. (APRIL 2006) Automated Segmentation, Classification, and Tracking of Cancer Cell Nuclei in Time-Lapse Microscopy. *IEEE Transactions On Biomedical Engineering,* Vol. 53, , 762-766.
7. Green, R. (01.05.2005) Acute leukaemia.
8. Hassan, R.: Diagnosis and outcome of patients with Acute Leukemia. *Haemotology department.* Malaysia, Universiti Sains Malaysia.(1996)
9. Markiewicz, T., Osowski, S., Marianska, A, B. & Moszczyn'ski, L. (July 31 - August 4, 2005) Automatic Recognition of the Blood Cells of Myelogenous Leukemia Using SVM. *International Joint Conference on Neural Networks,*. Montreal, Canada, IEEE
10. Yi, F., ChongXun, Z., Chen, P. & Li, L. (September 1-4, 2005) White Blood Cell Image Segmentation Using On-line Trained Neural Network. *Engineering in*

Medicine and Biology 27th Annual Conference. Shanghai, China, Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference.

11. Bell, A. & Sallah, S. (2005) *The Morphology of Human Blood Cells*, Abbott Ltd, USA.
12. Ismail, W., R. Hassan, et al. (2010). Detecting Leukaemia (AML) Blood Cells Using Cellular Automata and Heuristic Search. *Intelligent Data Analysis (IDA) 2010*, Tuscan, USA, Springer-Verlag Berlin Heidelberg 2010.
13. Ismail, W., R. Hassan, et al. (2010). Detecting Leukaemia (AML) blood cells using Genetic Algorithms. *2nd Student conference on Operational Research(SCOR 2010)*, Nottingham, UK
14. Puiri, V and Scott, F.: Morphological Classification of Blood Leucocyes by Microscope Image. *CIMSA 2004 - IEEE lntemational Conference on Computationvl Intelligence far Memrement Systems and Applications.* Boston, MA. USA, IEEE. (2004).
15. Zamani, F. & Safabakhsh, R. :An unsupervised GVF Snake Approach for White Blood Cell Segmentation based on Nucleus. *8th International Conference on Signal Processing, ICSP* Guilin, China.(2006)
16. Jiang, K., Liao, Q. G. & Dai, S.-Y. A Novel White Blood Cell Segmentation Scheme Using Scale-Space Filtering And Watershed Clustering. *Second International Conference on Machine Learning and Cybernetics.* Xi'an, Proceedings of the Second International Conference on Machine Learning and Cybernetics.(2003).
17. P. Ghosh and M. Mitchell (2006) Segmentation of Medical Images Using a Genetic Algorithm. *GECCO' 06.* Seattle, Washington.
18. Bosch, A., A. Zisserman, et al. (2007). Image Classification using Random Forests and Ferns. Computer Vision 2007 ICCV Rio de Janeiro, Brazil, IEEE 11th International Conference.
19. K.Sarojini, D. K. Thangavel, et al. (2010). "Supervised Feature Subset Selection based on Modified Fuzzy Relative Information Measure for classifier Cart." International Journal of Engineering Science and Technology (Vol 2(5)): 2456-2465.
20. Tjoa, S., W. S. Lin, et al. "Transform Coder Classification for Digital Image Forensics."
21. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.