Detection and classification of the Acute Myeloid Leukemia cells in the images of white blood cells

TRAN VAN NHAN

School of Information Science
Japan Advanced Institute of Science and Technology
September, 2016

Master's Thesis

Detection and classification of the Acute Myeloid Leukemia cells in the images of white blood cells

1410218 TRAN VAN NHAN

Supervisor : Associate Professor Atsuo Yoshitaka
Main Examiner : Associate Professor Atsuo Yoshitaka
Examiners : Professor Masato Akagi
Professor Jianwu Dang

School of Information Science
Japan Advanced Institute of Science and Technology

August, 2016

Contents

.

This dissertation was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Vietnam National University, Ho Chi Minh City.

List of Figures

List of Tables

Acknowledgements

First, I would like to thank my supervisor, associate professor Atsuo Yoshitaka of the School of Information at Japan Advanced Institute of Science and Technology. I would like to express my sincere gratitude to the support of my master study, for his patience, motivation, enthusiasm. His guidance helped me in all the time of my study.

Second, I would also like to thank all the students belonging to the Yoshitaka's lab and the member of Vietnamese community, who supported me about the study and the life in the first time I came to Japan.

Finally, I must express my very profound gratitude to my parents and to my girlfriend for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of studying and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Author

Tran Van Nhan

Chapter 1

Introduction

In recent decades, image processing has been applied to many applications of life. In particular, imaging applications are emerging as a new opportunity for innovation at the meeting point between medicine and computer science. Many software and research groups focus on the development of image processing applications for medical images. Collaboration with clinicians has allowed the extraction of useful information contributing to more efficient diagnosis. By the help of image processing, we can extract many useful information from medical images in order to assist and improve patient diagnosis especially in the cancer area.

1.1 Background

Blood is fundamental to human life. Biologically, blood is essential for maintaining homeostasis, that is keeping the bodys status stable. So, this is a very important of our body. Blood has four main elements to ensure that completing its functions.

- Red blood cells

  Transport oxygen from the lungs to organs and peripheral site.

- White blood cells

  Defensive role in destroying invading organisms and assist in the removal of dead or damaged tissue cells.

- Platelets

  Assist in the clotting process

- Plasma

  Carries nutrients, metabolites antibodies and the proteins involved in blood clotting

White blood cell composition and concentration in the blood gives valuable information and plays a crucial role in the diagnosis of different diseases. White blood cells fall into five categories: Neutrophil, Eosinophil, Basophil, Monocyte and Lymphocyte. These cells

provide the greatest defense against infections, and their individual concentrations can help specialists to distinguish between the presence or not of severe pathologies.

Leukaemia was recognized in 1845, by Bennett in Scotland and Virchow. Leukemia is a cancer of white blood cells, where the disease basically develops in the bone marrow, which is the spongy tissue that fills the inside region of the bones. Leukaemia is a disease of unknown cause where the bone marrow produces large numbers of abnormal cells white blood cells that stop developing before maturity. Most acute leukaemia patients are referred to specialist units for evaluation and treatment. Leukaemia can be diagnosed by blood tests and bone marrow tests. Currently the best available treatment is chemotherapy, which unfortunately also kills normal body cells along with the cancerous ones. A bone marrow and a blood test can help determine the different leukaemia types, allowing doctors to decide on the best choice of treatment.

Blood cells are made in the bone marrow by stem cells. The bone marrow is the soft material in the centre of bones. Diagram in Figure 1.1 shows the development process of AML cells. In the diagram, stem cells are immature (primitive) cells. There are two main types in the bone marrow: myeloid and lymphoid stem cells. Stem cells constantly divide and produce new cells.

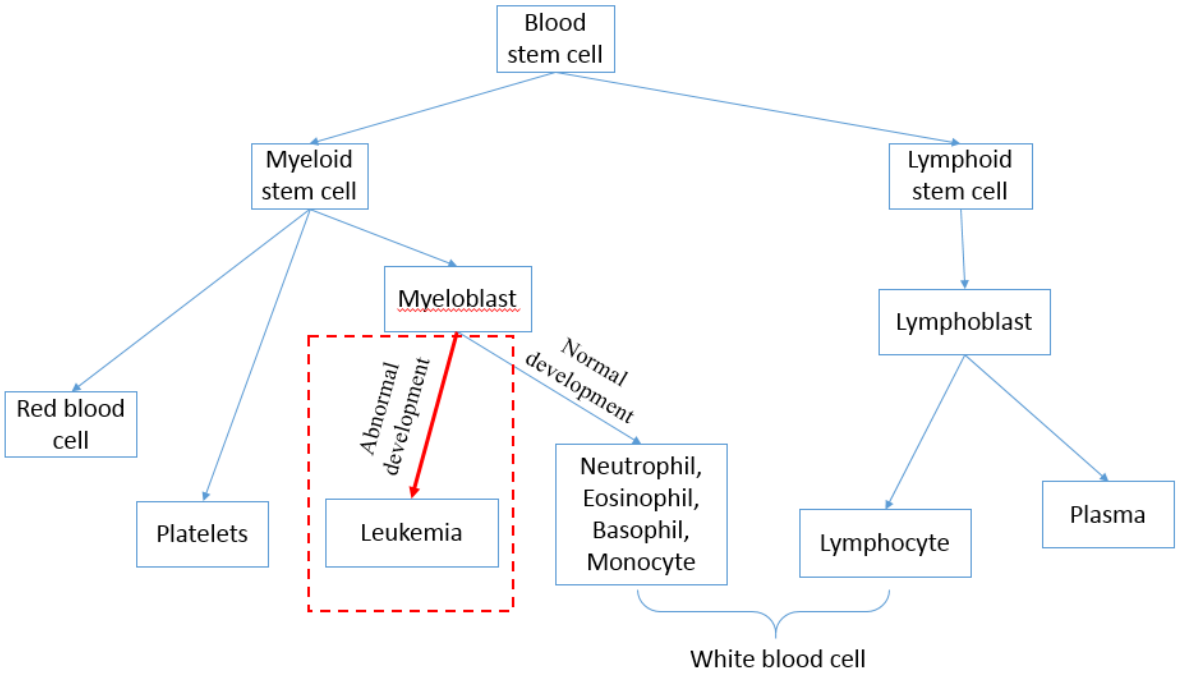


Figure 1.1: Diagram of blood cell and leukemia cell production [1]

There are four major different forms or types of leukemia, which develop in cancer patients according to the growth speed and the improper overproduction of leukemia cells: acute lymphocytic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL), chronic myeloid leukemia (CML) [2]. This study will focus on AML type. Most commonly, acute leukaemia patients are referred to specialist units for evaluation. Treatment is based on chemotherapy through the veins, lasting four to six months, which also kills normal body cells. Leukaemia can be diagnosed by blood tests while a bone marrow test serves to decide on the best choice of treatment. The descriptions of each type were described as below [1].

- Acute Lymphoblastic Leukaemia (ALL)

  It occurs mainly in adults. A very small number will infects the children.

  The survival rates vary by age: 85% in children and 50% in adults.

- Acute Myeloid Leukaemia (AML)

  It develops in both adults and children.

  The five-year survival rate is 40%.

- Chronic Myeloid Leukaemia (CML)

  It occurs mainly in adults. A very small number will infects the children.

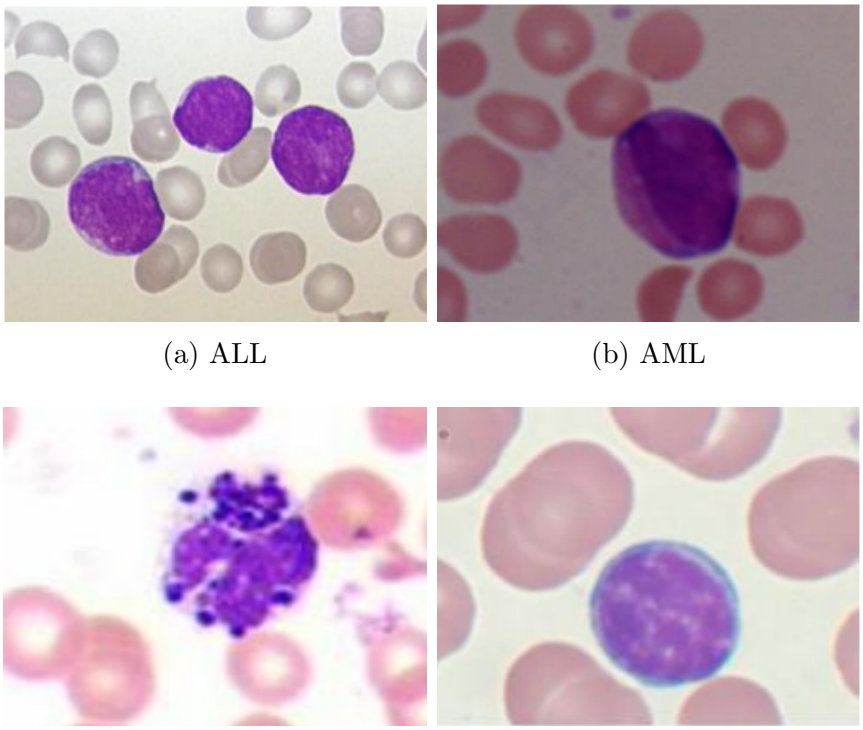
  The five-year survival rate is 90%.

- Chronic Lymphocytic Leukaemia (CLL)

  Most commonly it affects adults over the age of 55. It sometimes occurs in younger adults, but it almost never affects children.

  The five-year survival rate is 75%.

AML is a general form of acute leukaemia that is increasingly common with progressing age but may occur in all age groups. Acute myeloid leukemia is the second frequent type (after acute lymphocytic leukemia) of leukemia diagnosed in children. About 15% of children from birth to 19 years of age diagnosed with leukemia have acute myeloid leukemia. The risk of acute myeloid leukemia is closely associated with age. About 90% of acute myeloid leukemia is diagnosed in middle age. Incidence rate of acute myeloid leukemia is high in men and women over 50 years.



(a) ALL (b) AML

(c) CML (d) CLL

Figure 1.2: The examples of the leukemia subtypes [1]

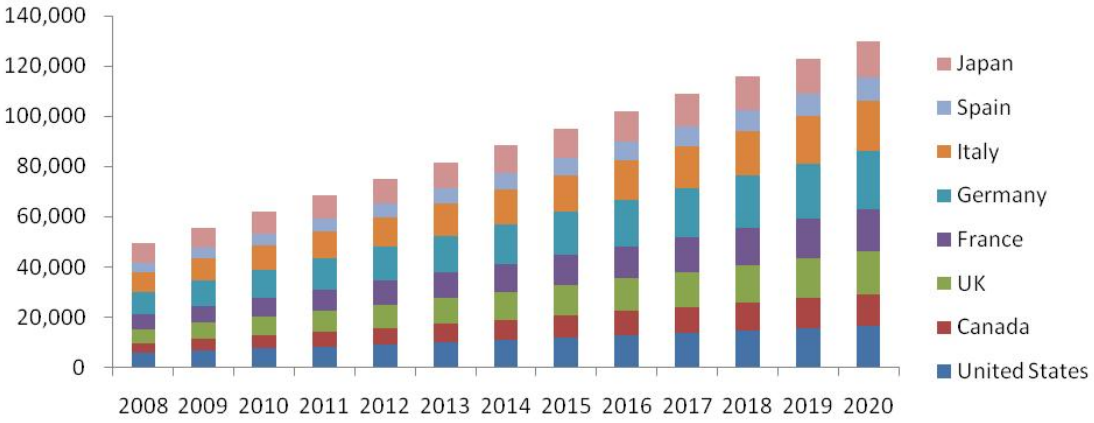


Figure 1.3: Acute myeloid Leukemia diagnosed population [3]

This thesis focuses on AML, a serious illness caused by the abnormal growth and development of early nongranular white blood cells. In AML, the bone marrow makes large numbers of abnormal immature white blood cells. The immature cells are called as blast cells. The recognition of the blast cells in the bone marrow of the patients suffering from myeloid leukemia is a very important step. It is followed by categorizing into subtypes which will allow the proper treatment of the patients. In 1971, the diagnosis of leukemia cells was based on the morphology [4].

There are two classification schematas for AML, the French-America-British (FAB) and the World Health Organization (WHO) system. There are 8 subtypes of AML according to the FAB system. The diagnosis of AML subtypes was based on the morphological analysis of peripheral blood and bone marrow [4]. The FAB classification relied on appearance of leukemia cells under the microscope after routine staining. They classified AMLs into subtypes from M0 to M7. This was based on the type of cell from which the leukemia developed and the level of maturity of the cells.

The treatment of AML is primarily based on the use of intensive chemotherapy. This is usually administered in four or five doses, each approximately one week apart and the most commonly used drugs include cytosine arabinoside, daunorubicin, idarubicin, 6-thioguanine, mitoxantrone or etoposide [1].

All AML subtypes (FAB M0 M7) are treated similarly except for the promyelocytic M3 variation associated with the t(15;17) translocation in which ATRA is added to the initial chemotherapy. Since the drugs are myelotoxic, with limited selectivity between leukaemia and normal cells, there is considerable risk of marrow failure, thus prolonged and intensive supportive care is required. A key concept developing in AML therapy is that of basing the treatment program of individual patients on their risk status [1].

1.2 Related works

From the images of blood smear microscope slides, the researchers have to detect and locate the AML cells. The image segmentation algorithms were used to solve this problem. There are many methods which were applied to segment the cells. However, they can be divided into 5 small methodologies, which are thresholding-base, region-based, edge-base, clustering-base and other methods.

- **Thresholding-base**

  Thresholding-based is a group of image techniques. Thresholding is the simplest method of image segmentation and usually applied for detecting the nuclei of blood cells.

  Patil, et al. [5] used the median filtering to remove the noise component in blood image, then the RGB images were converted into HSV color space and applied the Otsu's method. This method could not segment the cytoplasm of white blood cells.

  In [6], the pixel-based segmentation via bi-modal thresholding was presented. M.D. Joshi et al.[7], proposed automatic Otsus threshold blood cell segmentation method along with image enhancement and arithmetic. Both of above approaches were used for segmentation ALL cells, which belong to Acute Leukemia type.

  In [8], Putzu et al. presented a complete and fully automated method for WBC identification and classification. This approach isolated the whole leucocyte and then separates the nuclei and cytoplasm by using the color conversion and Zack's threshold method.

  The problem of thresholding-base methods is the value of threshold. To overcome it, these method were combined with another methods. The work in [9] and [10] used Otsu's method and morphology to segment the nuclei cells. In this case, it based on the shape of the nuclei to remove the other components. Another way, [11] presented a thresholding method by using Atanassov's intuitionistic fuzzy and interval type II fuzzy set theory. This method had the good result when compared to other threshold methods.

  The thresholding-based methods was easy affected by the noise of other component, the lack of sensitivity and the shape because of the optimal value of threshold. However, this method easy to implement and efficient with some simple cases. It did not need prioi information or calculation. This methods were often used to segment the nuclei of leukemia cells.

- **Region-based**

  In [12], Byoung et al. proposed two different schemes for segmenting the nuclei and cytoplasm of white blood cells (WBC). They used stepwise merging rules for nuclei segmentation and boundary removal rules and gradient vector flow snake for cytoplasm segmentation. Some results were slightly over-segmented because of a

weak difference between the cytoplasm and the background or connected red blood cells (RBC).

These methods only detected the nuclei of cell after that using some region growing method to detect the cytoplasm. However, they could not detect the complete cytoplasm. All of above method are used for segmentation WBC cells and ALL cells, which belong to Acute Leukemia type.

- **Clustering-based**

  A group of clustering methods were also applied to solve the problem. It attempts to achieve partition such that objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible.

  In [14], Zhang et al. presented the K-means clustering method based on the result of color correction methods. This method depended on the number of the initial clusters that were chosen.

  Sarrafzadeh et al. [15] used K-means clustering method by selecting A* & b* components as features. They proposed a method that applied region growing with a stop condition to separate the connected cells. This combination segmented the nuclei and the cytoplasm of white blood cells.

  Varghese et al. [16] also used the K-means method for clustering, but they combined the clustering method and layer subtraction method with RGB images.

  Instead of using a single clustering method, Jabar et al. [17] used a hybrid clustering techniques and mean shift for automated detection acute leukemia cells. They analyzed three hybrid clustering method, which were FCM, EKM and CKM. Then based on the results, they chose the EKM methods.

- **Edge-based**

  In [19], Mahmood et al. was presented the method that combining the Canny edge detection and the Hough transform algorithm.

  Again et al. [20] used edge enhancement and Candy's method for detecting the nuclei of AML cells. The algorithms were applied on sub-images and the results shown that these still manage to present higher accuracy than other systems employing similar techniques.

- **Other-based**

  In [21], Ruberto et al. used mean-shift procedure in order to search the clustering. Then the samples of ALL cells were used for training with support vector machine (SVM) model. They used SVM model to correctly classify all the pixels of a given test images. The results of this method depended on the choice of training set.

  For the AML cells, follow by W. Ismail [22], the combination of Cellular Automata and Heuristic Search was applied. Although there are a few methods which were proposed, however, there are also some problems that need to solve such as the overlapping of blood cells, the noise of images, etc.

Based on the results of the detection process, the classification process identify the subtypes of AML cells. Subtypes of AML can be classified M0 to M7 according to French-AmericanBritish (FAB) classification. The blast cell of these subtypes are different in terms of size, shape, an amount of cytoplasm, shape and number of nucleus and the constituent in the cytoplasm. Especially, it is necessary to identify a subtype M3 (acute promyelocytic leukemia) with others because the treatment is different between subtypes.

In this process, the features which extracted from leukemia cells have a significant effect on the accuracy of classification.

D. Goutam, et al. [23] performed feature extraction based on mean, median, variance, high and low for RGB that consists of 15 attributes. Then, the neural network was used to learn and classify.

The work in [24] extracted the gray level co-occurrence matrices texture features and shape features from the nuclei of cells. These features are energy, homogeneity, entropy and correlation.

1.3 Motivation and Goal

It is important to accurately determine the subtype of leukaemia, since treatment may differ. In case of wrong diagnosis, patients face complications and may die.

The early recognition of blast cells in the bone marrow of patients suffering from AML, during the developmental stage of the illness, is extremely important for appropriate treatment [25] Clinicians have to identify these abnormal cells under a microscope, in order to decide if a patient with suspected leukaemia would need a bone marrow transplant. In addition, they have to calculate the number of blast cell to confirm the diagnosis. Before classification of AML, it is necessary to recognise the types of blast cells which can be observed, and how they may differ from promyelocytes (M3 subtypes) [4].

Currently, the whole process was manual in nature and thus was time consuming and exhausting. Nowadays, there are several research groups focusing on the development of image processing application for medical images that collaborates with the clinicians. With a blood cell image of leukemia patient, the process of detection and classification depends on human look and it takes up to a few days. Haematologists are very difficult to identify the correct subtypes, given the morphological similarities they share.

The haematologists examine and count blast cells from blood smear under a microscope and attempt to identify the subtypes of blast cells. They need to diagnose AML based on the cytogenetic testing which takes from three to five days. The process is extremely tiring and time-consuming process. The work in this thesis aims to assist haematologists in detecting and classifying leukaemia cells. A quick diagnosis allows patients to receive treatment properly.

The goals of this study is to investigate the development of an automate method for the

detection and classification leukemia cells from the blood cells images that are captured as microscope images. The result of this study assist the heamatologist's diagnosis of AML cases to provide the correct form fo treatment. Besides, it also helps the diagnostic process to reduce in terms of its time span from a few days to a matter of a few hours and the cost of all processes. Moreover, this method limited the impact of human beings during the diagnostic process.

To achieve the goal, this study focus on the targets as following:

- Detect and locate the nuclei and the cytoplasm of AML cells from the input images.
- Extract the features from the AML cells.
- Classify the AML cells into 4 subtypes as M1, M2, M3, M5. The most important issue is the classification of an image into M3 or the group of the remaining subtypes, as AML M3 requires different treatment.

There are two parts of the AML cell, called nuclei and cytoplasm. The nucleis can be easily segmented because the contrast between the nuclei and other components of the images are high. However, the difference of color between background stain and cytoplasm is minimal, thus it is difficult to extract cytoplasm from the background. Therefore, in this study, the author will analysis the feature of AML cells and propose a method to solve this problem.

The big challenge of classification stage is the similarity of leukemia cells, so a suitable extracted features and machine learning model are chosen to increase the accuracy of the classification process.

1.4 Structure of Thesis

Structure of this paper is as follows:

- **Chapter 1: Introduction**

  In this chapter, the background of Acute Myeloid Leukemia are described. Besides, the advantages and drawbacks of related works are discussed. After that, the motivation and objectives of this study are described.

- **Chapter 2: Proposed method**

  The process consists of two stages: detection and classification. First, the location of AML cells was detected by using the color conversion and the gradient magnitude in detection stage. Then, the features of AML cells were extracted and applied for training and testing in classification stage. Finally, this stage classifies the AML cells into 4 subtypes. The details of proposed method, which were used, are explains in chapter 2.

- **Chapter 3: Experiment results**

  In this chapter, the author evaluate the proposed method with real dataset. The information of the dataset are described. The result of detection method is compared with a existing method. The results of detection stage and classification stage are also reported and analyzed in this chapter.

- **Chapter 4: Conclusion**

  Finally, a summary of presented methods will be shown in Chapter 5. Besides, the author discuss about the improvement ability of the proposed method and the future work of this study.

Chapter 2

Proposed method

2.1 Overal system

To achieve the goal of this study, a system was presented. There are two main stages in this system. Figure 2.1 shows the diagram of the system.

The first stage is detection. In this stage, it located all the AML cells that appearing in the input images. Different to the existing works, they only focus on detecting the nuclei of AML cells, this study detect and locate both the nuclei and the cytoplasm of AML cells. This work is very difficult because the characteristics between the cytoplasm and the background are similar, so the proposed method will solve this problem. The detail of proposed method was explained in the next chapter.

The second step is classification. This stage classifies the AML cells into four subtypes. After the detection stage, the AML cells were segmented. The author extract the features from the nuclei and the cytoplasm of AML cells and distribute them into a feature vector. This vector used to train for the learning model and test the accuracy of the classification.

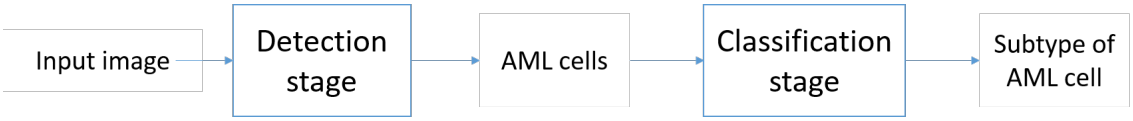


Figure 2.1: The diagram of system

Following this system, from an input image, the AML cells which appearing in the image were located and were classified into 4 subtypes.

2.2 Detection stage

Other methods aim first to detect the nuclei of AML cells, which are more important than other components. The author proposed a method for both the nuclei and cytoplasm detection based on the color conversion and the improvement of threshold technique with the gradient magnitude of images. Figure 2.2 shows the flow diagram of the proposed method, which consists of 6 small steps.

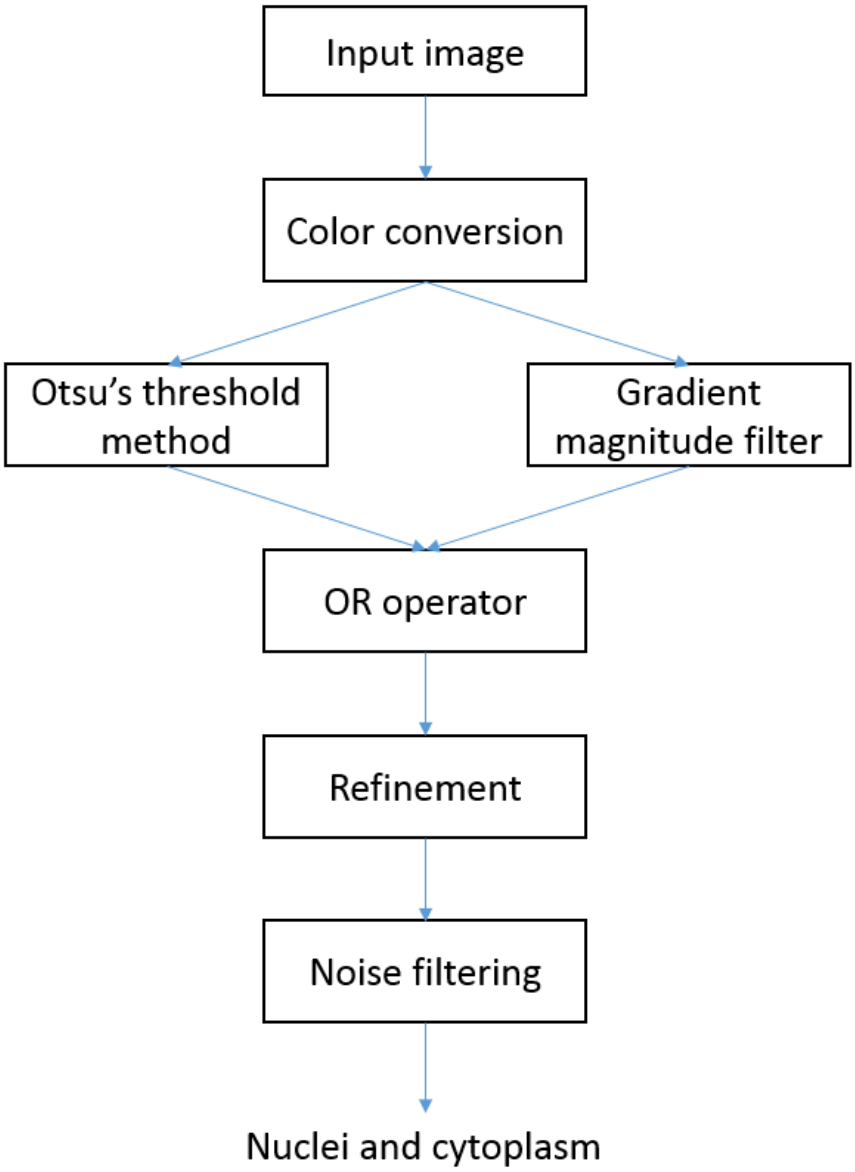


Figure 2.2: Flow diagram of proposed method in detection stage

Firstly, the color conversion step was the step that preprocessing the input image. The purpose of this step was that increasing the contrast between the AML cells and another component such as background, white blood cell.

The AML images, which were captured by microscopes, were usually in RGB color space. In this step, the RGB color space was converted to CMYK space. In fact, AML cells are more contrasted in the Y component of CMYK color model because the yellow color was presented in all elements of the image, except AML cells. Redistribution of image grey levels is necessary to make subsequent segmentation process easier.

Based on the value of the red, green and blue components, the conversion calculated the value of each component in CMYK model following the equations as below.

The red (R), green (G), blue (B) values are divided by 255 to change the range from 0..255 to 0..1.

$$R' = R/255 \tag{2.1}$$

$$G' = G/255 \tag{2.2}$$

$$B' = B/255 \tag{2.3}$$

The black (K) color is calculated from the red (R'), green (G') and blue (B') colors:

$$K = 1 - max(R', G', B') \tag{2.4}$$

The cyan color (C) is calculated from the red (R') and black (K) colors:

$$C = \frac{1 - R' - K}{1 - K} \tag{2.5}$$

The magenta color (M) is calculated from the green (G') and black (K) colors:

$$M = \frac{1 - G' - K}{1 - K} \tag{2.6}$$

The yellow color (Y) is calculated from the blue (B') and black (K) colors:

$$Y = \frac{1 - B' - K}{1 - K} \tag{2.7}$$

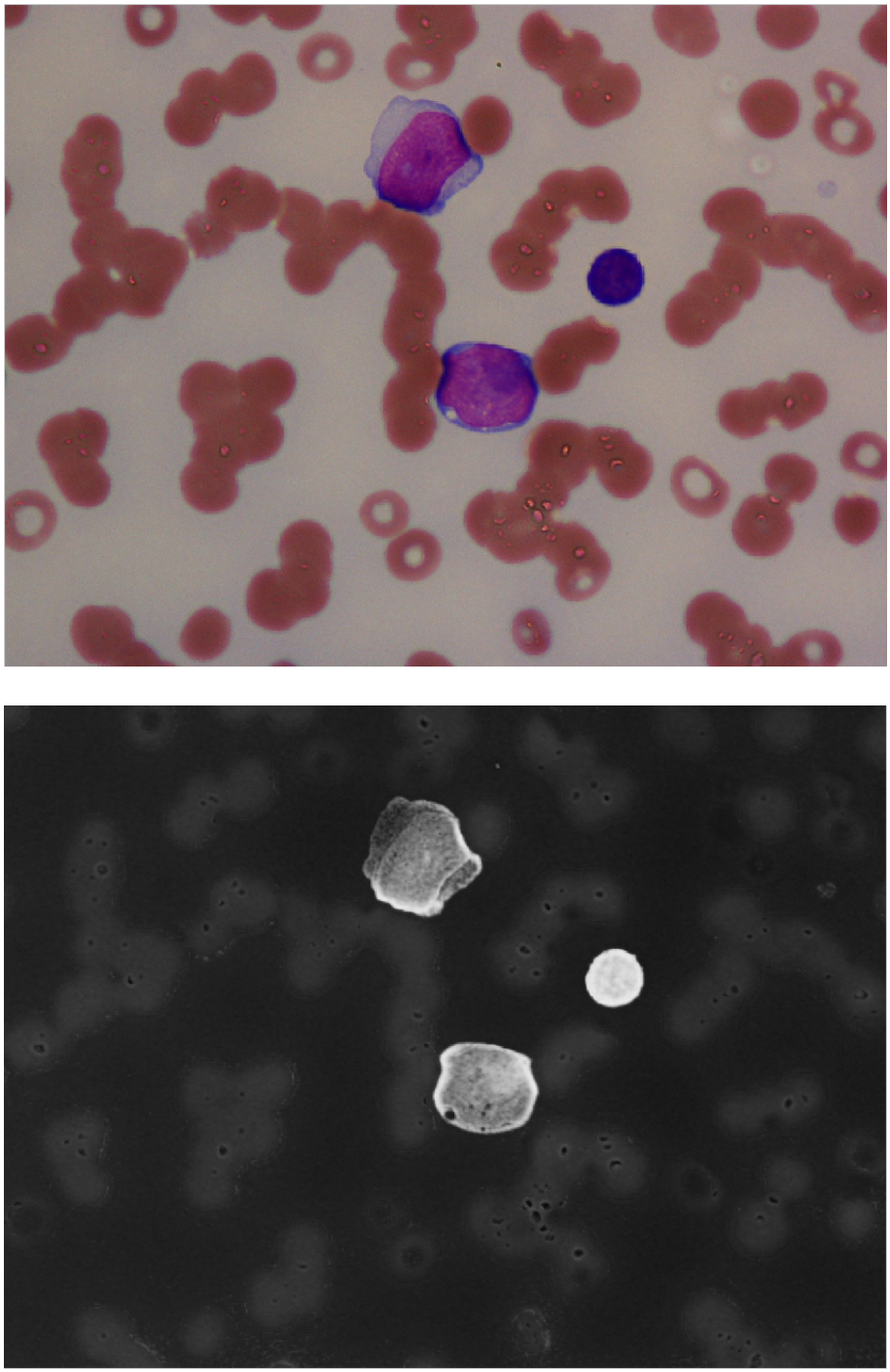


Figure 2.3: Top to bottom: original sample image, Y component image

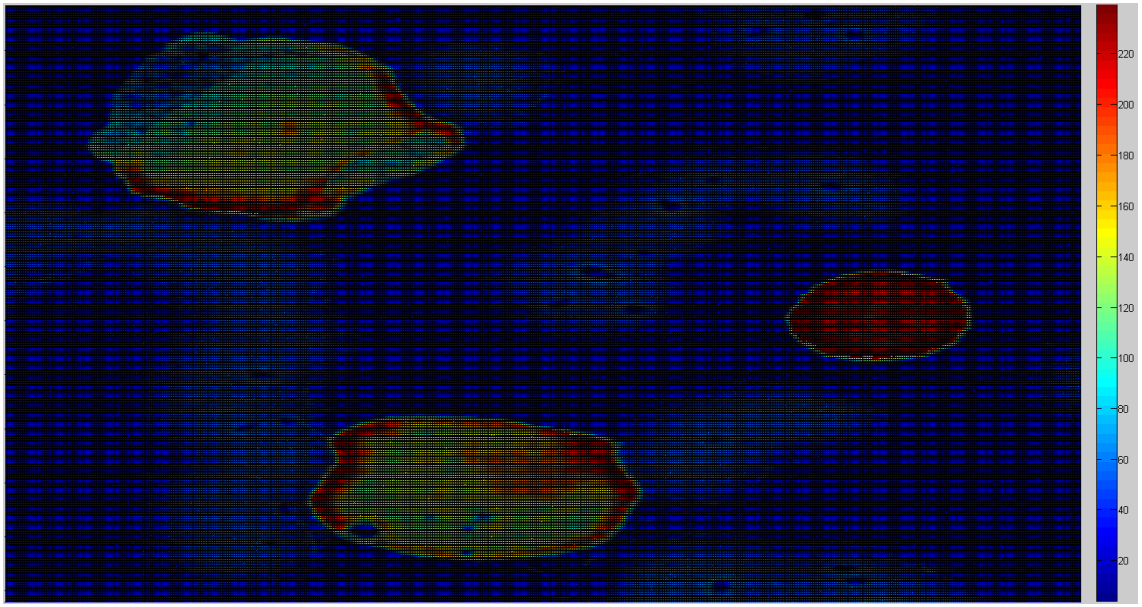


Figure 2.4: The redistribution of grey level image

Segmentation is achieved by using an automatically calculated threshold. Here, the threshold valued was used based on the Otsu's method [27]. This is the basic threshold method based on a simple idea: find the threshold that mininizes the weighted within-class variance. The value of threshold depend on each image. Therefore, this method adapt with the change of the subtype of AML cells.

Assume that I is the input image of this step. Let $I(x, y)$ be the coordinates of each pixel in an image. After the value of Y component was calculated, the author normalize it into the range [0..255] as a grey level image.

Otsu's method assumes two classes, C_0 and C_1 (background and objects or foreground respectively). The threshold T, allowing the best separation of classes in grey levels, would be the best threshold.

$$I(x, y) = \begin{cases} 1, \text{ if } I(x, y) \geq T \\ 0, \text{ otherwise} \end{cases} \tag{2.8}$$

Let the pixels of a image be represented in L grey levels [1, 2, ..., L]. The number of pixels at level α is denoted by n_i and the total number of pixels by:

$$N = \sum_{i=1}^{L} n_i \tag{2.9}$$

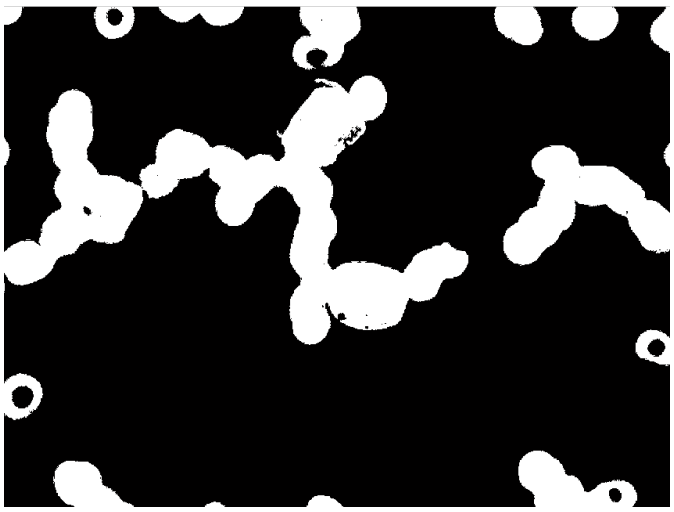
Pixels belong to two classes C_0 and C_1 separated by a grey scale threshold T. Pixels with intensity level $[1, ..T]$ belong to class C_0, while pixels with intensity $[T + 1, ..L]$ to class C_1. Hence, pixels with intensity below the threshold T as grey scale defined as partitioning an image between "black", and "white" otherwise.

The value of threshold T is calculated based on the equation:

$$\sigma_w^2(t) = w_0(t)\sigma_0^2(t) + w_1(t)\sigma_1^2(t) \tag{2.10}$$

where: $w_{0,1}$ are the probabilities of the two classes separated by a threshold t and $\sigma_{0,1}^2$ are variances of these two classes.

In this step, the author detect the strong regions, which contain the nuclei of AML cell. When this threshold method combined with the color conversion, the red blood cells were eliminated from the image. Figure 2.5 shows the difference of Otsu's method results between the grey-scale input and the Y component input.

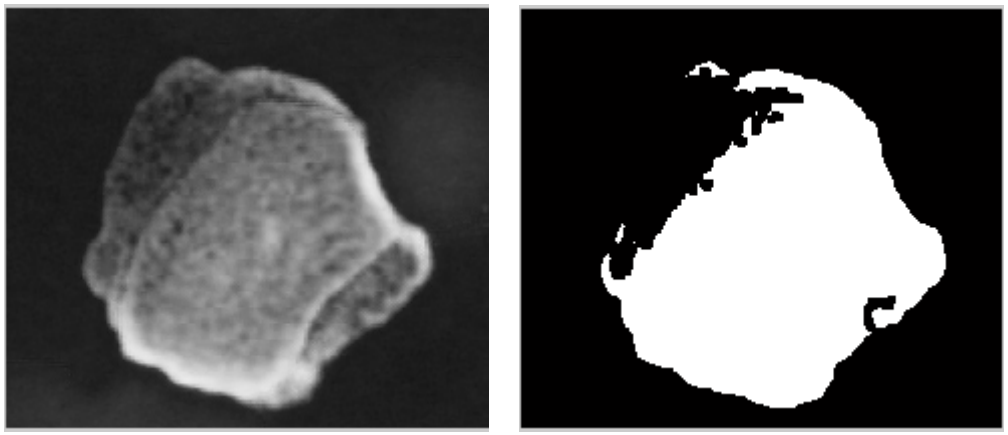


(a) From the grey-scale image

(b) From the Y component image

Figure 2.5: The results of Otsu's method with different input image

After applying threshold method, the author segmented the regions of object. In some cases, this threshold method cannot segment the cytoplasm because of the similar intensity between the cytoplasm and background. However, this step can help to detect the cytoplasm when combined with the gradient magnitude by OR operator.



(a) The Y component image (b) Result of threshold method

Figure 2.6: Demonstrate the threshold method

The problem of threshold method is the value of threshold. When the value is changed, the result will be changed too. If the value is increased, so result will not cover the cytoplasm of AML cells. Because the intensity between the cytoplasm and the background are similar, so if the value is decreased, the noise regions will be also detected. Therefore, in this study, the author proposed a method that solving this problem.

Based on the idea that the gradient magnitude between AML cells region and the background is different, we design a mask filter to calculate the gradient of image. After that, we choose a threshold and filter the region of cytoplasm. The final result after applying it to threshold method is encouraging. An algorithm was presented to combine the result of this step with the result of Otsu's method. This step segmented exactly the cytoplasm and the nuclei of AML cell.

Figure 2.7 visualizes the change of gradient magnitude. The changes between the background and the cytoplasm are shown clearly.

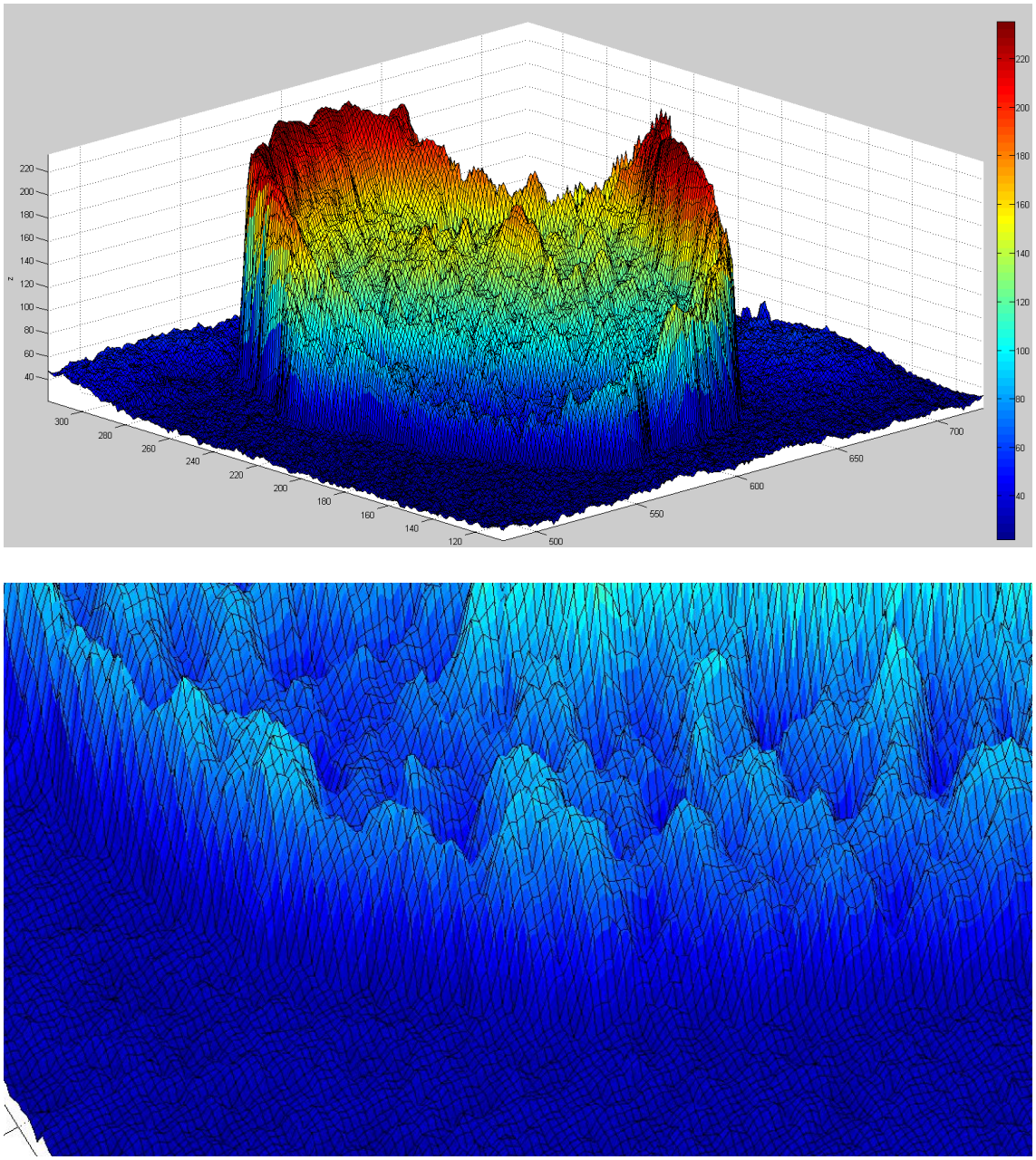


Figure 2.7: The visualization of the change of gradient magnitude

The gradient of a two-variable function at each image point is a 2D vector with the components given by the derivatives in the horizontal and vertical directions. This study uses two 3x3 kernels which are convolved with the image to calculate approximations of the derivatives - Gy for horizontal changes, and Gx for vertical. The value of two kernels are:

$$Gx = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}, Gy = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{2.11}$$

At each point in the image, the resulting gradient is combined to give the gradient magnitude, using the equation:

$$G = \sqrt{Gx^2 + Gy^2} \tag{2.12}$$

The value of threshold T is set to filter by computing the mean of all values in G. The result is shown in the left image of Fig.2.8. The value of each point in the result image Ir is calculated by:

$$\forall (x, y) \in Ir, \;\; Ir(x, y) = \begin{cases} 1 & G(x, y) \geq T \\ 0 & G(x, y) < T \end{cases} \tag{2.13}$$

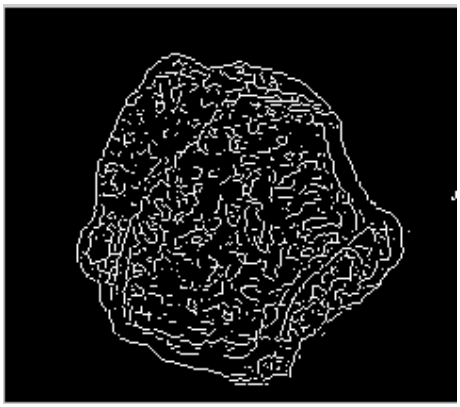


Figure 2.8: The result of gradient magnitude calculating

The definition of the points of the result are "weak point". It belongs to the cytoplasm regions. These points were disjointed so they had been connected together. All of the weak points surround the strong regions. Therefore, based on the location of strong regions, the author connect these points to the cytoplasm region in the next step.

After that, the author used the OR operator to combine the result of Otsu's method and the result of gradient magnitude. In some cases, the change of the gradient magnitude between cytoplasm and background was not clear so when we applied the threshold, the output images lost some parts of the cytoplasm.

In fact, the region of cytoplasm that could not be segmented by threshold method is covered by the result of gradient magnitude filter. In the refinement step, the author

connected all points in the result image of gradient magnitude that belong to the AML cells.

The author proposed an algorithm to achieved this objective based on the entropy of the pixels around the strong regions and the distance from these points to the nearest points in strong region of threshold method result.

The process of this algorithm:

- Use a sliding window 3x3 to scan over the area around the strong regions based on the radius of the strong area. With each position, we calculated the ratio between the number of pixel and the mask.
- Choose a threshold to filter the candidate.
- Connect all pixels inside the windows that satisfying the condition of threshold. After that, these windows are considered as a block of connected point.
- Calculate the distance from these blocks to the nearest points belong to the strong region.
- Combine all the blocks that the distance less than a half of the diameter of strong region.

After applying the algorithm, there were some holes inside the result. A technique is used to filled these holes and achieved the nuclei and cytoplasm regions.

(a) The result of OR operator (b) The result of connecting algorithm (c) The result after refinement step

Figure 2.9: The demonstration of refinement step

When the proposed method applied the gradient magnitude filter to Otsu threshold method, there were some regions in the images satisfy the condition and it became the noise of images. Therefore, we set a condition to eliminate them. In the noiseless region, the value of intensity is low and the size is small. Based on this, the constraints for eliminating a region:

$$\begin{cases} \text{area} \leq max(\text{all region})/4 \\ mean(\text{intesity of region}) < mean(\text{intensity of image}) \end{cases} \tag{2.14}$$

This condition eliminated almost the noise in the binary images and kept all AML cells. The value of parameters in the equation were chosen based on the heuristic search. After the author analyzed all the result images of refinement step, there were some AML cells, which are connected together. So the author set the value of threshold that satisfying the condition to eliminate the noise and keep the AML cells. Figure 2.10 shows an example of the image with the noise regions.

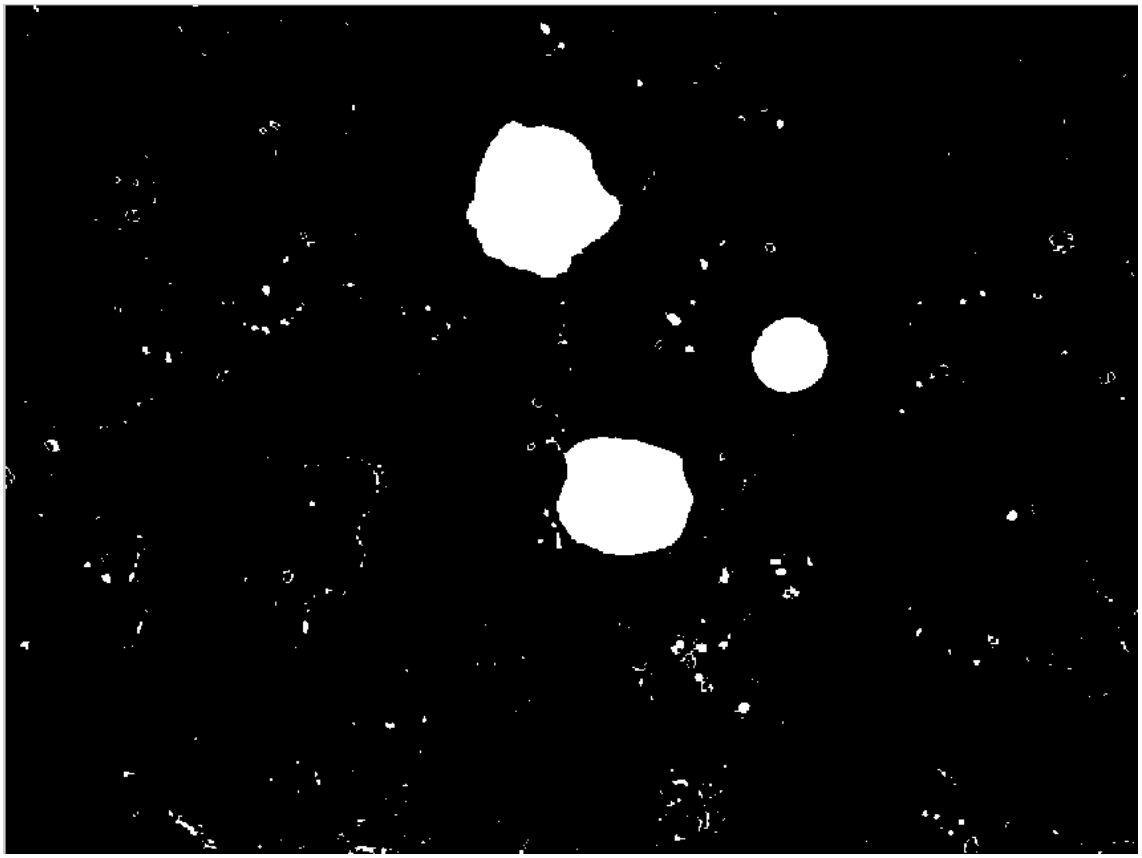


Figure 2.10: An example of the image with the noise regions

As we can see in Figure 2.11, after applying the elimination condition, all the noise regions are removed in the image. In the result binary images, the AML cells are presented by the white region.

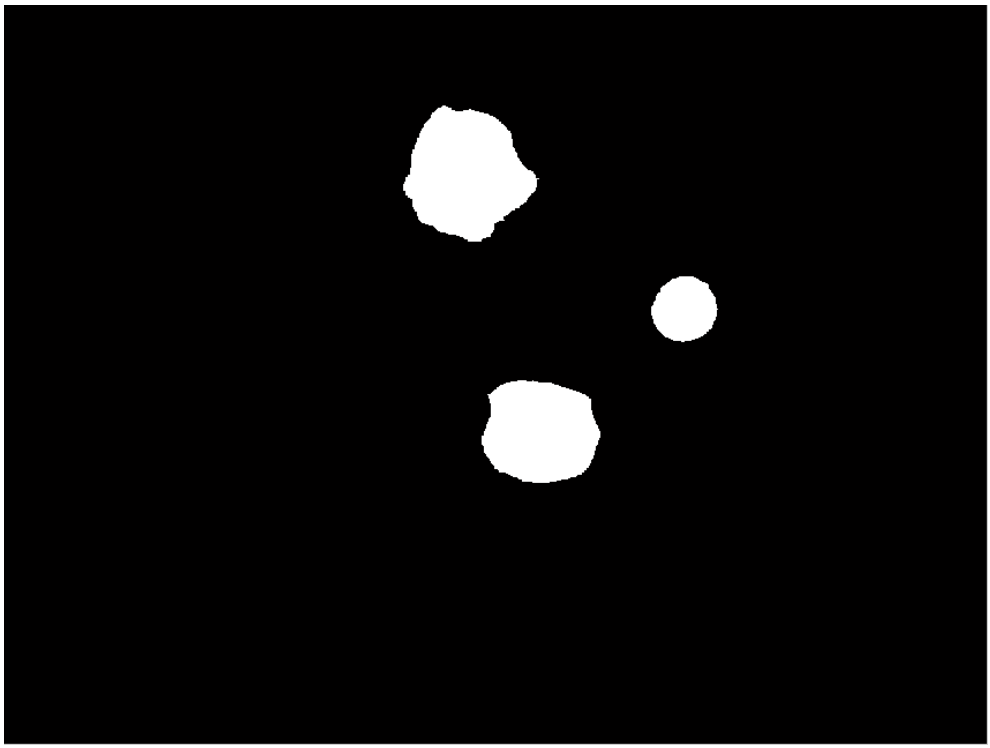


Figure 2.11: The result after eliminating the noise regions

In some cases, when the proposed method tried to connect the weak points belong to the cytoplasm with the strong region, there are some regions that belong to the others component like background, red blood cells. These regions were connected to the cytoplasm of AML cells.

So, the result of this step contained the unwanted regions, which not belong to the AML cells. This issue let the wrong result when locating the region of AML cells. Therefore, an algorithm was proposed to solve it. This algorithm removed the unwanted regions based on the distance of the point, which belong to the boundary of AML cells.

Figure 2.12 shows the connecting of the noise region with the AML cells. The green line present the result of the proposed method. The red circle borders the noise region. The reasons of this case is the connection between the red blood cells and the cytoplasm.

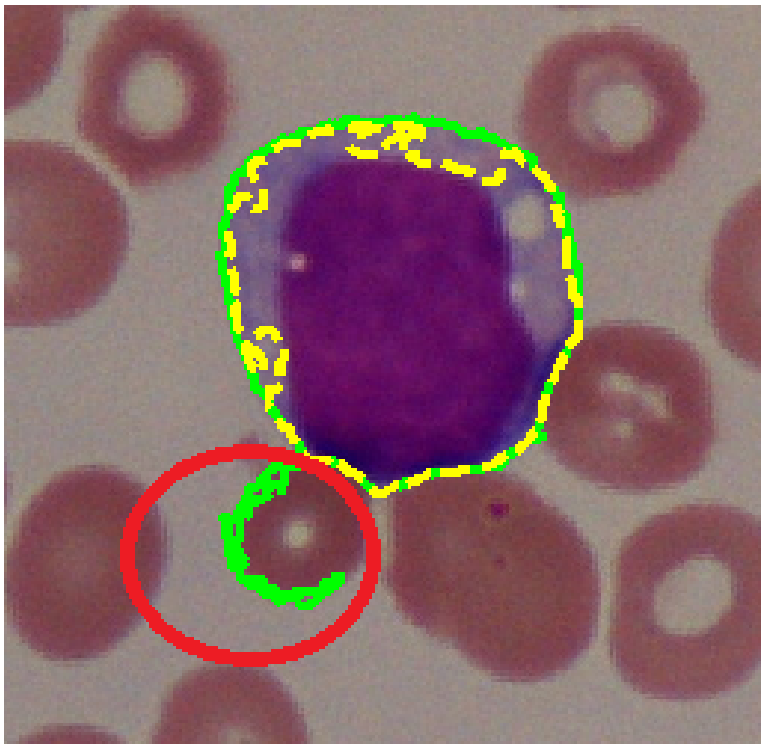


Figure 2.12: The noise region connect to the AML cell

In the noise regions, the distance between the points on which the boundary is small. The main idea of this algorithm is tried to connect the point and build a bridge which separating the noise regions and the AML cells.

So, for each point, the author open the boundary at this point following the direction inside the object. This step is repeated k times, after that, if the opened points are connected together, the author set the value of these point to 0 (background). Figure 2.13 shows an example of the algorithm.

(1)	1	1	1	(11)
(2)	1	1	1	(10)
(3)	1	1	1	(9)
(4)	(5)	(6)	(7)	(8)

K=1

(1)	*	1	*	(11)
(2)	*	1	*	(10)
(3)	*	*	*	(9)
(4)	(5)	(6)	(7)	(8)

K=2

(1)	(*1)	**	(*7)	(11)
(2)	(*2)	**	(*6)	(10)
(3)	(*3)	(*4)	(*5)	(9)
(4)	(5)	(6)	(7)	(8)

Figure 2.13: An example of noise removing algorithm

In this example, the points (**) were connected by points (*1) (*7) and (*2) (*6). Two point (*1) and (*7) is not continuous. It is same in case of (*2) and (*6). After that, we set the value to 0.

0	0	0	0	0
(2)	1	1	1	(10)
(3)	1	1	1	(9)
(4)	(5)	(6)	(7)	(8)

The bridge separated the noise regions and object region.

Figure 2.14: An example of connected bridge

The result of this algorithm is shown in Figure 2.15. The noise region was removed from the result image.

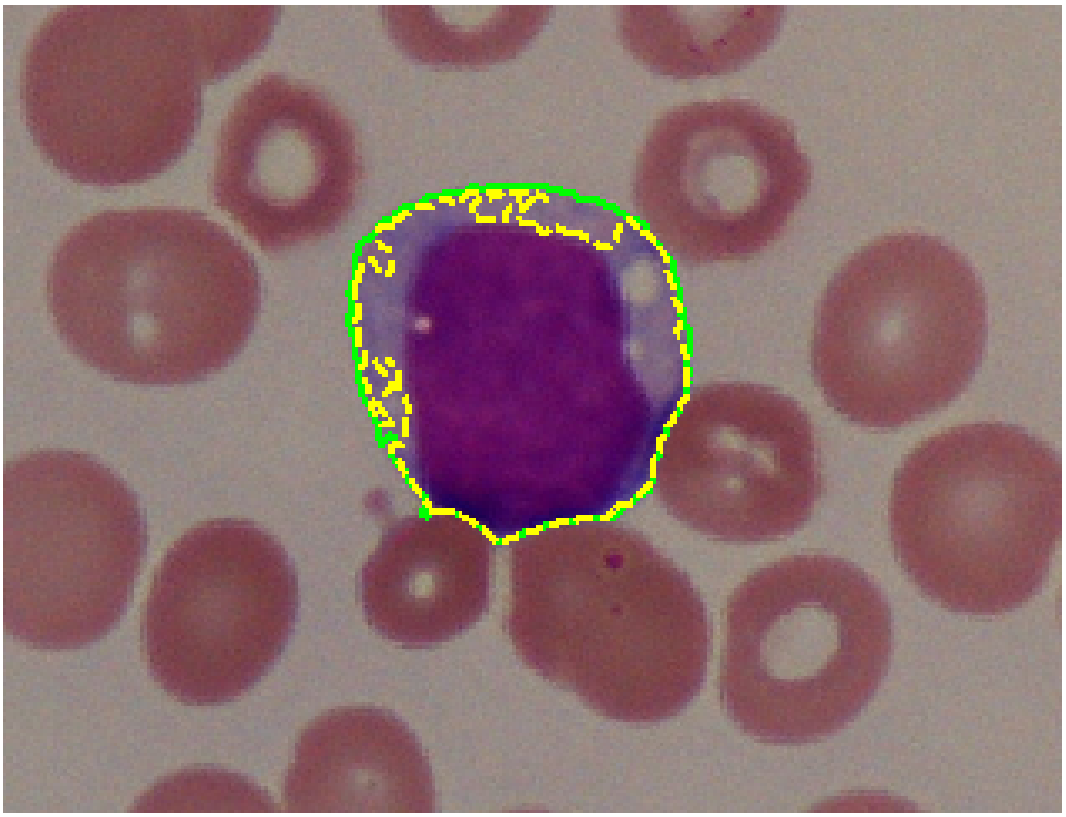


Figure 2.15: The result of connected bridge algorithm

After all above steps, the author received a binary image which contain the region of AML cells. The author located the AML cells by using the Connected Component Analysis (CCA) [26] technique in binary image. It provided the information about the position of all points inside AML cells and the boundary of AML cells. Figure 2.16 shows the result of CCA technique. In the figure, the green boundary border the AML cells.

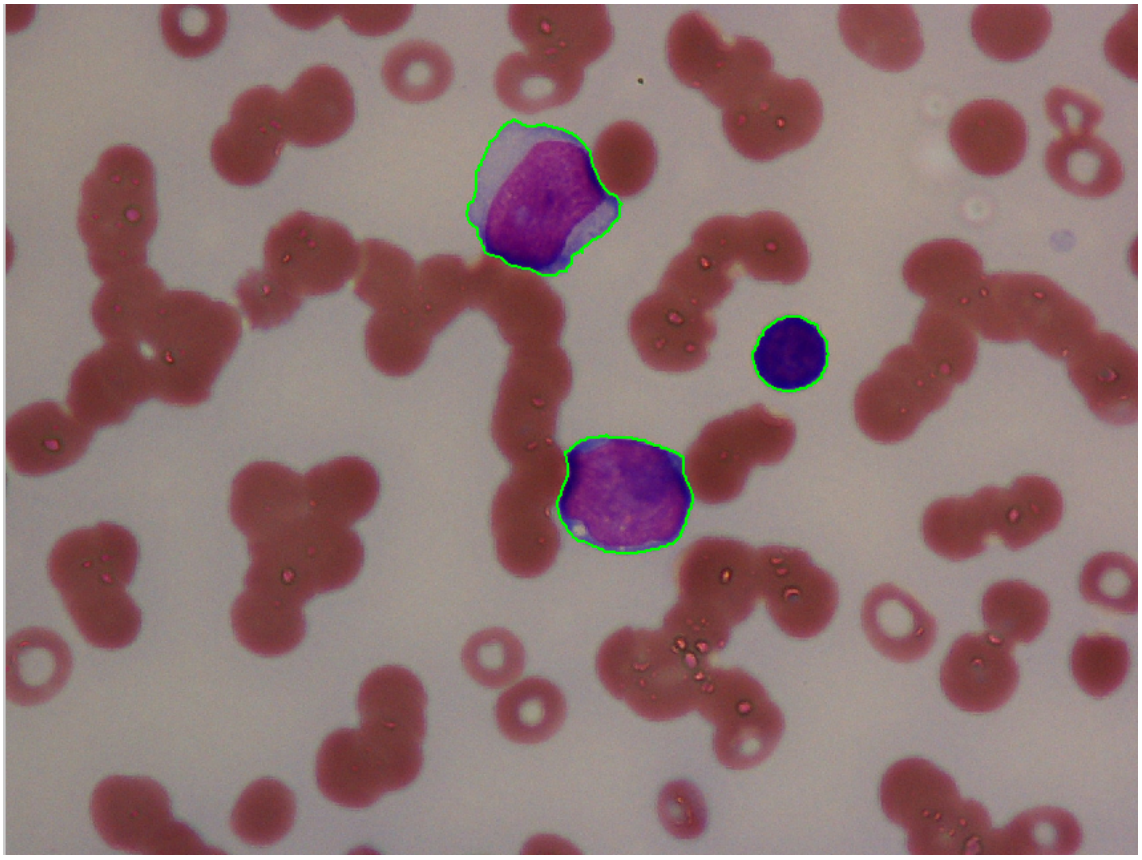


Figure 2.16: The final result of detection stage

2.3 Classification stage

Classification is the last step in the experimental methodology presented in this thesis. The objective of classification is to distinguish between M1, M2, M3 and M5 subtypes of AML cells. this chapter presented the method, including feature extraction and classification of blast cell into M1, M2, M3 and M5 subtypes.

The characteristics of four subtypes M1, M2, M3, M5 are described as below [29].

- M1: Myeloblastic without maturation

  $\geq$ 90% of myeloid cell lines are blasts.

  The blast cells show few granules but may show Auer rods.

- M2: Myeloblastic with maturation

  30-89% of myeloid cell are blasts.

  >10% are promyelocytes.

  <20% are monocytes.

  Show multiple cytoplasmic granules.

- M3: Promyelocytic

  Hypergranular promyelocytes with heavy to dust like granules.

  Frequent Auer rods, nucleus often blooded.

  Microgranular variant may occur.

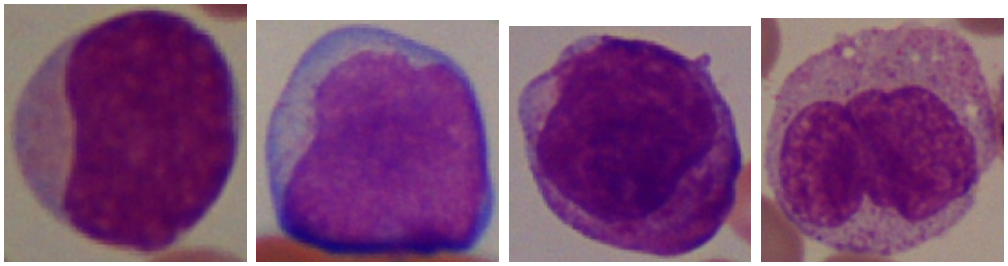
  Blast cells show multiple Auer rods.

- M5: Monoblastic monocytic

  >80% of a myeloid cell line are monoblasts, promonocytes or monocytes.

  In M5a, 80% of myeloid cell lines are monoblasts.

  In M5b, <80% were monoblast and the remainder are promonocytes or monocytes.



(a) M1 (b) M2 (c) M3 (d) M5

Figure 2.17: The examples of each AML subtype

After the detection stage, the location of the AML cells was located. Next, the author extracted features from the component of AML cells and have a feature vector. The author applied this feature vector for learning model. This learning model was used for training and testing the AML cells and classify them into 4 subtypes. The diagram of classification stage is shown in Figure 2.18.

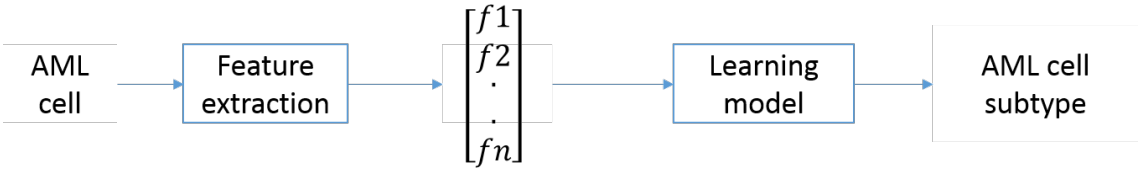


Figure 2.18: The diagram of classification stage

In extracting features step, the goal is to transform the images into data and then extract information reflecting the visual patterns that pathologists refer to, while simultaneously extracting the descriptors that are most relevant to the subsequent classification process. After the region of AML cells were located, the author extracted the features based on the color of each pixels. The author had the set of pixels:

$$S = \{p | p \in \text{AML cell region}\} \tag{2.15}$$

The author defined: p is a pixel, (x, y) is the coordinate of p, $red(p)$ is the value of red component of p, $green(p)$ is the value of green component of p, $blue(p)$ is the value of blue component of p, $ilu(p)$ is the value of illumination after converting RGB image to grey image, $f(p)$ is the value of a component which one of red, green, blue or illumination.

The author extracted the color features from the AML cells. These features show clearly the difference of AML subtypes. The descriptors which were used are: mean, median, standard deviation, low and high values. These descriptors were extracted from images in three color components (red, green, blue).

- Mean:

$$\bar{p} = \frac{\sum f(p)}{\#p} \tag{2.16}$$

- Median:

$$\underline{p} = \frac{max_{f(p)} + min_{f(p)}}{2} \tag{2.17}$$

- Standard deviation:

$$\sigma = \sqrt{\frac{1}{\#p} \sum_{i=1}^{\#p} \left(f(p_i) - \bar{p}\right)^2} \tag{2.18}$$

- Sum of high values:

$$H\left(\underline{p}\right) = \sum_{i=1}^{\underline{p}} h(p_i) \tag{2.19}$$

- Sum of low values:

$$L\left(\underline{p}\right) = \sum_{i=1}^{\underline{p}} (1 - h(p_i)) \tag{2.20}$$

where $h(p) = \begin{cases} 0, \text{ if } f(p) < 127 \\ 1, \text{ otherwise} \end{cases}$

The total number of extracted features is 15. The author put it into a vector and used it to classify. The AML cell develop to new stage in M3 subtype so the color of the AML cells will be changed. In that case, the accuracy when classifying M3 and M5 increases.

Next, the author extracted the features based on the histogram of AML cells. It shows the distribution of color component in AML cells images. This distribution is different from each AML subtypes because of the cytoplasm and nuclei. The author defined h is the histogram of the image. The author calculated in the grey images of AML cells. The descriptors which were used are: mean, variance, entropy.

- The mean:

$$\overline{p_h} = \frac{\sum_{i=0}^{255} i * h(i))}{\sum_{i=0}^{255} h(i)} \tag{2.21}$$

- The variance:

$$\text{variance} = \frac{\sqrt{\sum_{i=0}^{255} h(i) * (i - \overline{p_h})^2}}{\sum_{i=0}^{255} h(i)} \tag{2.22}$$

- The entropy:

$$\text{entropy} = \sum_{i=0}^{255} h(i) * log2(h(i)) \tag{2.23}$$

Based on the color gradient, we extracted the average descriptors in three components (red, green, blue). We defined g(p) is the magnitude of p.

- The average:

$$\text{average} = \frac{\sum g(p)}{\#p} \tag{2.24}$$

When using addition histogram, the author analyzed the detail of the cytoplasm and nuclei. The ratio between cytoplasm and nuclei is different from M5 and M1, M2. A vector which has 21 dimensions is built.

Now, the descriptors which were based on only color and histograms frequently do not provide information regarding the mutual position of the pixels. Therefore, it is necessary to consider both the intensity distribution and the position of the pixels in object. The following descriptors are evaluated: angular second moment, contrast, correlation, variance, inverse difference moment, sum variance, sum entropy, entropy, difference variance, difference entropy, information measures of correlation, maximal correlation [30].

These features were extracted to increase the accuracy of M1 and M2 classifying. The author recognized that the position of some regions in which same the intensity of each pixel is different from M1 and M2. So the author consider both the intensity distribution and the position of the pixels in the object.

Finally, the author build a vector which has 33 dimensions. This vector was applied to the learning model for training and testing.

We used the support vector machine (SVM) model for classifying because this model is particularly suitable for binary classification problems for which the separation between classes depends on a large number of variables. In this test, the linear kernel was chose to test. For the linear kernel function, the parameters were tuned using optimization techniques in order to find out the maximum accuracy value. The value of parameters c were 1e-3.

In order to classify 4 AML subtypes, the author used multi-binary classification SVM model. The process of the SVM model consists of 3 steps:

- Building 4 different binary classifiers, each trained to separate one class from the rest. For a classifier, the objects, which belong to class, are labeled positive value and the others are negative.
- Calculating the confidence value of an object in tested dataset with each classifier. The confidence value can be interpreted as the distance from the separation plane to the object.
- Assigning the object to the class whose confidence value is the largest for this object.

Chapter 3

Experiment results

3.1 Dataset

The dataset consists of 301 real images, 1280 by 960 pixels in size. They all were taken from patients suffering from AML. They were provided by Department of Haematology in the Universiti Sains Malaysia (USM) in Kota Bahru, Kelantan, Malaysia.

There are 4 subtypes of AML cells in this dataset, which are M1, M2, M3, M5. There are totally 643 AML cells inside all images. Table 3.1 shows the number of available images for each individual subtype.

The images were taken with the same lighting and contrast but different microscope zoom magnifications of 40x and 60x. The examples of images from the dataset are shown in the Fig. 3.2.

Each image has three main components: red blood cells, AML cells and the background. The input images have at least one AML cells. All AML cells belong to one of four subtypes: M1, M2, M3, M5.

Table 3.1: Dataset sample

AML subtype	Images	Cells
M1	44	82
M2	125	249
M3	90	154
M5	42	158
Total	**301**	**643**

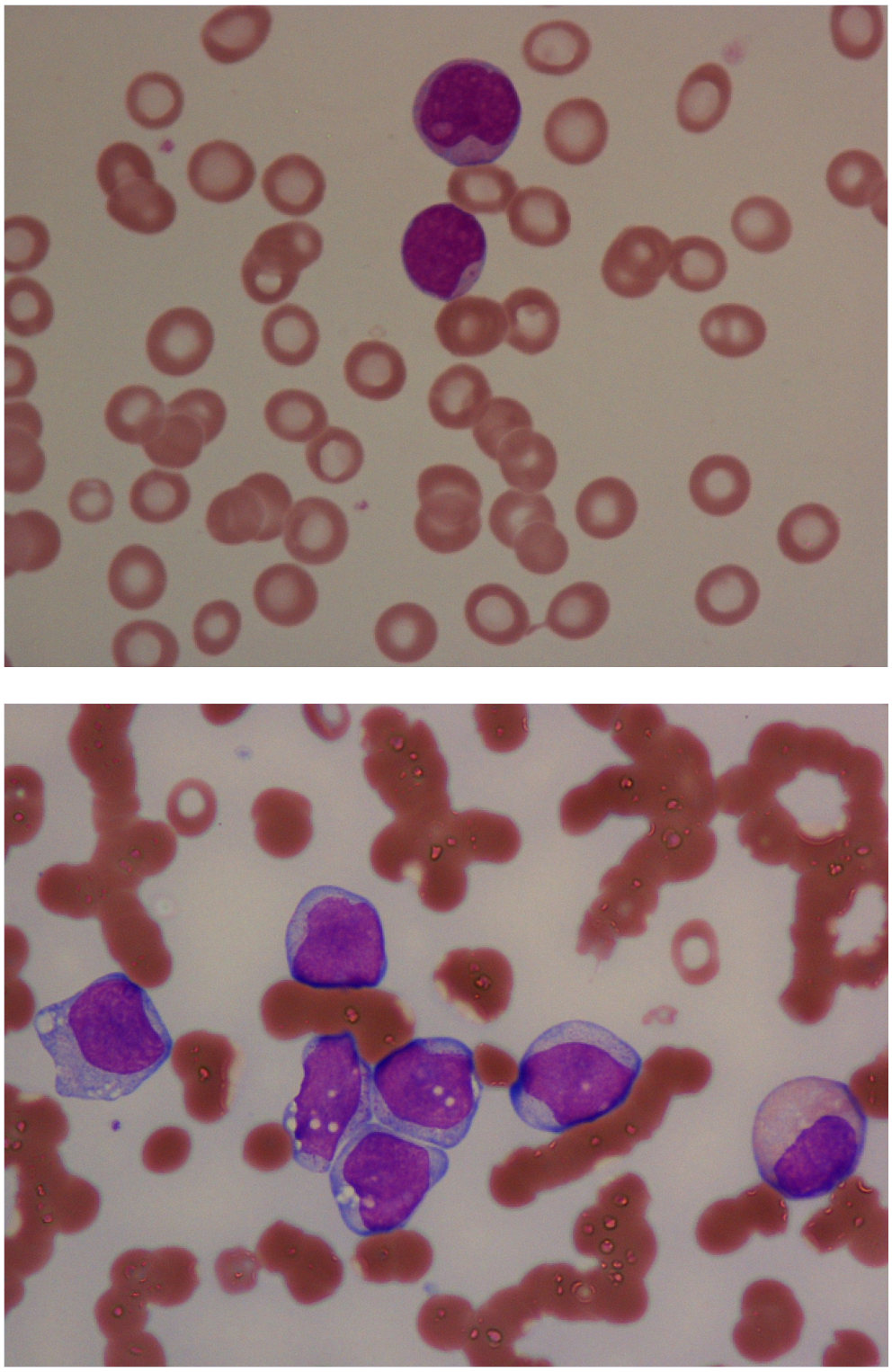


Figure 3.1: Example of a bone marrow image of M1, M2 subtype

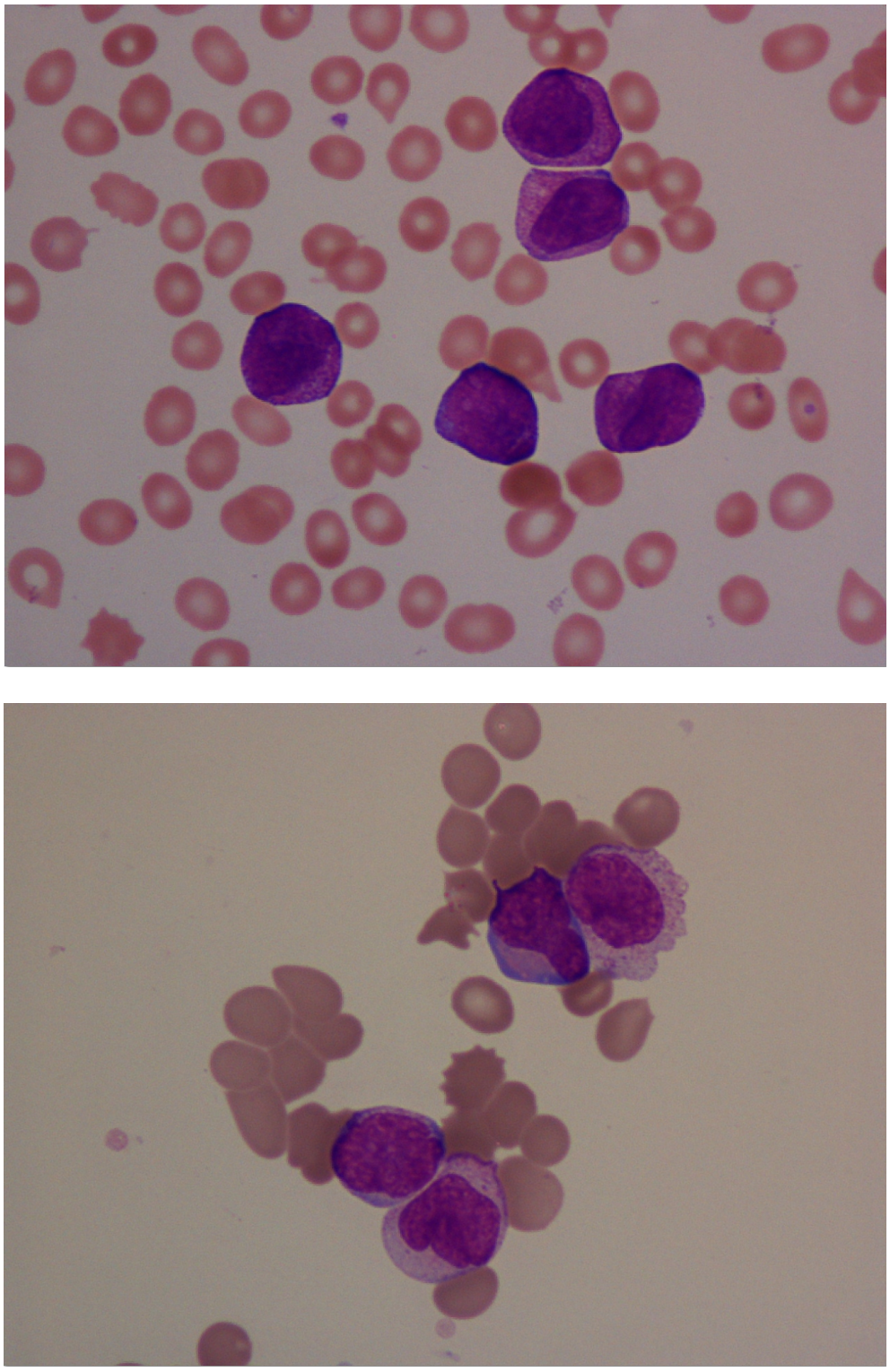


Figure 3.2: Example of a bone marrow image of M3, M5 subtype

3.2 Experiment result

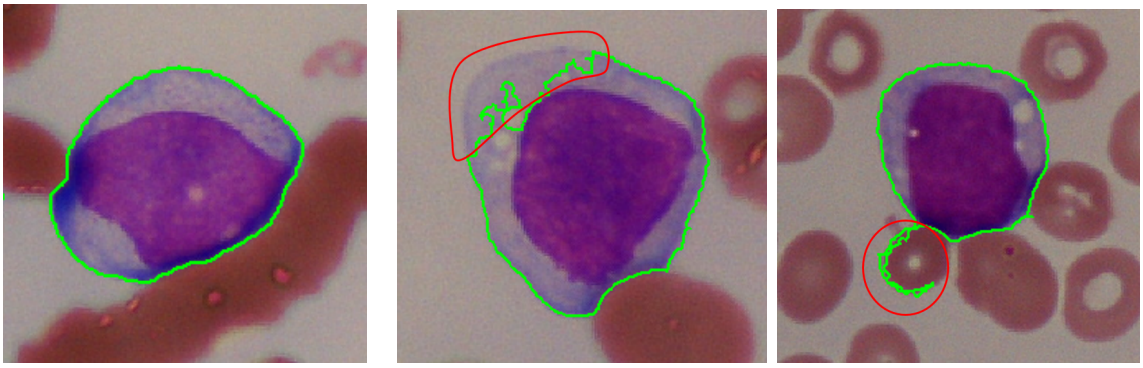
This study evaluated the proposed method in each stage. In detection stage, the author compared the position of the results with the ground-truth. The ratio between the successful case and all the cells in dataset was used to evaluate. In classification stage, the author based on the number of cells that were exactly classified. The accuracy of each subtype was calculated.

3.2.1 Detection stage

The author evaluated the performance of the proposed method by the state of AML cells which were segmented. Firstly, from the input images, the author built the ground-truth for every AML cells. After that, the author compared the overlap ratio between the results and the ground-truth. The overlap ratio was calculated by the equation:

$$\text{Overlap} = \frac{\#\text{ pixels in result}}{\#\text{ pixels in ground-truth}} \tag{3.1}$$

According to the measurement, the author divided the results of detection stage into two cases: successful case and failed case. Only detected regions with Overlap ≥ 0.9 are accepted as successful cases, the rest will be failed cases. In the successful cases, the complete nuclei and cytoplasm were detected. In the failed cases, the author divided them again into two small cases: partial case and noise case. In the noise cases, the author detected completely the nuclei and the cytoplasm but the author also located some regions which did not belong to the AML cells. In the partial cases, the author do not detect the complete cytoplasm. Some parts of cytoplasm are recognized as the other components.



(a) The successful case (b) The partial case (c) The noise case

Figure 3.3: The demonstrates of the result of detection stage

The performance of proposed method is shown in table II. We properly individuated 533 of 643 AML cells, for an average accuracy of 92.4%.

Table 3.2: The performance of detection stage

Dataset	Cells	Successful case	Failed case	
			Partial case	Noise case
M1	82	75 (91.5%)	1 (1.2%)	6 (7.3%)
M2	249	228 (91.6%)	11 (4.4%)	10 (4.0%)
M3	154	152 (98.7%)	0 (0.0%)	2 (1.3%)
M5	158	139 (88.0%)	9 (5.7%)	10 (6.3%)
Total	**643**	**594 (92.4%)**	**21 (3.3%)**	**28 (4.3%)**

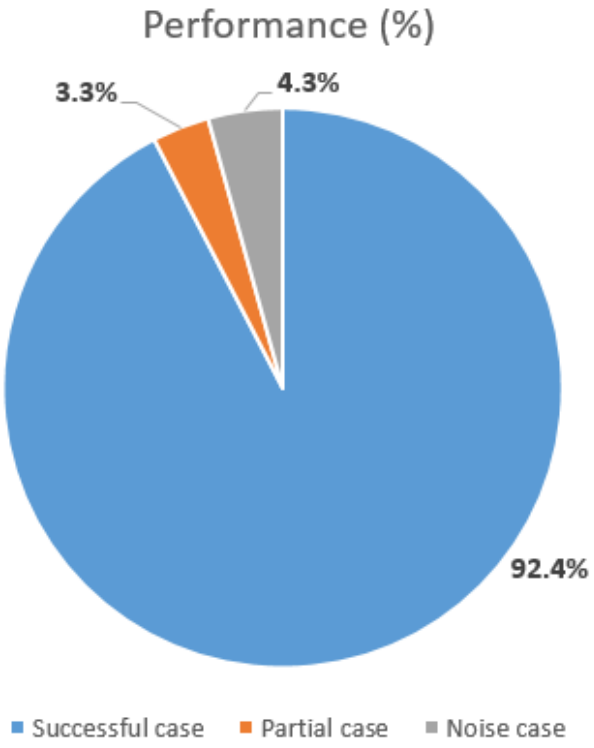


Figure 3.4: Performance for all dataset

Although the proposed method can solve successfully almost the case of problem, but there were also some cases that could not solve by the algorithm in proposed method.

In the noise elimination, there is an existent problem. If the distance of the bridge is long, the algorithm could not get the bridge and separate the noise region. In this case, it is very difficult to remove the noise. The reason are the point in the red blood cells. Figure 3.5 represents the example of this case.

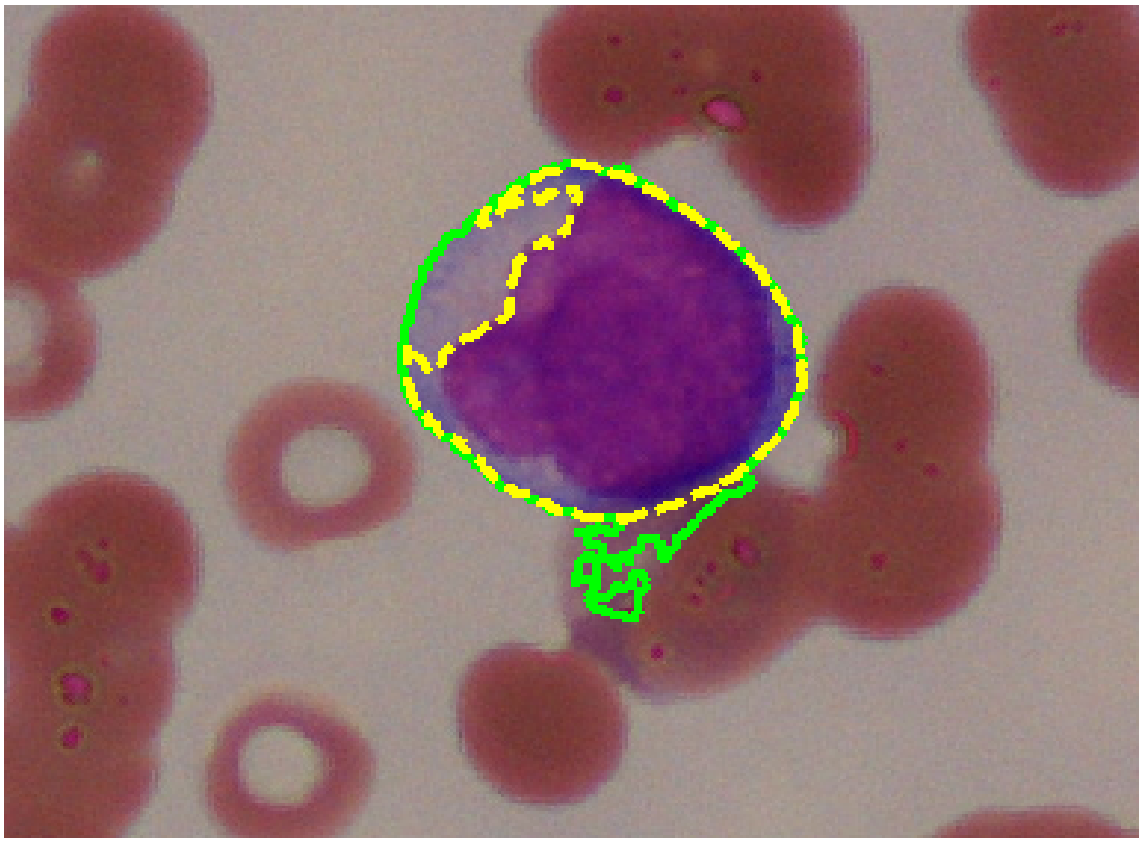


Figure 3.5: The failed case of the bridge algorithm

In the refinement step, when the connecting algorithm was applied, there were some cases which it could not solve completely. The main reason are the position between the weak points. When the distance was too far, the sliding window could not connect them.

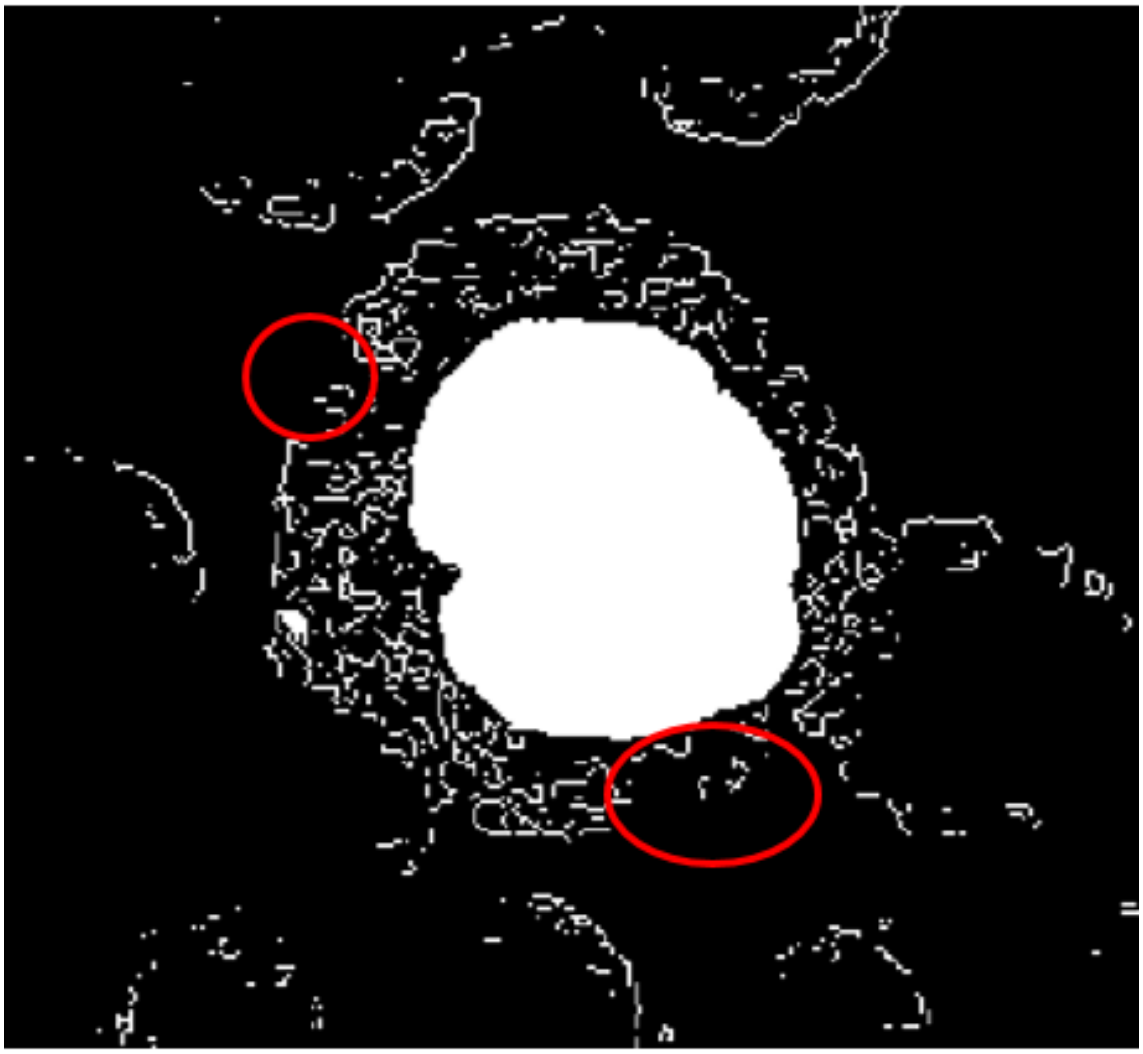


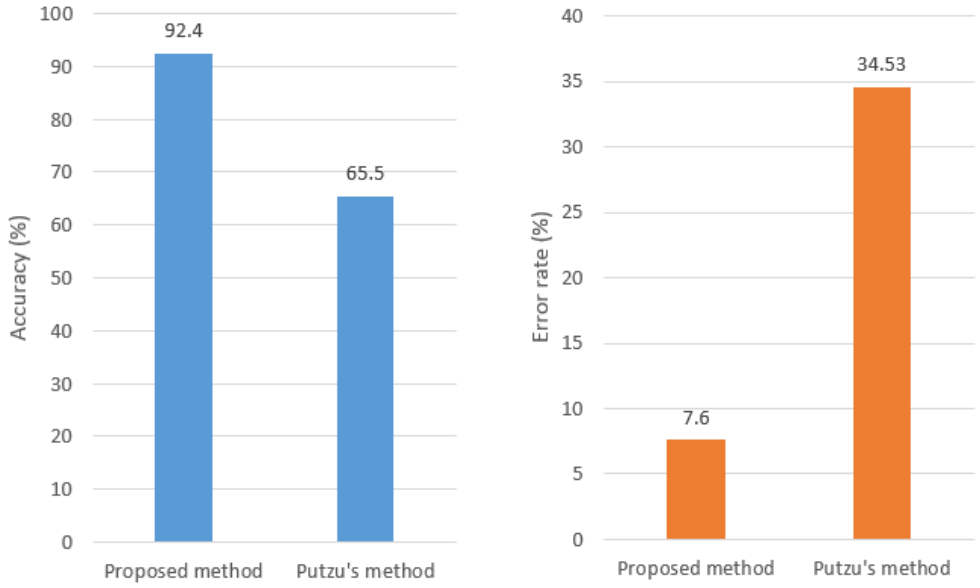
Figure 3.6: The failed case of the refinement step

Currently, there have never been any research that focus on segmenting both the nuclei and cytoplasm of AML cells. Therefore, the author compared our method with one in [8]. This method also used the color conversion and the threshold value based on the triangle method [28]. Figure 3.7 shows the comparison of two methods. The accuracy is calculated by the equation:

$$\text{Accuracy} = \frac{\text{Number of cell of the successful case}}{\text{total cell in dataset}} \tag{3.2}$$

The error rate is calculated by the equation:

$$\text{Error rate} = \frac{\text{Number of cell of the failed case}}{\text{total cell in dataset}} \tag{3.3}$$



(a) The accuracy (b) The error rate

Figure 3.7: The comparision of two methods

the author used two parameters above for comparing the proposed method with existing method. The proposed method increased the accuracy from 65.5% to 92.4% and decrease the error rate from 34.53% to 7.6%. These are encouraging experiment results.

Figure 3.8 shows the final results of two methods. The green lines border the AML cells which are the result of the proposed method. The yellow lines border the AML cells which are the result of the Putzus results.

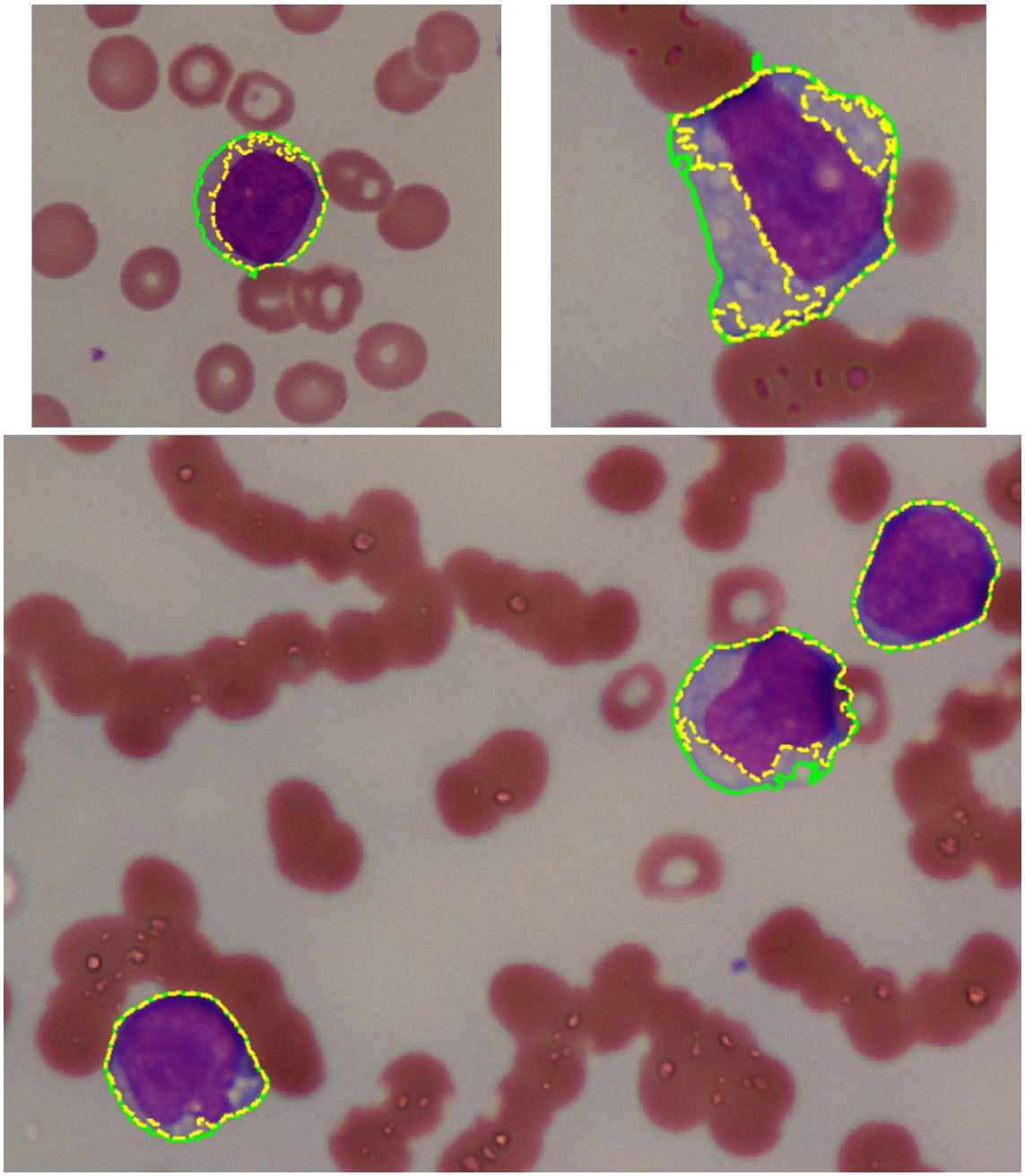


Figure 3.8: The final results of two methods. The green lines are the proposed method results and the yellow lines are the Putzus results

3.2.2 Classification stage

In addition to the type of algorithm used to induce the learning model, the performance of a model also depends on the size of the training and the test sets decrease. It depends on their specific composition, resulting in higher variance. Therefore, the author used the cross-validation technique for evaluating the experiment result of classification stage. Cross-validation is a technique partitioning the original sample into a training set to train the model, and a test set to evaluate it.

The original dataset is randomly divided into 10 equal size folds. The cross-validation process is repeated 10 times, using one different sub-sample as the validation data for testing the model and the remaining 9 sub-samples as the training data each time. Finally, the 10 performance from the folds are averaged to produce a single estimation.

Figure 3.9: 10-folds example.

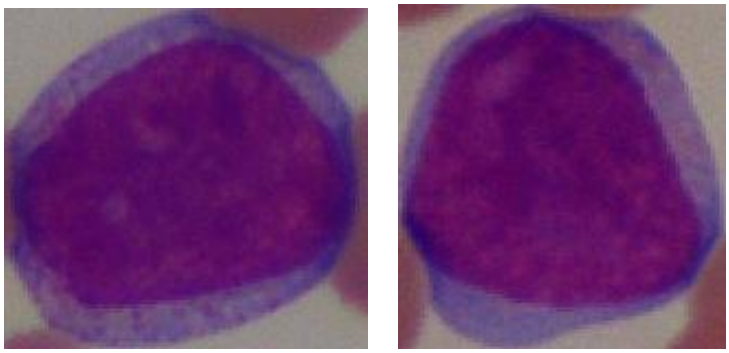
Table 3.3: The experimental result of classification

	M1	M2	M3	M5	All Dataset
Total cells	76	231	150	141	594
Cells for testing	8	23	15	14	60
Accuracy of average validation	92.3%	90.4%	97.2%	94.1%	93.5%

The accuracy of M3 classification was achieved 97.2%. This is an acceptable results for the diagnosis white blood cells. It will help the doctor to give the treatment for their patients.

In classification stage, this study focused on the accuracy of M3 classification. The classification of M3 subtypes is very important for the diagnosis. The reason for this is that identifying M3 subtype is of great importance because patients suffering from it require different treatment than patients suffering from the other subtypes. In particular, in M3 cases All-Trans-Retinoic-Acid (ATRA) is added to the initial chemotherapy, so it affected to the final results of the blood test. Show some wrong case of classification.

In some cases, the classification model could not identify exactly the subtype of AML cells. In M1 and M2 subtypes, the accuracy of classification are not good. Because it is the two first continuous stage of the AML cells life. So, the characteristics of the cytoplasm and the nuclei are not clear. This is a big challenge for extraction step.



(a) An example of M1 (b) An example of M2

Figure 3.10: Example of wrong classification of M1 and M2 subtype

Chapter 4

Conclusion

Leukemia is a cancer of white blood cells that affect the blood forming cells in the body. Acute Myeloid Leukemia (AML) is a form of leukemia and are caused by replacement of normal bone marrows with leukemic cells, which cause a drop in red blood cells, platelets, and normal white blood cells. Early classification of the subtype of AML cells is necessary for proper treatment management.

In this study, the author presented a system that detecting and classifying the AML cells. There was two stage in this system: detection and classification. In detection stage, the author proposed a method to detect the nuclei and cytoplasm of AML cells based on the change of gradient magnitude to filter the region of cytoplasm. Although it can not solve all the cases of the problems, but it worked and achieve the encouraging results with almost cases of dataset. In classification stage, the color features, histogram features and texture features were extracted from the AML cells. The SVM learning model was applied for training and validating the dataset.

We tested the system with 301 images which total 643 AML cells. The proposed method was demonstrated to improve the detection performance when compared to another method. Experimental results confirmed that the proposed method in detection stage can efficiently segment the nuclei and cytoplasm of AML cells. We can also use this proposed method for another type of cell such as white blood cell, another type of leukemia cell.

There are some AML cells that connected together, so the author did not count the number of AML cells. This study focuses on the classification of AML cells, following that, the proposed method did not separate the connected cell. However, the counting of AML cells can help to diagnose the treatment for their patients. Therefore, in the future, the author will separate the connected cell by using the watered transform method.

Bibliography

[1] A. V. Hoffbrand, J. E. Pettit, P. A. H. Moss, *Essential Haematology*, Fourth Edition, Publisher: Blackwell Science, ISBN: 0-63-205153-1, 2001.

[2] C. Haworth, A. Heppleston, M. Jones, et al., *Routine bone inarrow examination in the management of acute lymphoblastic leukeamia of childhood*, J Clin Pathol, 1981.

[3] Marketsandmarkets.com, *Acute Myeloid Leukemia Therapeutics Market in G8 Countries (2010 - 2020)*, Report Code: UC 1705, 2016.

[4] R. Hassan, *Diagnosis and outcome of patients with Acute Leukemia*, In: Haemotology department, University Sains Malaysia, Malaysia, 1996.

[5] T. G. Patil, V. B. Raskar, *Automated Leukemia Detection By using countor Signature method*, International Journal of Advance Foundation and Research in Computer, vol. 2, iss. 6, 2015.

[6] A. Khashman, E. Al-Zgoul, *Image Segmentation of Blood Cells in Leukemia Patients*, Computer engineering and applications, pp. 104-109, 2010.

[7] M. D. Joshi, A. H. Karode, S.R. Suralkar, *White Blood Cells Segmentation and Classification to Detect Acute Leukemia*, International Journal of Emerging Trends & Technology in Computer Science, vol. 2, iss. 3, pp. 147-151, June 2013.

[8] L. Putzu, G. Caocci and C. Ruberto, *Leucocyte classification for leukaemia detection using image processing techniques*, Atrtifical Intelligence in Medicine, vol. 62, pp. 179-191, 2014.

[9] S. Nazlibileka, D. Karacorb, T. Ercanc, et al.,*Automatic segmentation, counting, size determination and classification of white blood cells*, Measurement Journal, vol. 55, pp. 58-65, 2014.

[10] J. Rawat, A. Singh, H. S. Bhadauria, *An approach for leukocytes nuclei segmentation based on image fusion*, International Symposium on Signal Processing and Information Technology, pp. 456-461, 2014.

[11] T. Chaira, *Accurate segmentation of leukocyte in blood cell images using Atanassov's intuitionistic fuzzy and interval Type II fuzzy set theory*, Micron Journal, vol. 61, pp. 18, 2014.

[12] C.K. Byoung, J. Gim, J. Nam, *Automatic white blood cell segmentation using stepwise merging rules and gradient vector flow snake*, Micron Journal, vol. 42, pp. 695-705, 2011.

[13] D. Senthilbabu, S. Maheswari, *White Blood Cell Segmentation using hybrid segmentation methods*, International Journal of Engineering Research & Technology, vol. 3, iss. 2, pp. 2223-2227, 2014.

[14] C. Zhang, X. Xiao, X. Li, et al., *White Blood Cell Segmentation by Color-space-based K-means clustering*, Sensors Journal, vol. 14, pp. 1128-1147, 2014.

[15] O. Sarrafzadeh and A.M. Dehnavi, *Nuleus and cytoplasm segmentation in microscopic images using K-means clustering and region growing*, Adv Biomed Res Journal, vol. 4, pp. 174, 2015.

[16] A. K. Varghese, *Automated Screening System for Acute Myelogenous Leukemia Detection using Layer Subtraction*, Article published in International Journal of Current Engineering and Technology, Vol.5, No.5, pp. 3285-3289, 2015.

[17] F. Jabar, W. Ismail, R.A. Salam, R. Hassan, *Image segmentation using a hybrid clustering technique and mean shift for automated detection acute leukemia blood cells images*, Journal of Theoretical and Applied Information Technology, vol. 76, pp. 88-96, 2015.

[18] M. M. Amin, S. Kermani, A. Talebi, et al., *Recognition of Acute Lymphoblastic Leukemia Cells in Microscopic Images Using K-Means Clustering and Support Vector Machine Classifier*, Journal of medical signals and sensors, vol. 5, pp. 49-58, 2015.

[19] N. H. Mahmood, P. C. Lim, S. M. Mazalan, et al., *Blood cells extraction using color based segmentation technique*, International Journal of Life Sciences Biotechnology and Pharma Research, vol. 2, no. 2, pp. 233-240, 2013.

[20] S. Agaian, M. Madhukar, A. T. Chronopoulos, *Automated Screening System for Acute Myelogenous Leukemia Detection in Blood Microscopic Images*, IEEE Systems Journal, vol. 8, iss. 3, pp. 995-1004, 2014.

[21] C. D. Ruberto, A. Loddo, L. Putzu, *Learning by Sampling for White Blood Cells Segmentation*, International Conference of Image Analysis and Processing, vol. 9279, pp. 557-567, 2015.

[22] W. Ismail, R. Hassan, et al., *Detecting Leukaemia (AML) Blood Cells Using Cellular Automata and Heuristic Search*, Advances in Intelligent Data Analysis IX, vol. 6065, pp. 54-66, 2010.

[23] D. Goutam, S. Sailaja, *Classification of Acute Myelogenous Leukemia in Blood Microscopic Images using Supervised Classifier*, IEEE International Conference on Engineering and Technology, doi. 978-1-4799-1854-6, March 2015.

[24] S. Mohapatra, D. Patra, S. Satpathy, *Automated leukemia detection in blood microscopic images using statistical texture analysis*, International Conference on Communication, Computing & Security, pp. 184-187, 2011.

[25] N. Theera-Umpon, P. D. Gader, *System-level training of neural networks for counting white blood cells*, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 32, iss. 1, pp. 48-53, 2002.

[26] R.C. Gonzalez, R.E. Woods, *Digital Image Processing*, second edition, ch. 2, sec. 2.5, pp. 66-69.

[27] N. Otsu, *A Threshold Selection Method from Gray-Level Histograms*, IEEE Transactions on Systems, Man, and CyberneticsIEEE Transactions on Systems, Man, and Cybernetics, vol. 9, iss. 1, pp. 62-66, 1979.

[28] G. Zack, W. Rogers, S. Latt, *Automatic measurement of sister chromatid exchange frequency*, The journal of Histochemisty and Cytochemistry, vol. 25, no. 7, pp. 741-753, 1977.

[29] A. Bell, S. Sallah, *The Morphology of Human Blood Cells*, Seventh Edition, Publisher: Abbott Laboratories, ISBN-10: 1090346018, 2005.

[30] Robert M. Haralick, K. Shanmugam, I. Dinstein, *Textural Features for Image Classification*, IEEE Trans Syst Man Cybern, vol. 3, no. 6, pp. 610-621, November 1973.

Publication

[1] Van-Nhan Tran, Waidah Ismail, Rosline Hassan and Atsuo Yoshitaka, *An automated method for the nuclei and cytoplasm of acute myeloid leukemia detection in blood smear images*, World Automation Congress, Puerto Rico, USA, August 2016.