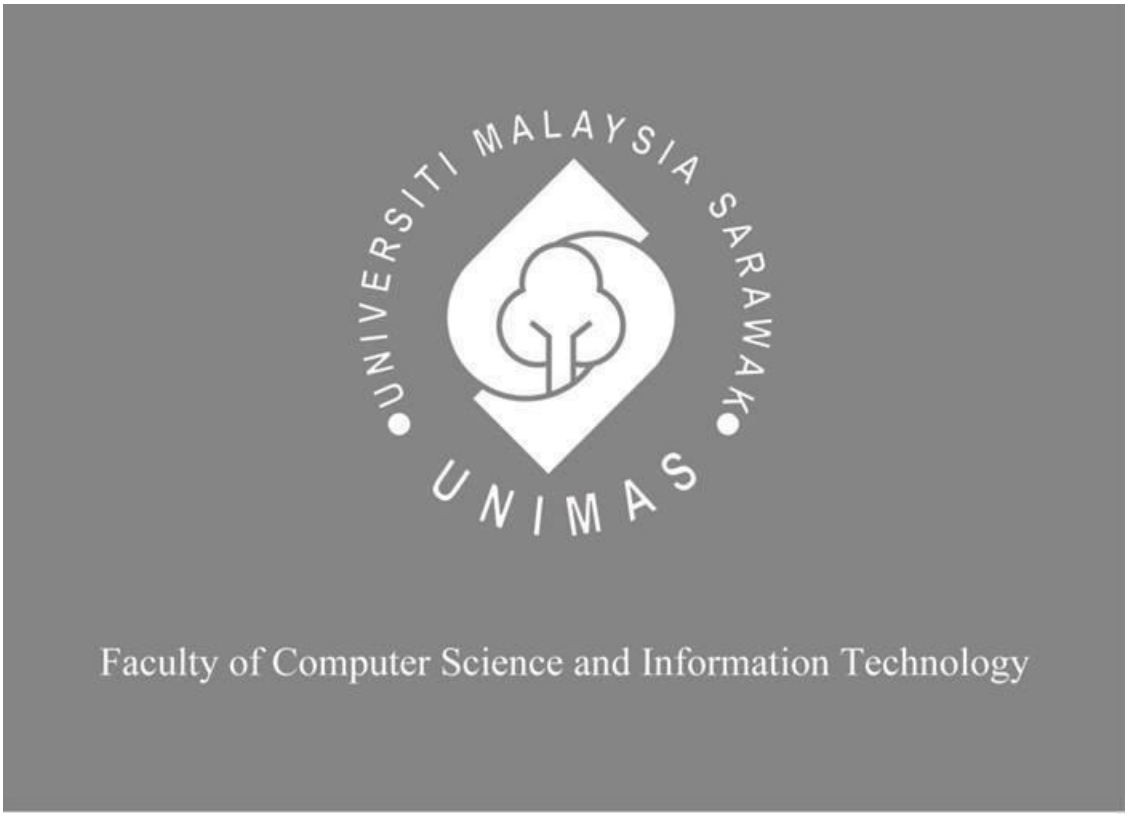


CLASSIFICATION OF MALARIA INFECTED ERYTHROCYTES USING IMAGE PROCESSING

Chang Jo Yee

Bachelor of Computer Science with Honours

(Multimedia Computing)

2020

CLASSIFICATION OF MALARIA INFECTED ERYTHROCYTES USING IMAGE PROCESSING

CHANG JO YEE

This project is submitted in partial fulfilment of the
requirements for the degree of
Bachelor of Computer Science and Information Technology

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2020

KLASIFIKASI ERITROSIT YANG DIJANGKITI MALARIA MENGGUNAKAN PEMPROSESAN IMEJ

CHANG JO YEE

Projek ini merupakan salah satu keperluan untuk
Ijazah Sarjana Muda Sains Komputer dan Teknologi Maklumat

Fakulti Sains Komputer dan Teknologi Maklumat
UNIVERSITI MALAYSIA SARAWAK
2020

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE CLASSIFICATION OF MALARIA INFECTED ERYTHROCYTES USING IMAGE PROCESSING

ACADEMIC SESSION: 2019 / 2020

CHANG JO YEE
(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

☐	CONFIDENTIAL	(Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)
☐	RESTRICTED	(Contains restricted information as dictated by the body or organization where the research was conducted)
☑	UNRESTRICTED	

Validated by

(AUTHOR'S SIGNATURE)

(SUPERVISOR'S SIGNATURE)

Permanent Address

Q1-02-12 GOODYEAR COURT 8,
JALAN USJ 14/1,
47630 SUBANG JAYA, SELANGOR

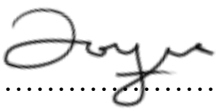
Date: 27.7.2020

Date: 27.7.2020

Note * Thesis refers to PhD, Master, and Bachelor Degree
** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

DECLARATION

I hereby declare that the work entitled Classification of Malaria Infected Erythrocytes using Image Processing submitted to the Faculty of Computer Science and Information Technology (FCSIT) is my original work under the guidance of my supervisor, Dr. Chai Soo See.



…….. 15/7/2020 ……..

(Chang Jo Yee) Date

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor Dr. Chai Soo See, who has provided invaluable guidance and continuous support in conducting this project. I also wish to thank my family and friends for their encouragement and support.

ABSTACT

Malaria parasites are known to have caused deaths worldwide, especially in African regions where resources are limited. Currently, malaria diagnoses are still done manually by trained experts. Hence, the use of computer-aided detection systems for malaria detection or identification in erythrocytes is a valuable approach in reducing the need for human resources. This project aims to use image processing to extract the features of infected erythrocytes and study three commonly used machine learning techniques in order to compare their performance in classifying malaria infected erythrocytes.

ABSTRAK

Parasit malaria diketahui menyebabkan kematian di seluruh dunia, terutamanya di kawasan Afrika yang mempunyai sumber daya terhad. Pada masa kini, diagnosis malaria masih dilakukan secara manual oleh pakar terlatih. Oleh itu, penggunaan sistem pengesanan komputer untuk pengesanan malaria di dalam eritrosit adalah pendekatan yang amat berharga demi mengurangkan keperluan sumber manusia. Projek ini bertujuan untuk menggunakan pemprosesan imej untuk mengekstrak ciri-ciri eritrosit yang dijangkiti malaria dan mengkaji tiga teknik pembelajaran mesin yang biasa digunakan untuk membandingkan prestasi mereka dalam pengklasifikasian eritrosit yang dijangkiti malaria.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

WHO	World Health Organization
CAD	Computer-Aided Detection
SVM	Support Vector Machine
k-NN	k-Nearest Neighbour
CNN	Convolutional Neural Network
CDC	Centre for Disease Control
ANOVA	Analysis of Variance
NIH	National Institute of Health
ReLU	Rectified Linear Unit

CHAPTER 1: INTRODUCTION

1.1 Introduction

Malaria parasites, scientifically named Plasmodium spp., are transmitted via female Anopheles' bites and infect the blood of victims. Besides, malaria transmission could also occur through organ transplant, needle or syringe sharing, and blood transfusion of contaminated blood. Out of more than hundred species of malaria parasites, five are known to infect humans: Plasmodium vivax, Plasmodium malariae, Plasmodium falciparum, Plasmodium ovale, and Plasmodium knowlesi (Poostchi, Silamut, Maude, Jaeger, & Thoma, 2018). It was reported that there were 219 million cases of malaria and an estimated 435000 deaths in 2017, with a high concentration of occurrence in African regions (World Health Organization (WHO), 2018).

This project aims to use image processing to extract the features of infected erythrocytes and study three commonly used machine learning techniques in order to compare and analyse them in classifying malaria infected erythrocytes.

1.2 Problem Statement

The current method widely used in classifying malaria infected erythrocytes is through manual classification under a microscope by health personnel. Usually, this requires blood samples to be transported to laboratories whereby it might take days before a diagnosis is obtained due to limited resources. Manual classification also has risks of human error.

1.3 Objectives

i. To extract features in malaria infected erythrocytes using image processing.

ii. To adopt three commonly used machine learning algorithms for classifying malaria infected erythrocytes.

iii. To compare and analyse the three algorithms based on their elapsed time, sensitivity, specificity, and accuracy for classification of malaria infected erythrocytes.

1.4 Methodology

The methodology for image processing consists of digital image acquisition, image pre-processing, erythrocyte detection and segmentation, and feature extraction and selection. Three machine learning algorithms will then be selected for the next step of malaria parasite classification. Their performance in terms of elapsed time, sensitivity, specificity, and accuracy will be measured and compared.

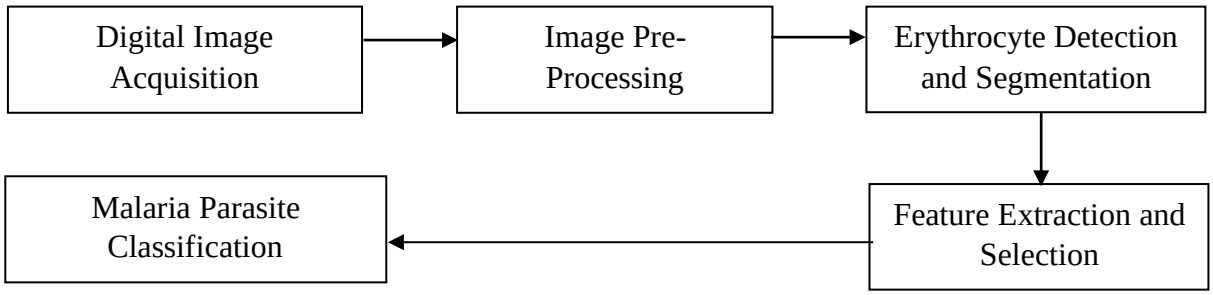


Figure 1.1. Image processing methodology

Once images of thin blood smears are acquired, they will then be pre-processed to improve the image quality and reduce variations that might complicate subsequent processing steps. To achieve this, the images may be resampled to a standardized size, undergo noise reduction using filtering, and lighting and colour normalization (Poostchi et al., 2018).

In erythrocyte detection and segmentation occurs the separation of image foreground (i.e. cells) and background, elimination of the background, and the segmentation of cell or

intercellular elements (Gamarra, Zurek, San-Juan, 2017). Other components in blood, namely leukocytes, platelets and other unwanted artifacts will be identified in this step and will not be processed further. Choosing a segmentation technique requires consideration of touching or overlapping cells that will complicate cell identification and segmentation.

Since each erythrocyte has varying circular shape, rotation, and scale, raw pixel information cannot be used. The feature extraction and selection step identifies the difference between features of uninfected and infected erythrocytes to be used in training the classifier (Savkare & Narote, 2012). According to Poostchi et al. (2018), extracted features types for thin blood smears include colour, texture, and morphologic characteristics.

Using the extracted and selected features with the best discriminating ability, the classification model is trained using a training dataset. It is then followed by using a test dataset to assess the models' performance (Gamarra et al., 2017). Typically used performance evaluations apply a statistical approach to describe diagnostic testing, namely sensitivity, specificity, and accuracy.

1.5 Scope

This project covers both research and development. The research covers the steps and the techniques in each step required in processing images of blood smears. It is proceeded with the classification of malaria infected and uninfected erythrocytes using three commonly used algorithms. Each algorithm will be compared for their elapsed time, sensitivity, specificity, and accuracy in the classifications. Prototype development will be done for the adoption of the algorithms for the purpose of data collection and comparison.

1.6 Significance of Project

Malaria is endemic to African regions with limited financial and human resources. This project aims to improve the methods of malaria diagnosis by reducing the needed resources through automated malaria diagnosis.

1.7 Project Schedule

	Task Name	Duration	Start	Finish
1	**Final Year Project**	**229 days**	**Fri 9/13/19**	**Wed 4/29/20**
2	**Final Year Project 1**	**90 days**	**Fri 9/13/19**	**Thu 12/12/19**
3	**Project Proposal**	**36 days**	**Fri 9/13/19**	**Sat 10/19/19**
4	**Brief Proposal**	**16 days**	**Fri 9/13/19**	**Sun 9/29/19**
5	Define problem and objectives	16 days	Fri 9/13/19	Sun 9/29/19
6	**Full Proposal**	**19 days**	**Mon 9/30/19**	**Sat 10/19/19**
7	Await brief proposal approval	6 days	Mon 9/30/19	Sun 10/6/19
8	Amend brief proposal	4 days	Sun 10/6/19	Thu 10/10/19
9	Determine scope and methodology	9 days	Thu 10/10/19	Sat 10/19/19
10	**Chapter 1: Introduction**	**6 days**	**Sun 10/20/19**	**Sat 10/26/19**
11	**Chapter 2: Literature Review**	**20 days**	**Sun 10/27/19**	**Sat 11/16/19**
12	Review journal articles and other documentations	7 days	Sun 10/27/19	Sun 11/3/19
13	Analyse existing techniques	13 days	Sun 11/3/19	Sat 11/16/19
14	**Chapter 3: Methodology**	**19 days**	**Sun 11/17/19**	**Fri 12/6/19**
15	Analyse methodologies in image processing, data collection, and speed and accuracy calculation	9 days	Sun 11/17/19	Tue 11/26/19
16	Determine requirements of prototype	10 days	Tue 11/26/19	Fri 12/6/19
17	**Review and Finalize Chapter 1, 2 and 3**	**5 days**	**Fri 12/6/19**	**Wed 12/11/19**
18	**Submit Final Year Project 1 Report**	**0 days**	**Thu 12/12/19**	**Thu 12/12/19**
19	**Final Year Project 2**	**138 days**	**Fri 12/13/19**	**Wed 4/29/20**
20	**Chapter 4: Implementation and Testing**	**116 days**	**Fri 12/13/19**	**Tue 4/7/20**
21	Train, test and evaluate classification models	72 days	Fri 12/13/19	Sun 2/23/20
22	Develop matlab prototype	50 days	Mon 2/17/20	Tue 4/7/20
23	**Chapter 5: Conclusion and Future Work**	**7 days**	**Wed 4/8/20**	**Wed 4/15/20**
24	**Review and Finalize Project**	**12 days**	**Thu 4/16/20**	**Tue 4/28/20**
25	**Submit Final Year Project Report, Source Code, Installation Kits, User Manual and Paper**	**0 days**	**Wed 4/29/20**	**Wed 4/29/20**

Figure 1.2. Gantt chart task table for project timeline

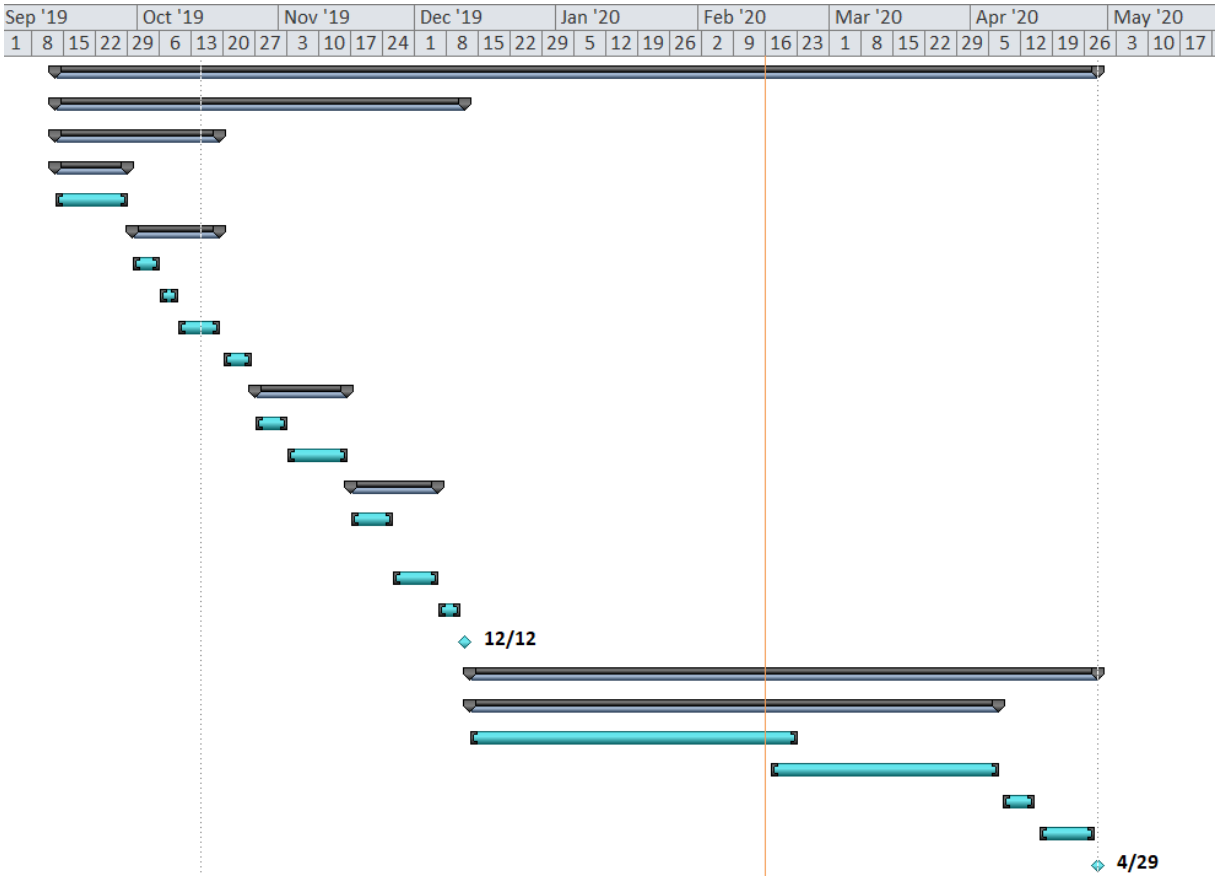


Figure 1.3. Gantt chart for project timeline

1.8 Expected Outcome

A comparison of three commonly used machine learning algorithm and a MATLAB prototype for classification of malaria infected erythrocyte.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

At present, there are many existing computer-aided detection (CAD) systems for malaria detection or identification in erythrocytes which mostly follow the pipeline of image pre-processing, segmentation, feature extraction, and classification. Each step uses a combination of techniques, and the varying techniques used in each step greatly impacts the final classification outcome, including errors that can propagate to the next step. In malaria CAD, thick and thin blood smears are used, but because of the ability of detecting malaria species and life cycle stage besides malaria detection, thin blood smears are more widely researched. In order to better understand the requirements of the project, Classification of Malaria Infected Erythrocytes Using Image Processing, this chapter covers the review of existing systems and the study of other related information.

2.2 Existing Systems

Four existing malaria CAD are selected based on their classification techniques, namely support vector machine (SVM), k-nearest neighbour (k-NN), and convolutional neural network (CNN).

2.2.1 Parasite detection, life cycle stage and species identification using SVM

Savkare and Narote (2015) proposed a MATLAB application that could detect malaria infected erythrocyte and identify the life cycle stage and malaria species using SVM classifier. The processes involved in the system include image pre-processing, cell segmentation, feature extraction, and classification. The dataset used by Savkare and Narote (2015) is Giemsa-stained thin blood smears, with a constraint for infected blood smears consisting of only P. vivax and P. falciparum species.

The image pre-processing techniques used by Savkare and Narote (2015) consists of converting RGB image to grayscale, removing unwanted pixels using Median filter, and smoothening and enhancing edges using Laplacian filter with second order derivative.

Next, in the step of segmenting cells from the background, Laplacian filtering and Otsu's thresholding are performed. Besides, watershed transform is also applied to separate overlapping or touching cells which in turn will aid erythrocyte counting. Using real-world analogy, watershed is the divide lines that separate adjacent catchment basins. In digital images, the catchment basins contain a group of connected pixels, where the divides are watersheds. The segmentation effects can be seen in Figure 2.1.

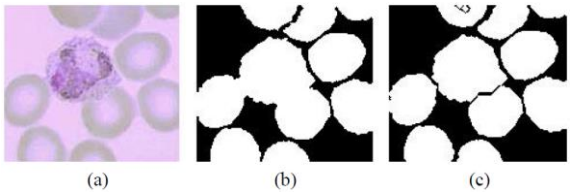


Figure 2.1. (a) Original image, (b) Binary mask of segmented cells, (c) Separation of overlapping cells (Savkare & Narote, 2015)

Then, in feature extraction, parameters of shape feature, texture, and colour are extracted and used to train the SVM classifier. As mentioned by Savkare and Narote (2015), radial basis

function (RBF) kernel is used in the classifier for the best results due to its greater flexibility in fitting data and its ability to model other kernels. After training the classifier, the system was tested, and the results obtained are 99.43% correct infected erythrocytes detection, 96.42% correct life cycle stage identification, and 80% correct species identification.

2.2.2 SVM based automated malarial parasite recognition

Another system that uses SVM was proposed by Muralidharan, Dong, and Pan (2016). Their approach consists of segmentation, feature extraction, feature selection, and SVM classification. Whole slide images of thin blood smears stained with Wright-Giemsa solution are used instead of conventional light microscopic images to standardize the image scale and quality. Because image quality is already standardized, pre-processing is not required; hence, in the first step, individual cells are segmented.

According to Muralidharan et al. (2016), an image will go through detection of connected components and thresholding is done to morphologically segment the cells. Areas larger than the threshold indicates overlapping cells. Hence, in order to segment them individually, Muralidharan et al. (2016) took advantage of the circular shape of erythrocytes and applied Hough transform to find circles, their centres, and radii. As shown in Figure 2.2, the segmented cells have many variations.

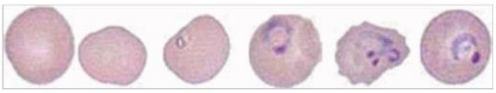


Figure 2.2. Normal and infected samples show variations (Muralidharan et al., 2016)

Once each cell is segmented, feature extraction and selection are performed. In this case, 76 features belonging to five categories were extracted: Haralick texture features, grey level run length matrix properties, histogram features, shape features, and Hu's moments. According

to Muralidharan et al. (2016), the purpose of grouping the features into categories is to overcome the curse of dimensionality, improve classifier performance, and reduce training time. After that, the features need to be narrowed down and filtered based on their discriminating ability, whereby a feature selection method of Kullback-Leibler distance is used.

The cell samples with the selected features are then used to train the SVM classifier, in which the classifier will find a hyperplane that best separates the two classes (i.e. infected and uninfected). As stated by Muralidharan et al. (2016), SVM with RBF kernel is used in case of non-separable data. Next, to assess the classifier performance, ten-fold cross validation is used with the intention of obtaining the model with least errors and also to prevent overfitting. Overall, the system achieved an accuracy of 95.5%, sensitivity of 96.43%, and specificity of 94.61%. It was mentioned by Muralidharan et al. (2016) that SVM is considered more robust than neural networks.

2.2.3 Detection of malaria species and life cycle stages using k-NN

Nanoti, Jain, Gupta, and Vyas (2016) conducted a study for detection of malaria parasites species and life cycle stage from images of thin blood smears utilizing k-NN as the classifier. The techniques implemented in image pre-processing, segmentation, feature extraction, and classification were discussed. According to Nanoti et al. (2016), the dataset of blood smear images consisting of erythrocytes infected with P. falciparum, P. vivax, P. ovale, and P. malariae was acquired from Centre for Disease Control (CDC) and WHO.

Firstly, the images underwent pre-processing such as geometric mean filter, median filter, and forward discrete curvelet transform. In this step, Nanoti et al. (2016) also took note of Schuffner's dot which are brown pigments caused by malaria infection and highlighted them using partial contrast stretching.