IMAGE ENHANCEMENT AND FEATURE EXTRACTION TECHNIQUES FOR INTELLIGENT THALASSEMIA SCREENING PROCEDURES

by

NURUL ZHAFIKHA BINTI NOOR RASHID
(1131310627)

A thesis submitted in fulfilment of the requirements for the degree of
Master of Science in Biomedical Electronics Engineering

School of Mechatronic Engineering
UNIVERSITI MALAYSIA PERLIS

2016

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful.

Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis. This thesis owes its existence to the help, support and inspiration of several people. First and foremost, I would like to express my sincere appreciation to my supervisor, Prof. Dr. Yusoff Mashor for his supervision during this research. His guidance and motivations always keep me focused on the research objectives and have been precious in the development of this thesis content.

I would also like to express my appreciation to the medical supervisor Prof. Dr. Rosline Hassan, haematologist of Hospital Universiti Sains Malaysia and all the staff for providing the constructive guidance during the research study. I would like to convey my gratitude to the Ministry of Higher Education (MOHE) for the scholarship that helps my aspiration became possible.

I would like to express my deepest gratitude to my beloved husband, Muhammad Shafiq and my precious daughter, Hana Insyirah for keep believe in me and bear with me all the times. Not forgotten to my beloved parents, Noor Rashid Hamid and Zaleha Abd Rahman, my beloved parents-in-laws, Mezan Kassim and Siti Salasiah, and the rest of my family for the prayer, love, motivation and encouragement that inspire me to strive harder in achieving the dreams.

My sincere thanks must also go to all my colleagues especially to Dr. Aimi Salihah, Dr. Rafikha Aliana and my fellow friends for all the support and guidance in terms of knowledge and motivation during this whole period. Your help and encouragement really means to me. Thank you very much.

Nurul Zhafikha bt Noor Rashid, January 2016.

TABLE OF CONTENTS

CHAPTER 4: RESULTS & DISCUSSION

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ALL	Acute Lymphocytic Leukemia
AML	Acute Myelogenous Leukemia
ANN	Artificial Neural Network
BMP	Bitmap
BR	Bayesian Regulation
CBC	Complete Blood Count
CFS	Correlation Feature Selection
DNA	Deoxyribonucleic Acid
FN	False Negative
FP	False Positive
Hb	Haemoglobin
Hct	Hematocrit
HSI	Hue, Saturation, Intensity
HUSM	Hospital Universiti Sains Malaysia
InfoGain	Information Gain
LM	Levenberg-Marquardt
LVQ	Learning Vector Machine
MCV	Mean Corpuscular Volume
MCH	Mean Corpuscular Haemoglobin
MCHC	Mean Corpuscular Haemoglobin Concentration
MLP	Multilayer Perceptron
PCA	Principle Component Analysis
PSNR	Peak-signal-to-noise-ratio

RBC	Red Blood Cell
RDW	Red cell Distribution Width
RGB	Red, Green, Blue
ROI	Region of Interest
SCG	Scale Conjugate Gradient
SFS	Sequential Forward Selection
SRGAE	Seed Region Growing Area Extraction
SVM	Support Vector Machine
TN	Truth Negative
TP	Truth Positive
WBC	White blood cell
WHO	World Health Organization

LIST OF SYMBOLS

A_c	Area of segmented RBC
SF_b	Bright stretching factor
B_c	Blue pixels of RBC
μ_{pq}	Central Moment
$f(x,y)$	Color level of image in the location of [x,y]
$in(x,y)$	Color level for input pixel
q_k	Color level of input pixel
$out(x,y)$	Color level for output pixel
P_k	Color level of output pixel
SF_d	Dark stretching factor
dL	Decilitre
max	Desired maximum color levels in output image
min	Desired minimum color levels in output image
fL	Femtolitre (a fraction of one-millionth of a litre)
$\nabla V(\underline{x})$	Gradient
gm	Grams
G_c	Green pixels of RBC
$\nabla^2 V(\underline{x})$	Hessian matrix
HN	Hidden node
$[i,j]$	Image pixel location
$f[i,j]$	Image pixel value
I	Input node

$\nabla^2 f$	Laplacian operator
minTH	Lower threshold value
μ	Marquardt adjustment parameter
maxB	Maximum color level of blue
maxG	Maximum color level of green
maxR	Maximum color level of red
maxRGB	Maximum color level of RGB
f_{max}	Maximum color level values in the input image
$\overline{B_c}$	Mean blue of RBC
$\overline{G_c}$	Mean green of RBC
$\overline{R_c}$	Mean red of RBC
$\overline{I_c}$	Mean intensity of RBC
μm	Micrometer
minB	Minimum color level of blue
minG	Minimum color level of green
minR	Minimum color level of red
minRGB	Minimum color level of RGB
f_{min}	Minimum color level values in the input image
m_{pq}	Moment
NminTH	New lower stretching value
NmaxTH	New upper stretching value
$out_{rgb}(x,y)$	New RGB pixel value
α	Normal alpha gene
β	Normal beta gene
β^o	No beta gene formation

$in_{RGB}(x,y)$	Original RGB pixel value
R_c	Red pixels of RBC
β^+	Some beta gene formation occur
θ	Theta
$maxTH$	Upper threshold value
$\sigma^2 B_c$	Variance blue of RBC
$\sigma^2 G_c$	Variance green of RBC
$\sigma^2 R_c$	Variance red of RBC
$\sigma^2 I_c$	Variance intensity of RBC

Teknik-teknik Peningkatan Imej dan Pengekstrakan Ciri untuk Prosedur Saringan Talasemia Pintar

ABSTRAK

Talasemia adalah penyakit darah yang diwarisi dan memberi kesan kepada pengeluaran sel darah merah (RBC) di dalam badan. Kekurangan RBC yang terdiri daripada 99% sel darah secara keseluruhan akan menjejaskan fungsi utama sebagai agen pembawa oksigen. Ciri morfologi RBC memainkan peranan penting dalam diagnosis penyakit tersebut. Buat masa ini, siasatan mikroskopik dijalankan secara manual dalam mengenal pasti kehadiran sel talasemia oleh pakar hematologi melalui mikroskop cahaya. Walau bagaimanapun, prosedur manual ini boleh menghasilkan keputusan yang kurang tepat, memerlukan tenaga kerja yang intensif dan memakan masa kerana ia terlalu bergantung kepada pengalaman dan kemahiran pakar hematologi. Oleh itu, objektif utama penyelidikan ini adalah membangunkan prosedur saringan talasemia pintar berdasarkan sampel darah. Satu kaedah yang efisien telah dibina dengan menggunakan teknik pemprosesan imej termasuklah peningkatan imej, peruasan imej dan pengekstrakan ciri bagi mendapatkan peruasan RBC sepenuhnya. Terdapat lima jenis teknik peningkatan kontras yang telah diaplikasikan ke atas imej asal RGB secara berasingan dan dua teknik dipilih untuk digunakan dalam kaedah cadangan iaitu teknik peningkatan kontras gelap dan separa. Seterusnya, peruasan sel darah diteruskan dengan penukaran kepada model warna HSI sebelum imej itu diproses dengan menggunakan kaedah kelompok purata-*k* bergerak dan kaedah penapisan median. Prestasi peruasan bagi sel darah yang baik telah diperolehi daripada kombinasi kaedah peningkatan kontras gelap dan komponen keamatan. Kemudian, imej RBC yang telah diruas sepenuhnya diperolehi selepas komponen lain yang tidak dikehendaki seperti RBC yang bersentuhan, WBC dan platelet berjaya dibuang dengan menggunakan kaedah pertumbuhan rantauan. Tiga ciri utama diekstrak daripada RBC individu berdasarkan bentuk ringkas, warna dan bentuk kompleks. Pengekstrakan ciri ini telah menghasilkan 31 ciri-ciri dan dijadikan sebagai data masukan dalam rangkaian perseptron berbilang lapisan. Terdapat dua proses klasifikasi RBC dalam kajian ini. Proses klasifikasi pertama adalah mengklasifikasikan jenis RBC kepada normal atau tidak normal (talasemia). Proses klasifikasi yang kedua pula adalah mengklasifikasikan sel normal dan tiga jenis talasemia yang dinamakan α-talasemia, β-talasemia trait dan β-talasemia. Algoritma *Levenberg-Marquardt* telah menghasilkan prestasi klasifikasi yang terbaik setelah dibandingkan dengan algoritma regulasi *Bayesian* dan algoritma skala konjugat kecerunan. Kemudian, ciri-ciri yang telah diekstrak itu akan diproses dengan menggunakan empat kaedah pemilihan ciri secara berasingan iaitu khi kuasa dua, kenaikan maklumat, korelasi ciri pemilihan dan analisis komponen utama. Peratus ketepatan klasifikasi yang tertinggi diperolehi bagi kedua-dua klasifikasi RBC dengan menggunakan ciri-ciri yang telah terpilih daripada penapis khi kuasa dua. Hal ini disebabkan oleh ciri-ciri terpenting telah digunakan sebagai data masukan. Prestasi klasifikasi yang optimum telah dicapai dengan kejituan pengesahsahihan dan kejituan ujian masing-masing dengan 98% dan 96.8% untuk klasifikasi yang pertama. Manakala, kejituan pengesahsahihan dan kejituan ujian masing-masing dengan 95.7% dan 94.4% diperolehi dalam klasifikasi yang kedua. Maka, pembangunan teknik peningkatan imej dan teknik pengekstrakan ciri RBC dalam penyelidikan ini memberikan satu alternatif yang cekap dalam menganalisis dan mengklasifikasikan sampel darah.

Image Enhancement and Feature Extraction Techniques for Intelligent Thalassemia Screening Procedures

ABSTRACT

Thalassemia is an inherited blood disease that effects the production of red blood cells (RBC) in the body. The deficiency of RBC that constitutes 99% of blood cells will affect their main function as oxygen carrier. The morphological features of RBC play a crucial role in medical diagnosis. Currently, the microscopic investigation to identify the present of any thalassemia cells is performed manually by haematologists through visual identification under a light microscope. However, the manual procedure yields inaccurate results, labour-intensive and time-consuming since it is highly dependent on the haematologists experience and skill. Thus, the main objective of this research is to develop an intelligent thalassemia screening procedure based on blood samples. Essentially an efficient method using the image processing techniques including image enhancement, segmentation and feature extraction have been constructed in order to obtain a fully segmented RBC. There are five contrast enhancement techniques have been applied to the original RGB image, separately and two techniques were selected to be used in the proposed procedure that are Dark Contrast and Partial Contrast techniques. Then, the segmentation of blood cells proceed with the conversion of HSI color space before the image being processed using moving *k*-mean clustering and median filter techniques. Good segmentation performance for blood cell has been obtained from the combination of dark contrast technique and intensity component. Then, fully segmented RBC image was obtained after the unwanted components such as overlapping RBC, WBC and platelet were successfully removed using seed region growing technique. Next, three main features were extracted from the individual RBC that are simple shape, color and complex shape based features. The features extraction has produced 31 features that have been fed as inputs to the Multilayer Perceptron (MLP) network. There are two processes of RBC classification have been performed in this research. The first process was carried out to classify the type of RBC into normal or abnormal (thalassemia). Then, the second process was continued to classify the normal blood and three types of thalassemia cells namely α-thalassemia, β-thalassemia trait and β-thalassemia. The Levenberg-Marquardt (LM) algorithm produces the best classification performance compared to Bayesian Regulation (BR) and Scale Conjugate Gradient (SCG) algorithms. Next, the extracted features were processed using four feature selection methods that are Chi-Squared, Information Gain, Correlation-based Feature Selection and Principal Component Analysis, separately. The highest classification accuracy was obtained for both RBC classifications using the selected features from Chi-Squared filter. This is due to selected significant features were used as inputs in the network. The optimal classification performance has been achieved with validation accuracy of 98% and testing accuracy of 96.8% for first classifier (11 features). While, validation and testing accuracies of 95.7% and 94.4%, respectively have been achieved in the second classifier (14 features). Therefore, the development of image enhancement and feature extraction techniques in this research provides an efficient alternative in analyzing and classifying the blood sample.

CHAPTER 1

INTRODUCTION

1.1 Background

Haemoglobinopathies, or haemoglobin disorder, is one of the inherited blood diseases characterized by the presence of abnormal haemoglobin in the blood. The cause of this abnormality are due to the inheritance of mutated genes that interfere with the formation of haemoglobin molecules in the red blood cells with its primary function as a transporter of oxygen to the whole body (Ferry, 1923). If the oxygen has not been transported sufficiently, it will cause further damage or functional disability of certain organs and tissues.

World Health Organization (WHO) estimates that approximately 2.9% of thalassemia and 2.3% of sickle cell disease are the carriers of the abnormal haemoglobin condition (Memish, Owaidah & Saeedi, 2011). In fact, the prevalence of these diseases is greater than cystic fibrosis and haemophilia (Anionwu & Atkin, 2001). Over 300,000 to 500,000 babies are died each year due to the severe haemoglobin disorders (WHO, 2011). As a result of global migration patterns, the diseases are spreading aggressively throughout the countries including Malaysia. Figure 1.1 shows the global distribution of haemoglobinopathies in terms of births per 1,000 affected infants.