THALASSAEMIA DETECTION USING CBR ALGORITHM VIA MOBILE DEVICEE

NUR FAEZAH BINTI OMAR

A thesis submitted in partial fulfilment of the requirements
For the award of the degree of
Bachelor of Computer Science (Software Engineering)

Faculty of Computer System & Software Engineering
UNIVERSITI MALAYSIA PAHANG

MAY 2011

ABSTRACT

This thesis proposes a Case-Based Reasoning model for medical diagnosis, particularly for Thalassaemia diagnosis. CBR Algorithm is an algorithm that can solve a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation in a new situation. CBR is suit for solving the problem in Thalassaemia cases. The model is designed and prototype is developed to test the diagnosis accuracy of the model. This application is focus on Beta Thalassaemia and Haemoglobin E Trait only. For platform, this application will be using the mobile device such as PDA. Moreover, this application will be using the programming language Visual Basic.Net applied in Visual Studio 2008. The methodology has chosen is rapid application development (RAD) where this method is the archive with apparel search requirement. The results show that the Case-Based Reasoning model has a great potential to be implemented in diagnosing Thalassaemia cases

ABSTRAK

Tesis ini mencadangkan sebuah model penaakulan berasaskan kes untuk diagnosis perubatan, terutamanya untuk diagnosis Thalassaemia. CBR Algoritma adalah suatu algoritma yang dapat menyelesaikan masalah baru dengan menggunakan situasi yang sama sebelum itu dan dengan menggunakan semula maklumat dan pengetahuan ke dalam situasi yang baru. CBR adalah satu teknik yg sesuai untuk menyelesaikan masalah terutama sekali untuk penyakit Thalassaemia. Model ini direka bentuk dan prototaip dibangunkan untuk menguji ketepatan diagnosis model. Aplikasi ini tertumpu kepada Beta Thalassaemia dan Hemoglobin E-Trait. Untuk platform, aplikasi ini akan menggunakan PDA. Selain itu, aplikasi ini akan menggunakan bahasa pengaturcaraan Visual Basic.Net dan dibangunkan menggunakan Visual Studio 2008. Metodologi yang digunakan untuk melaksanakan projek ini adalah Rapid Application Development atau dikenali sebagai "RAD". Keputusan kajian menunjukkan bahawa model penaakulan berasaskan kes mempunyai potensi besar untuk diterapkan dalam mendiagnosis kes Thalassaemia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX

LIST OF ABBREVIATIONS

ANCRONYM	MEANING
AI	Artificial Intelligence
BP	Back-Propagation
CBR	Case-Based Reasoning
CE	Compact Edition
HBA2	Haemoglobin A2 Estimation
Hb	Haemoglobin
HBF	Foetal Haemoglobin
HTAA	Hospital Tenku Ampuan Afzan
MCH	Mean Corpuscular Haemoglobin
MCHC	Mean Corpuscular Haemoglobin Concentration
MCV	Mean Corpuscular Volume of red Cells
PDA	Personal Digital Assistant
PCV	Packed Cell Volume
RAD	Rapid Application Development
RBR	Rule-Based Reasoning
TDA	Thalassaemia Detection Application
TRBC	Total Red Blood Count
VB	Visual Basic

CHAPTER 1

INTRODUCTION

1.1 Introduction

Thalassaemia is a term that refers to a group of genetic disorders characterized by insufficient production of haemoglobin. There are two proteins involved in the production of haemoglobin, alpha and beta. If there is a deficiency in either of these proteins the red blood cells do not form properly and cannot carry adequate amounts of oxygen to all parts of the body. This then results in organs that are starved for oxygen and unable to function properly. Because of this, patients have to get blood transfusions, usually every two to three weeks. These blood transfusions are done at a hospital and can take anywhere from six to eight hours or more. After some time, the blood cells break down and leave iron in the patient's body. This iron will bind to the major organs of the body, such as the liver or heart. If left alone, the iron will overload these organs until they will not be able to do their job creating other health problems for the patient.

In Malaysia, one out of 20 people are carriers of the disease without realizing it. This accounts for 600,000 to a million Thalassaemia carriers. There were 4,768 Thalassaemia patients as of last year based on the National Thalassaemia Register

launched in 2007. In Malaysia, there were 345 clinics nationwide with the equipment to conduct screening. Those who should undergo the screening are those who have a family history of Thalassaemia, teenagers and young adults, in order of priority. Teenagers and young adults were the most suitable for undergoing the blood test screening (New Straits Times, 2010)

There are two types of Thalassaemia which is Thalassaemia minor and Thalassaemia major. Thalassaemia minor is a genetic blood condition. Patients with Thalassaemia minor are sometimes said to have "Thalassaemia trait," and they are often non-symptomatic. Although someone with this condition may not experience adverse symptoms, the trait can be passed on to a child, and if the other parent also carries the trait, the child could develop Thalassaemia minor by inheriting a bad gene from one parent, or a more severe form of the disease by inheriting the gene from both parents. Patients diagnosed with Thalassaemia minor do not exhibit symptoms and are simply carriers of the defective gene. Thalassaemia major, also known as beta Thalassaemia major or Cooley's anaemia, is a genetic blood disorder that causes the body to manufacture an abnormal form of haemoglobin. Symptoms of Thalassaemia major that a parent may notice include poor appetite, and increased infections. As the child matures, other symptoms may include delayed growth, bone deformities in the face, and an extended abdomen caused by liver and spleen swelling. Without treatment, the blood disorder can result in heart failure and liver problems. Thalassaemia major can be diagnosed using blood tests.

For this application named as "Thalassaemia Detection using Case-Based Reasoning via Mobile Device", this application is build for the clinical staff especially the haematologist in the Hospital Tengku Ampuan Afzan (HTAA) to detect their patient who has the Thalassaemia disease. The main purpose of building this application is to make easier for the clinical staff especially the haematologist in the HTAA to detect whether the patients has Thalassaemia disease or not.

For the platform, this application will be using the mobile devise that operate in Window Mobile operating system. The best device that suit for this application is Personal Digital Assistant (PDA). PDA is a mobile device that functions as a personal information manager. Most PDA can synchronize the data with applications on a user's personal computer. This allows the user to update all the information on their computer.

The main purpose of using the PDA is to make easier for the clinical staff especially the haematologist to bring the device everywhere because of PDA is smaller, portable to bring everywhere and lighter than laptop. Besides that, they can easily detect and know whether the patient has the Thalassaemia disease or not by inserting all the data's that required for computing and calculating process.

This application will be using the Case-Based Reasoning (CBR) algorithm. CBR is Artificial Intelligence (AI) approach learning and problem solving based on past experience. A new problem is solved by finding a similar past case and reusing it in the new problem situation by reusing information and knowledge of that situation (Aamoldt, 1994). CBR also is an approach to incremental, sustained learning, since a new experience is retained each time a problem has been solved, making it immediately available for future problems. . A previously experienced situation, which has been captured and learned in a way that it can be reused in the solving of future problems, is referred to as a past case, previous case, stored case, or retained case. Correspondingly, a new case or unsolved case is the description of a new problem to be solved. CBR is not only a powerful method of computer reasoning but also pervasive behaviour in everyday human problem solving.

1.2 Problem Statement

There are some problems occur in real life especially in the laboratory. The first example situation is while doing the diagnosing process in the laboratory, the haematologist not straight away gets the result whether the patients have the Thalassaemia disease or not. The haematologist needs to wait for the result until the screening test had been done in the laboratory. By doing this application, the haematologist can predict the result earlier before the screening test had been done in the laboratory. It will decrease the time while waiting for the result.

The second example situation is that not all the haematologist bring along their laptop all the time with them in the hospital. The problem that might be occurring from this situation is when the patient wants to ask the haematologist about the result after diagnosed. Patient had to wait for the result whether they have the Thalassaemia disease or not. By applying this application using mobile device which is smaller, portable to bring everywhere and lighter than laptop, patients can check through the haematologist and easily know whether they have the Thalassaemia disease or not. The haematologist can predict whether the patient has the Thalassaemia disease or not. Besides that, it can save their time while waiting for the answer.

1.3 Objective

In order to develop an application of Thalassaemia Detection by using the CBR algorithm via mobile device, the overall objectives of this system are:

i. To develop a mobile application to detect a Thalassaemia disease.
ii. To apply the CBR technique in medical diagnosis for classifying the Thalassaemia disease.
iii. To evaluate the diagnose accuracy using real cases obtained from Hospital Tenku Ampuan Afzan (HTAA).

1.4 Project Scope

The project scopes that have been identified for this project are divided into three categories which are user module, data and system environment. The user in this system is the Doctor especially the haematologist in the hospital. The data and requirement are taken from Pathology Department in Hospital Tenku Ampuan Afzan (HTAA) which is located in Kuantan Pahang. This application is focus on Beta Thalassaemia Haemoglobin and E Trait only. This application will be using the CBR algorithm which is focus on retrieves and reuse process only. For platform, this application will be using the mobile device such as PDA. Moreover, this application will be using the programming language VB. Net applied in Visual Studio.

1.5 Thesis Organization

This thesis is divided into 5 chapters which is chapter 1 is about the introduction of the system. In this chapter, the problem statement, objective and scope will be identified. For Chapter 2, this chapter will discuss about the review of the literature that related to the system. For Chapter 3, this chapter will discuss about the materials and methods that will be using in the system. This chapter is all about the framework and approach of the project. It explains about software and also the hardware that used to develop the system. For Chapter 4, it will discuss about the implementation of the system. For Chapter 5, it will discuss about the result and discussion of the system. The result included the project limitation, result analysis and suggestion for the project enhancement. While for the last chapter which is Chapter 6, it will discuss about the conclusion of the system. This chapter will briefly summarize the overall developed project.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The Thalassaemias are a group of genetic or inherited blood disorders that share in common one feature, the defective production of haemoglobin, the protein that enables red blood cells to carry oxygen and carbon dioxide. There are many different disorders with defective haemoglobin synthesis and, hence, many types of Thalassaemia. Thalassaemia is never 'caught' by another person in the way that a cold or flu is transmitted. People with Thalassaemia disease and trait are born with it.

The most familiar type of Thalassaemia is Beta Thalassaemia. It involves decreased production of normal adult haemoglobin (Hb A), the predominant type of haemoglobin from soon after birth until death. All haemoglobin consists of two parts which are heme and globins. The globins part of Hb A has four protein sections called polypeptide chains. Two of these chains are identical and are designated the alpha chains. The other two chains are also identical to one another but differ from the alpha chains and are termed the beta chains. In persons with Beta Thalassaemia, there is reduced or absent production of Beta globins chains.

There are two forms of Beta Thalassaemia. They are Thalassaemia minor and Thalassaemia major which is also called Cooley's anaemia. For Thalassaemia minor, the individual with Thalassaemia minor has only one copy of the Beta Thalassaemia gene together with one perfectly normal beta-chain gene. The person is said to be heterozygous for Beta Thalassaemia. Persons with Thalassaemia minor have at most mild anaemia with slight lowering of the haemoglobin level in the blood. This situation can very closely resemble that with mild iron-deficiency anaemia. However, persons with Thalassaemia minor have a normal blood iron level unless they have are iron deficient for other reasons. No treatment is necessary for Thalassaemia minor. In particular, iron is neither necessary nor advised.

Thalassaemia major which is also called Cooley's anaemia, the child born with Thalassaemia major has two genes for Beta Thalassaemia and no normal beta-chain gene. The child is homozygous for Beta Thalassaemia. This causes a striking deficiency in Beta chain production and in the production of Hb A. Thalassaemia major is, therefore, a serious disease (Medicine net. Inc, 2010).

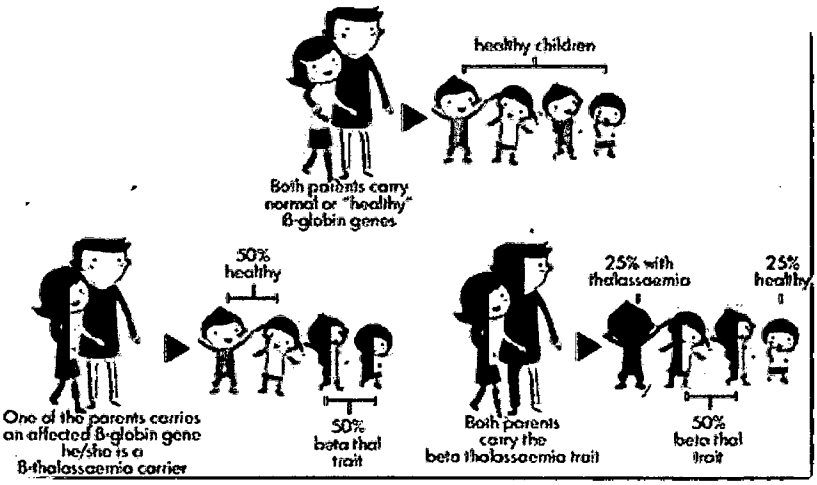

Figure 2.1: Thalassaemia Traits (Medi-Vance Healthcare, 2009)

Individuals with Beta Thalassaemia trait usually have evidence of microcytosis and increased levels of haemoglobin A2. Haemoglobin F is sometimes elevated as well. Individuals with alpha Thalassaemia trait usually have evidence of microcytosis and normal levels of haemoglobin A2 and F. Sometimes trait status cannot be determined by these screening tests alone. Results may be ambiguous for a variety of reasons. If iron deficiency is detected, an individual must be retested after completing iron supplement therapy. Occasionally, DNA testing that directly examines the alpha and/or beta globins genes is necessary. DNA testing is the only way to determine silent alpha Thalassaemia trait and the related haemoglobin trait called haemoglobin Constant Spring. DNA testing may also be necessary in order to allow for the option of prenatal testing (thalassemia.com, 2003)

2.2 Recent Works

After doing researches regarding the requirements of the application, a study of current system was conducted for comparison and inspiration. The study was conducted on current information of Case-Based Reasoning (CBR) algorithm.

2.2.1 Lockheed - CLAVIER

The first commercially fielded CBR application was at Lockheed, Palo Alto (Hennessy & Hinkle, 1992). Modern aircraft contain many elements that are made up from composite materials. These materials require curing in large autoclaves. Lockheed, the US aerospace company, produce many such parts. Each part has its own heating characteristics and must be cured correctly. If curing is not correct the part will have to be discarded. Unfortunately, the autoclave's heating characteristics are not fully understood. This is complicated by the fact that many parts are fired together in a single large

autoclave and the parts interact to alter the heating and cooling characteristics of the autoclave (Watson and Marir, 1995).

Operators of Lockheed's autoclaves relied upon drawings of previous successful parts layouts to inform how to layout the autoclave. However, this was complicated by the fact that layouts were never identical because parts were required at different times and because the design of the composite materials was constantly changing. Consequently operators had to select a successful layout they thought closely matched and adapt it to the current situation. This closely resembled the CBR paradigm and when Lockheed decided to implement a KBS to assist the autoclave operators they decided upon CBR. Their objectives were to:

- Reuse previously successful loadings
- Reduce the pressure of work on one or two experts,
- Secure the expertise of the experts as a corporate asset,
- Help to train new personnel.

The development of CLAVIER started in 1987, and it has been in regular use since the autumn of 1990. CLAVIER searches a library of previously successful autoclave layouts. Each layout is described in terms of:

- Parts and their relative positions on a table
- Tables, and their relative positions in the autoclave, and
- Production statistics such as start and finish times, pressure and temperature

CLAVIER finds substitutes for parts in a layout that do not match, and it recommends new layouts to operators. In adapting new layouts from previous ones CLAVIER:

- Creates new layouts by adapting pieces of previous layouts
- Minimizes the number of required parts not included in the layout,
- Maximizes the number of high-priority parts included in the layout, and
- Maximizes the total number of parts in a layout.

CLAVIER acts as a collective memory for Lockheed and as such provides a uniquely useful way of transferring expertise between autoclave operatives. In particular the use of CBR made the initial knowledge acquisition for the system easier. Indeed, it is doubtful if

it would have been possible to develop a MBR system since operatives could not say why a particular autoclave layout was successful. CLAVIER also demonstrates the ability of CBR systems to learn. The system has grown from 20 to over 150 successful layouts and its performance has improved such that it now retrieves or adapts a successful autoclave layout 90% of the time.

2.2.2 British Airways - CaseLine

CaseLine is a first generation technology demonstrator used by British Airways (BA) to assess the potential of CBR (Magaldi, 1994). CaseLine assists Boeing 747-400 technical support engineers with aircraft fault diagnosis and repair between aircraft arrival and departure. It advises on past defects and known successful recovery and repair procedures. When a fault in a Boeing 747-400 is detected or suspected, either by monitoring equipment or the pilots, details are transmitted to ground staff. The plane may only be scheduled to be on the runway for one hour during which time engineers have to identify the cause of the fault and effect repair. This is complicated because defects are often obscure and have complex and inconsistent causes. To delay the plane will disrupt schedules and costs thousands or pounds per minute. To let the plane take off with an unresolved fault could have catastrophic consequences (Watson and Marir, 1995).

CaseLine is implemented in ReMind. Users can input diagnostic information and control the search for available repair and recovery information. The system contains around 200 cases in early 1994 that describe previous failure instances and details of successful recovery actions. Three main search modes are provided:

- ATA Chapter - a simple two digit number referring to a fault in the plane's maintenance manual,
- EICAS Message - a precise but variable length alphanumeric text indicating a fault, and
- Reported Defect - a variable length string describing a fault.

These can be used alone or together for case retrieval using either nearest neighbour or inductive retrieval. Usually, a single case is retrieved if an exact match has been specified or several cases if a partial match was required. CaseLine helps engineers identify procedures that have the highest likelihood of success. The engineer is still obliged to use the aircraft maintenance manuals as a final authority and to follow approved procedures. But CaseLine does reduce costly delays by cutting out less productive routes to fault analysis and fault finding.

Initial assessment by BA states that "CBR has a set of in-built capabilities that complement a specific range of engineering problems. These are by nature, often more than just technical in origin, and require an awareness of many competing human, organizational and operational factors when posing solutions" (Magaldi, 94). In particular BA identifies three benefits of CBR:

- CBR is intuitive to both developers and users,
- CBR complements human reasoning and problem solving, and
- CBR retains the rich context of a problem situation - they discard nothing but simply index on different features of what they store.

The latter point is of particular legal interest. If an rule-based diagnostic system were developed its rules would represent distillate knowledge. The original reasons why a rule was created may become obscured with time. However, episodic cases are always heavily contextualized. If BA were sued for negligence using CaseLine they could demonstrate that engineers had followed procedures that had proved successful in case X - a simple defence. However, if BA used a rule-based system expert witnesses might have to prove that a rule was theoretically correct - a more complex defiance.

2.2.3 Legal & General – SWIFT

Legal & General (L&G) are a major UK provider of financial services. Their IT department has an annual budget of around 60 million. In 1993 the company was in the

process of down-sizing from dumb terminals attached to mainframes to PC based LANs. As part of a business process reengineering project they wanted to provide a streamlined service to employees purchasing PC's, peripherals, software or upgrades (Ian Watson and Farhi Marir, 1995).

The system was developed very rapidly using a KBS development methodology called the Client Cantered Approach (CCA). The CCA combines the linear stages of the Waterfall approach to software development with the iterative prototyping methods popular with KBS developers. The CCA explicitly encourages the involvement of all stakeholders in a project and emphasizes the need to consider maintaining the system.

A week was spent in April 1993 analyzing the existing process and reengineering it. This resulted in a very much simplified design whereby all L&G's employees would have access to a single point of contact for ordering PC products and upgrades. An essential component of the reengineered process was a KBS that would contain knowledge about L&G's IT strategy, their approved product range and the hardware or software options that different business units used.

Because of the volatile nature of the PC market it was decided that it would be important to make the maintenance of the KBS as easy as possible. It was decided that the managers of the service ideally should be able to maintain the knowledge-base themselves. Consequently, CBR was chosen as the knowledge representation paradigm that would best meet these constraints (Vargas and Raj, 93). Inference's CBR-Express/Case Point combination, Asymetrix Toolbook and Microsoft's Access were chosen to develop the demonstration system. After approximately one month's prototyping a demonstration system, called SWIFT, was shown at several seminars to stakeholders from all the business units within L&G.

These presentations were carefully organized as part of a comprehensive communications plan. They were professionally conducted and involved describing why the existing process had to be reengineered, what benefits the new process would deliver, and concluded with a demonstration of the CBR software supporting the process. These sessions were essential in obtaining the support of the whole of the company to the project.