DEVELOPMENT OF KNOWLEDGE-BASED EXPERT SCREENING SYSTEM FOR CERVICAL CANCER

LIM JIN CHOW

UNIVERSITI SAINS MALAYSIA

2016

DEVELOPMENT OF KNOWLEDGE-BASED EXPERT SCREENING SYSTEM FOR CERVICAL CANCER

by

LIM JIN CHOW

Thesis submitted in fulfillment of the requirements
for the degree of
Master of Science

September 2016

ACKNOWLEDGEMENT

Firstly, I would like to express my deepest gratitude and appreciation to my supervisor, Prof. Dr. Nor Ashidi bin Mat Isa; co-supervisors, Assoc. Prof. Dr. Umi Kalthum Ngah and Assoc. Prof. (Dr) Mohd Pazudin Ismail for their guidance and support which has been given to me throughout the completion of this project. They have fully dedicated their time in guiding me for my Masters project and also provide numerous suggestions that help me to overcome my project difficulties throughout my research work. Besides, they have provided useful and helpful information as a reference for my project as well. In addition, their effort to assist the application of ethical approval of Hospital Universiti Sains Malaysia (HUSM) for data collection is greatly appreciated and will always be something for me to remember.

My special thanks reach out to Hospital Universiti Sains Malaysia (HUSM) situated at Kubang Kerian, Kelantan and the board of ethical of HUSM for approving the ethical approval of this research so that the data collection of this project has been carried out successfully.

Last but not least, i would like to acknowledge School of Electrical and Electronic Engineering Universiti Sains Malaysia (USM) and Institute of Graduate Studies (IPS) for providing necessary supports and courses to aid the research students. Likewise, i also appreciate and thank the colleagues that have helped me on this project.

TABLES OF CONTENTS

CHAPTER FIVE: CONCLUSION AND FUTURE WORK

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AWFQ	Adaptive Weighted Fair Queuing
ASR	Age Standardize Rate
AAMBE	Average Absolute Mean Brightness Error
AC	Average Contrast
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under Curve
ASC	Atypical Squamous Cells
BMI	Body Mass Index
CIN	Cervical Intraepithelial Neoplasm
CIS	Carcinoma In Situ
CBR	Case-Based Reasoning
DNA	Deoxyribonucleic Acid
ESCUM	Expert System for Cervix Uteri Morphometry
ES	Expert System
FL	Fuzzy Logic
FES	Fuzzy Expert System
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
HUSM	Hospital Universiti Sains Malaysia
HGSIL	High Grade Squamous Intraepithelial Lesion
HPV	Human Papilloma Virus
HMLP	Hybrid Multilayered Perceptron

HVAC	Heating Ventilation and Air Conditioning
HLVK	High Level Vision Kernel
HUSM	Hospital Universiti Sains Malaysia
HIV	Human Immunodeficiency Virus
IH	Intrinsic Hypothesis
ICM	Intelligent Computing Method
LGSIL	Low Grade Squamous Intraepithelial Lesion
MOH	Ministry of Health Malaysia
NC	Nucleo-Cytoplasmic
KBS	Knowledge-Based Expert System
OR	Odd Ratio
PT	Priority Table
PSA	Prostate Specific Antigen
PCCM	Pittsburgh Cervical Cancer Screening Model
ROC	Receiver Operating Characteristic
RBR	Rule-Based Reasoning
RGB	Red Green Blue
SS	Scheme Scheduler
SES	Socio Economic Status
SMILE	Structural Modeling, Inference, and Learning Engine
TCP	Transmission Control Protocol
TP	True Positive
TN	True Negative
UDP	User Datagram Protocol
WHO	World Health Organization
WFQ	Weighted Fair Queuing

PEMBANGUNAN SISTEM SARINGAN PAKAR BERASASKAN PENGETAHUAN UNTUK BARAH PANGKAL RAHIM

ABSTRAK

Barah pangkal rahim adalah barah kedua serius dalam kalangan wanita. Barah pangkal rahim boleh dirawat pada peringkat awal. Oleh itu, pemeriksaan dan diagnosis memainkan peranan yang penting bagi pengesanan awal barah tersebut. Tugas pemeriksaan pesakit barah pangkal rahim memerlukan pengetahuan dan pengalaman pakar sakit puan. Biasanya terdapat beberapa masalah yang wujud dalam saringan manual palitan pap di mana kekurangan kepakaran ahli patologi dan pakar ginekologi serta proses kerja yang memakan masa yang lama dan keputusan tidak boleh diperolehi dalam tempoh yang singkat. Kadang-kadang manusia mempunyai kesukaran dan ketidakpastian dalam membuat keputusan ketika pemeriksaan. Oleh itu, penyelidik-penyelidik terdahulu telah membangunkan sistem pakar dalam saringan barah pangkal rahim dengan menggunakan penganalisis imej. Walau bagaimanapun, analisis imej memerlukan masa yang lama di mana spesimen imej diperlukan sebelum analisis imej boleh dilakukan. Untuk menangani masalah ini, pembangunan sistem pakar berasaskan pengetahuan dalam pemeriksaan pra-barah pangkal rahim diperkenalkan dalam kajian ini di mana sistem tersebut hanya terdiri daripada penggunaan soal selidik, tanpa penglibatan analisis imej dalam sistem ini. Selain itu, hasil klasifikasi boleh diperolehi dengan cepat daripada sistem ini. Dari hasil kajian ini, sistem yang dibina menunjukkan keupayaan yang baik dalam pengasingan kes-kes pemeriksaan pangkal rahim yang normal dan tidak normal, di mana sensitiviti dan spesifisiti yang diperolehi masing-masing adalah 95% dan 85%.

DEVELOPMENT OF KNOWLEDGE-BASED EXPERT SCREENING SYSTEM FOR CERVICAL CANCER

ABSTRACT

Cervical cancer is the second common cancer among women. Cancer can be cured at early stage. Hence screening and diagnosis plays an important role for early detection of cancer. The screening task of cervical cancer patient needs knowledge and experience of a gynecologist. Typically there are several problems exist for pap smear screening which is the shortage of expertise such as pathologists and gynecologists, and it involves plenty of manual work for the cervical cancer screening, hence the work process becomes time consuming and the results cannot be obtained in a short period. Sometimes human has difficulty and uncertainty in making decision for the screening results. Researchers have already developed an expert system in screening cervical cancer by using image analyzer. However, image analysis is time consuming time where the image specimen is needed before further image analysis can be done. In order to address the above mentioned problems, a knowledge-based expert system for pre-cervical cancer screening is introduced in this study which only consists the usage of questionnaire and without the involvement of image analysis in this system. Besides that, the classification result can be obtained spontaneously from this system. From the results of this study, the system shows high capability of segregating the abnormal cervical screening cases and normal screening cases, with sensitivity and specificity of 95% and 85% respectively.

CHAPTER ONE

INTRODUCTION

1.1 Background

Cervical cancer is a type of cancer that develops in the cervix, which is located at the lower part of the uterus (Cervical Cancer, 2015). Initially, it develops as a thin layer of epithelium cells in the cervix. Compared to other cancers, cervical cancer is the most preventable. This is because cervical cancer develops gradually with pre-cancer symptoms that can be detected at an earlier stage. If pap smear screening is done at regular intervals, cervical cancer symptoms can be detected earlier. Thus, this prevents the disease from developing into invasive cancerous cells and therefore allows only the infected part to be removed with ease (Cervical Cancer, 2015). The cervical pre-cancer stage is divided into three stages, which is known as the Cervical Intraepithelial Neoplasm one, that is (CIN I), CIN II and CIN III. CIN III is the invasive cancerous cell, which is also known as carcinoma in situ (CIS) and has the potential in developing into cancerous cell (Cervical Cancer, 2015).

In Malaysia, cervical cancer is ranked the second most top killer among women compared to other cancerous disease (MOH, 1999; Othman (2003)). Cervical cancer is caused by the human papilloma virus (HPV) (Bosch et al., 1997). HPV is a type of disease transmitted sexually. The risk factors associated with HPV infections are sexual intercourse, having multiple sexual partners and a poor socioeconomic status (Cuzick, 1999; Domingo et al., 2008; Pitts and Clarke, 2002). Women are encouraged to go for pap smear screening test at regular intervals. This is a

preventive method, which can reduce the mortality rate associated with cervical cancer (Cuzick, 1999).

1.2 Artificial Intelligence

In artificial intelligence, the machine has the ability to imitate human intelligence. Artificial intelligence (AI) consists of several different types of intelligent systems such as neural network, genetic algorithm, data mining and expert system. Artificial neural network (ANN) has the ability to imitate the human mind, which consists of trillions of neurons such as soma, dendrites and axon. Therefore, in ANN, the neurons are typically interconnected to each other by weighted links, imitating a biological neural network. There are three layers in ANN, which are the input layer, the hidden layer, and the output layer. Each neuron will receive the input signals from its weighted connections and then the output signals are transmitted through the outgoing connection. The neuron connections have numeric weights, which can be fine tuned and trained according to the condition of the environment (Negnevitsky, 2005).

Another AI system is the genetic algorithm (GA). GA is an artificial intelligent system that uses the stochastic search method. A superior chromosome will only appear when it is found that the population remains stable for a certain number of generations. The GA will only terminate itself when it is found that the generations have met a specified number. The best chromosomes in the population will be identified and if the solutions did not meet the criteria, the GA will be repeated itself (Negnevitsky, 2005).

Another AI system is the expert system (ES). ES is typically a computer software program that behaves like a human expert in performing a specific task. There are several types of expert systems in artificial intelligence such as the knowledge-based expert system, the rule-base expert system and the fuzzy expert system. The expert system has several key components which are the knowledge-based, data-base, inference engine, and user interface (Negnevitsky, 2005). The knowledge-based of the expert system contains the domain knowledge of the human expert. The domain knowledge is represented in the form of structured rules that is IF-THEN rules, whereby IF represents the condition set and THEN represents the action that is to be taken. The rule will be fired and the action will be triggered according to the condition set by the rule (Negnevitsky, 2005). The data in the database of the expert system will be used to match the IF condition part of the rules. The inference engine in the expert system provides the reasoning between the rules and data to find a solution (Negnevitsky, 2005). A user interface represents a communication link between a human and an expert system to exchange information in order to find a solution to a problem.

1.3 Problem Statements

Cervical cancer screening is an important screening method available that can detect HPV in its early stages before it develops into an invasive cancer cell and it is also known to help reduce the mortality rate in cervical cancer cases (Cuzick, 1999). However, manual pap smear screening test that is practiced in the medical field today, has some setbacks. Pap smear screening involves the acquiring of cervical smears manually from the cervix area and then smeared on the test slide (Cuzick, 1999). The smear will be examined manually under a microscope for abnormalities. The process

of manual screening is not only time consuming but involves plenty of manual work. On the other hand, the expert system has the capability to do fast screening without acquiring cervical scraps and a medical officer can receive some reference results, which can act as a reference before further detailed screening is to be done.

Secondly, in a certain condition a person can find some difficulty in making decisions even after diagnosing the results from the screening test. Therefore, in these types of cases the expert system can assist in giving a second opinion to medical officers for inference.

Furthermore, there is currently no expert system that is totally dependent on social and medical questionnaires only. This expert system also does not perform any image processing. Studies have shown that there is an expert system available, which uses knowledge-based image analyzer in detecting cervical cancer cells (Chan et al., 1996). This involves the image processing of cervical cancer cells and the minimum information required for input is a patient's historical detail, which is used in cervical cancer classification. However, if image processing is required in the study, the image in the knowledge-based system has to be transferred to the expert system. The screening process will take much longer where image processing is concerned. Besides this, there are only a limited number of medical experts who have the expertise to diagnose and analyze the cervical cancer cell image.

From the review of artificial intelligence that involved in medical screening, there is currently no expert system for cervical cancer screening which involves only questionnaires, without involvement of image processing. There are several studies that have the nearest similarity with this research which are the research of an expert system for the detection of cervical cancer cells using knowledge-based image

analyzer (Chan et al., 1996) and research from Ho et al. (2004) which is the data mining with induction technique implemented in the cervical cancer screening (Ho et al., 2004). Expert system of Chan et al. (1996) needs image processing of cervical scraps from the pre-work of attaining the patient's cervical scraps. Data mining system from Ho et al. (2004) also needs the pre-work of obtaining the genome information of a patient from the medical expertise for system analysis. Hence both studies need the pre-work to be done before further tasks can be performed by the system. These factors are the differences comparing to current research where the current expert system does not need any pre-work to be done for the system to perform cervical cancer screening and it purely involves questionnaires only for cervical cancer screening.

1.4 Objectives

Based on the above mentioned drawbacks, the objectives of this project are as follows:

1. To determine the need of social and medical questionnaires of the cervical cancer cases for the aspect of optimizing the screening effect.
2. To build an artificial intelligent expert system for cervical cancer screening.
3. To build an overall combined social and medical questionnaires knowledge-based expert screening system for cervical cancer.

The scope of this project focuses mainly on the screening aspects of the social and medical historical background of the user. Besides this, the project also tests the system's compatibility of the combined social and medical questionnaires that are risk factors associated with cervical cancer. Secondly, the scope of this project also

covers the development of the expert system in cervical pre-cancer screening by classifying the output as normal or as an abnormal outcome. The C++ programming language is used to develop the proposed algorithms.

1.5 Thesis Outline

This thesis comprises of five chapters that is Chapter 1, introduction; Chapter 2, literature review; Chapter 3, methodology; Chapter 4, result and discussion and Chapter 5, conclusion and future development work.

In Chapter 2, some relevant theories on the background of cervical cancer are discussed. The risk factors associated with cervical cancer are explained in detail. The cervical pre-cancer screening procedure is described in this chapter for its clarity when performing pap smear screening tests. Subsequently, the uses of artificial intelligence systems in diagnosing cervical pre-cancer disease are discussed followed by the introduction of the knowledge-based expert system. Finally the application of expert systems in other relevant fields other than medical field will be presented in Chapter 2.

The content in Chapter 3 shows the methodology of the development of the framework in the knowledge-based expert system. Details regarding the development of system content, system content flow, weight allocation and expert's knowledge acquisition will also be discussed explicitly in this chapter. Other than that, the data collection and weight tuning process will also be presented in Chapter 3.

This is followed by Chapter 4, which discusses the statistical analysis of the screening capabilities of the expert system. All results from the receiver-operating

characteristic curve (ROC), area under the curve (AUC), sensitivity and specificity will be discussed, explained, analyzed and compared in order to evaluate the performance and screening capability of the system.

Chapter 5 is the final chapter in this thesis where the work of the entire project is summarized and concluded. In this chapter, the study also discusses the future development work involved.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Cancer greatly impacts our society today and has become one of the major health issues in Malaysia. The Ministry of Health Malaysia shows that in Malaysia alone, cancer cases are ranked the third highest cause of deaths as compared to other medically certified deaths in the year 1999. Cervical cancer has been reported to be the third most popular disease, in comparison to other types of cancer and also has been identified as one of the second top ranked killer cancer among women (MOH, 1999; Othman, 2003)

According to the Planning Division Health Informatics Centre, Ministry of Health Malaysia, which conducted a cancer survey, neoplasm, a type of cancer has the fourth highest mortality rate compared to other medically certified deaths in the year 2014 in Malaysia which is showed in Figure 2.1 (MOH, 2014).

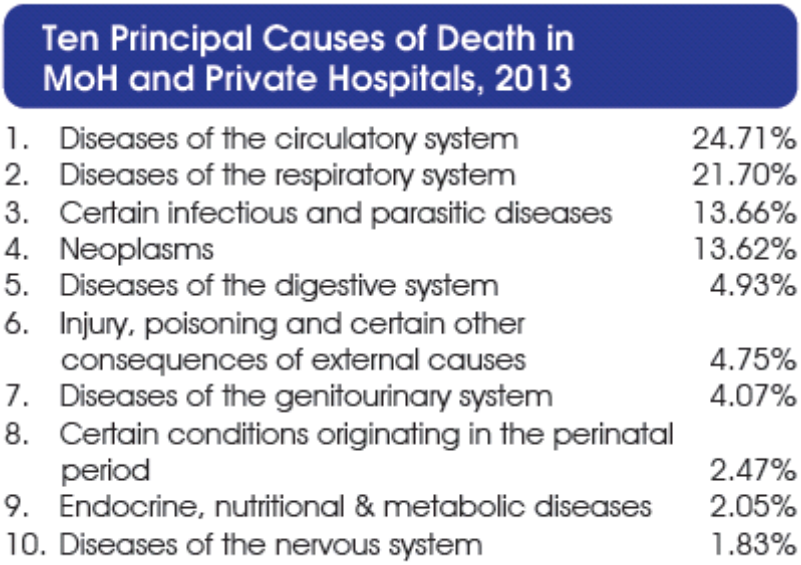

Figure 2.1: Health Facts 2014, Ministry of Health Malaysia. (MOH, 2014).

2.2 Cervical Cancer

2.2.1 Background of Cervical Cancer

Cancer can simply be defined as normal cells that have been transformed into abnormal cells. Cancer cells can be characterized as being invasive. Cervical cancer basically occurs at the location of the cervix, which is the lower portion of the uterus. Initially, it develops into a thin layer of cells that covers the cervix known as epithelium. Epithelium cells are of several types. The first type is called squamous cells and, the shape looks flat and scaly. Initially, cancer cells are most likely to begin their transformation process from the squamous cells stages. This phenomenon is termed as squamous cell carcinoma. The second type is called columnar cells. The shape looks like a column or a glandular. If cancer is found in the columnar cells, it is termed as adenocarcinomas. The third type of condition is a mixed phenomena

between squamous cells and adenocarcinomas. It is in this case that, these types of cells are termed as mixed carcinomas.

Cancer has two main stages, which are pre-cancer stage and cancerous stage. In the pre-cancer stage, the pre-cancer condition is termed as dysplasia. Dysplasia means that the squamous epithelial cell that covers the cervix begins to grow. After a pap smear is done, the results are divided into three types, atypical squamous cells (ASC), low grade squamous intraepithelial lesion (LGSIL), and high grade squamous intraepithelial lesion (HGSIL) (Cervical Cancer, 2015).

The outcome from a pap smear test is called colposcopy. The result of colposcopy is again categorized into three other sections called cervical intraepithelial CIN I, CIN II, and CIN III. CIN III has a high probability to develop into an invasive cancer, which is known as carcinoma in situ (Cervical Cancer, 2015).

In Malaysia, cancer of the cervix is the second most common cancer among women after breast cancer. The population of women in Malaysia in the year 2000 was approximately 10.5 million (WHO, 2003), and the total population grew to 14.9 million in the year of 2013 (Malaysia Population 2015, 2015). According to WHO, approximately 30% of these women are in the reproductive period or older and they are at risk of developing cervical cancer (WHO, 2003). In fact, Malaysia has more cases of cervical cancer when compared to other developed countries (Mokhtar, 2007). In peninsular Malaysia alone, the crude incidence rate for cervical cancer in women is 100,000. Out of this 100,000, the crude incidence rate for women is 13.4% who are between 15 and 49 years of age, and 62.9% who are between the later ages of 50 and 69 years old, (Lim and Yahya, 2003). The high risk group who are affected include early sexual exposure, having multiple sexual partners, promiscuity, HPV

infections, sexually transmitted disease, smoking and who are not economically sound (Cuzick, 1999; Domingo et al., 2008; Pitts and Clarke, 2002).

2.2.2 Risk Factor of Cervical Cancer

Human papillomaviruses (HPV) is a type of sexually transmitted disease that can cause cervical cancer and the development of cervical cancer is highly associated to certain risk factors. Medical research studies have shown that these risk factors can lead to the cervical cancer.

The first risk factor discussed is the age dependency associated with cervical cancer. According to Bosch et al. (1997), in developed countries, the younger age group has a higher risk of being affected by HPV when compared to other age groups. Furthermore, the data shows that the HPV rate has dropped to 12%, between the ages of 21 to 25 years old, and to 4%, between the ages of 51 to 55 years old (Bosch et al., 1997). However, an increase in HPV has become prevalent in a certain section of the population, which is the elderly group (Bosch et al., 1997; Meijer et al., 1991; Schneider and Koutsky, 1991). According to Patel et al. (2003), the age group of those who are 55 years and more, are reported to be within a growing population of HIV cases due to unsafe sexual intercourse, which could indirectly increase the number of HPV cases (Patel et al., 2003). To elaborate further the point, the elderly age group, has a higher risk in developing cervical cancer, Fox et al. (2008) states that elderly women who are from the age of 60 and above are significantly higher in cervical cancer cases (Fox et al., 2008).

According to Rotkin (1967), from the distribution table showing the menopause of cervical cancer patients, the mean age for both patients and controls is

directed towards the ages of 41.7 and 44.5 years old (Rotkin, 1967). This shows that women from the age of 41.7 also have a high possibility of being infected by cervical cancer. Further, according to Domingo et al. (2008), in Thailand, there is a study which concludes that HPV infection have been tested positive in the age group of women who are less than 25 years old (Domingo et al., 2008). Women in an early age group also are at risk of being infected by HPV. As a prevention plan for cervical cancer, in Malaysia, the government has introduced the pap smear screening programme for women within the age group of 20 to 65 years old (Domingo et al., 2008).

Ethnicity and race is another risk factor that is associated to cervical cancer. According to Domingo et al. (2008), in Malaysia, Chinese women have the highest age standardize rate (ASR) of cervical cancer cases, which shows a ratio of 28.8 per 100,000. The second highest is the Indian community, which post an ASR of 22.4 per 100,000, and Malay community with an ASR of 10.5 per 100,000 (Domingo et al., 2008). Statistic from the Second Report of National Cancer Registry showed that, ASR for Chinese women is 59.7 per 100,000, followed by Indian women with an ASR of 55.8 per 100,000, and Malay women being the lowest with an ASR of 33.9 per 100,000 (Lim and Yahya, 2003). Thus, historical statistical data shows that ethnicity and race are important factors in determining HPV cancer cases.

Menarche is another risk factor that is associated with cervical cancer. Menarche happens to women between the ages of 8 and 12 years. Data collection in a few western countries such as the United States showed that the mean age of menarche was reported in the age group of between 12.60 ± 2.0 years old in the year 1966, whereas in Great Britain, the mean age of menarche was reported in the age