DESIGN AND DEVELOPMENT OF BREAST CANCER DIAGNOSIS SYSTEM USING MACHINE LEARNING

MOHAMMAD NURHAQIM BIN KAMARUDDIN

B071710835

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2020

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Tajuk: DESIGN AND DEVELOPMENT OF MACHINE LEARNING IN PROGNOSIS OF BREAST CANCER

Sesi Pengajian: 2020

Saya MOHAMMAD NURHAQIM BIN KAMARUDDIN mengaku membenarkan Laporan PSM ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka (UTeM) dengan syarat-syarat kegunaan seperti berikut:

Laporan PSM adalah hak milik Universiti Teknikal Malaysia Melaka dan penulis.

Perpustakaan Universiti Teknikal Malaysia Melaka dibenarkan membuat salinan untuk tujuan pengajian sahaja dengan izin penulis.

Perpustakaan dibenarkan membuat salinan laporan PSM ini sebagai bahan pertukaran antara institusi pengajian tinggi.

**Sila tandakan (X)

X	Mengandungi maklumat yang berdarjah keselamatan atau

	Mengandungi maklumat TERHAD yang telah ditentukan oleh TERHAD organisasi/badan di mana penyelidikan dijalankan.
X	TIDAK TERHAD

Yang benar,

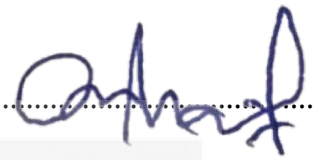
.....

MOHAMMAD NURHAQIM BIN KAMARUDDIN

Alamat Tetap : N0.14, JALAN RUMBIA 46, TAMAN DAYA 81100 JOHOR BAHRU JOHOR

Tarikh :12/6/2020

Disahkan oleh penyelia:



.....

EN NOOR MOHD ARIFF BIN BRAHIN

Cop Rasmi Penyelia:

NOOR MOHD ARIFF BIN BRAHIN
Jurutera Pengajar
Jabatan Teknologi Kejuruteraan Elektronik & Komputer
Fakulti Teknologi Kejuruteraan Elektrik & Elektronik
Universiti Teknikal Malaysia Melaka

Tarikh: 12/6/202

DECLARATION

I hereby, declare that this report entitled “Design and Development of Cancer Diagnosis Using Machine Learning” is the result of my own research except as cited in references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

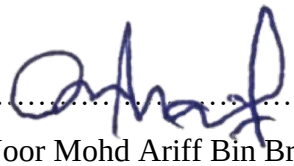
Signature: ……

Author:MOHAMMAD NURHAQIM BIN KAMARUDDIN

Date: 12/6/2020

APPROVAL

This report is submitted to the Faculty of Electrical and Electronic Engineering Technology of Universiti Teknikal Malaysia Melaka (UTeM) as a partial fulfilment of the requirements for the degree of Bachelor of Computer Engineering Technology (Computer System) with Honours. The member of the supervisory is as follow:

Signature: 

Supervisor: En Noor Mohd Ariff Bin Brahin

ABSTRACT

This research deals with the design and development of machine learning for diagnosis of cancer, which is then used for prognosis. It acts as an alternative to assist the pathologist in analyzing the cell physical characteristics under a microscope and determining whether the tissue removed is benign (non-cancerous) or malignant (cancerous) in the early detection. The initiative focuses primarily on early breast cancer screening, and aims to classify patients on the basis of tests. While the traditional method is good, early diagnosis of breast cancer can significantly improve the prognosis and the chance of survival, as it can help the clinical treatment of patients. The importance of cancer patients has inspired many biomedical and bioinformatics investigative teams to study the use of machine learning approaches. Variation of methods in this machine-learning, such as Decision-Tree Classifier, Logistic Regression, SVM, Gaussian-NB and Random Forest Classifier, which are widely used in cancer research in predictive models for efficient and accurate decision-making, can be crucial since we can choose the best model. Furthermore, it is important that machine learning algorithms are able to identify core features from very large datasets.

ABSTRAK

Projek ini adalah mengenai reka bentuk dan pengembangan pembelajaran mesin yang akan digunakan dalam diagnosis kanser. Ia bertindak sebagai alternatif untuk memberikan bantuan kepada ahli patologi yang bertugas menganalisis penampilan sel di bawah mikroskop dan menentukan sama ada tisu yang dikeluarkan adalah jinak (bukan barah) atau ganas (barah) untuk pengesanan awal. Projek ini terutama difokuskan pada pengesanan awal kanser payudara yang membantu memprioritaskan pesakit berdasarkan hasilnya. Walaupun kaedah tradisionalnya baik, diagnosis awal kanser payudara dapat meningkatkan prognosis dan peluang untuk hidup dengan ketara, kerana dapat memudahkan pengurusan klinikal pesakit berikutnya. Petunjuk untuk mengklasifikasikan pesakit barah berdasarkan keseriusan mereka telah mnarik perhatian banyak pasukan penyelidik dari bioperubatan dan bioinformatik untuk mengkaji penerapan kaedah pembelajaran mesin. Variasi kaedah yang digunakan dalam pembelajaran mesin ini seperti Decision-Tree Classifier, Logistic Regression, SVM, Gaussian-NB dan Random Forest Classifier yang telah banyak diterapkan dalam penyelidikan kanser untuk pengembangan model ramalan yang menghasilkan keputusan yang efektif dan tepat, dapat menjadi penting kerana kita boleh memilih yang terbaik untuk menjadi model. Lebih-lebih lagi, kemampuan alat pembelajaran mesin untuk mengesan ciri utama dari set data yang sangat kompleks menunjukkan kepentingannya.

DEDICATION

I would like to dedicate this project to my supervisor En Noor Mohd Ariff Bin Brahin who have guided me throughout this project process. Moreover, special thank you to my family Kamaruddin Bin Md Din, my father and Marsiah Binti Minhat, my mother who always give supports and always encourage me to finish this project. Without their prayers I could not able to finish this project in time and their unconditional love make me become more motivated also not easily give up. Also, not to forget I would like to dedicate this work to Universiti Teknikal Malaysia Melaka (UTeM) and Faculty of Electrical and Electronic Engineering Technology that gives me opportunity to complete my study here. Also, I would like to dedicate my brothers and friends who always surround me with positive vibes and always be beside me through thick and thin especially Mohammad Syafiq, Mohammad Firdaus, Annisya Aziz.Thank You.

ACKNOWLEDGEMENTS

Firstly, I might want to offer my thanks to Allah S.W.T that without His endorsement I probably won't total my four year college education venture.

Next, the most notable individual that assume job in helping me to finish my undertaking, I might want to offer my earnest thanks to my manager En Noor Mohd Ariff Bin Brahin. He calmly directs me and give me support until this task totally done. Without him these would not even conceivable to be done on schedule.

In conclusion yet not least, to all talks in FTKEE that share information to me. Additionally, an uncommon gratitude to my relatives for giving me all the help and invigorate me to prop up so as to complete this task. I might want to thank to any individual who in a roundabout way and legitimately include to this task, no one but Allah can reimburse all your benevolence and May Allah give us his approval.

TABLE OF CONTENTS

CHAPTER 5 : CONCLUSION AND RECOMMENDATION

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.0 Introduction

This chapter explains the overview of the study and the main purpose of the project. The chapter includes background of the study, problem statement of the project that wanted to build, objectives that goal at the end of the project and the scope of the study that will be conducted.

1.1 Background Study

One of the most common cancers in women and the second most common cancer in general has been breast cancer for decades. Early detection is vital when it comes to cancer disease. The tumour size and how far it has spread are the main factors that decide a patient's outcome. Furthermore, scientists applied variations of methods such as early-stage screening methods, which they can contribute to detection of cancer before they cause symptoms. They have grown new methodologies for the early forecast of malignancy treatment result that . With the coming of new advances in the field of medication, measures of malignant growth information have been accumulated and are accessible to the exploration network who need to look for an answer. In any case, the exactness of forecast of an infection result is one of generally interesting and trying for doctors who need to locate the most ideal apparatus for clinical exploration purposes. Thus, Machine Learning strategies have become a mainstream instrument for clinical analysts. The utilization of Artificial Intelligence (AI) and information mining strategies has changed the entire procedure of bosom malignant growth Diagnosis and Prognosis (Gupta, Kumar and Sharma, 2011) .

Generally, a finding of cancerous tumour was normally will be controlled and examined by pathologists. A pathology report is a clinical record composed by pathologist which will analyze infection by dissect research facility tests and survey cells, tissues, and organs. These reports will give a conclusion dependent on the pathologist's assessment of an example tissue taken from the patients tumour or cell cores. By at that point, the pathologists can make sense of whether the tissue is noncancerous or malignant. A cancerous tumour can

be characterize into two classes which is malignant and benign. A dangerous malignant growth tumour is a disease cell that can develop and spread to different pieces of the body while considerate malignancy tumour implies the tumor can develop however does not spread. The datasets tests that been utilized in this task is from UCI Machine Learning Repository which separated by University of Wisconsin. Highlights of the cell cores are processed from a digitized picture of a fine needle suction (FNA) of a tumour mass. They depict attributes of the cell cores present in the picture. Fine needle goal (FNA) is a sort of biopsy methodology. In fine needle goal, a meager needle is embedded into a region of irregular showing up tissue or body fluid. As with different kinds of biopsies, the example gathered during fine needle yearning can help make a conclusion or preclude conditions, for example, cancer. Fine needle desire is commonly viewed as a sheltered system (Mendoza et al., 2011).

This project centers around how we can break down the datasets given and execute the prepared model for new information from new patients. All the expense and time can be diminished by utilizing this AI which plan to help the pathologists to decide an exceptionally precise determination for a bosom disease understanding. Their constraint can be settled by utilizing the guideline of computer vision framework which will improves the clinical methodology particularly in early location of bosom disease. So as to anticipate the malignant growth analysis, Machine Learning utilizing certain calculations to test and check the noteworthy highlights as it will process and break down the most ideal conclusion of the disease tumour. The forecast procedure is utilizing AI with the help of Scikit-learn library and different calculations. AI is supplanting a traditional strategy in characterizing and identifying malignancy tumour.

1.2 Problem Statement

For the past few years, breast cancer became one of steadily increase disease that need attention from all people especially scientists and medical researchers. One of the reasons of increase in percentage of people who got affected is lack of approach in early detection. In Figure 2, a bar chart shown by The Global Cancer Observatory was indicator for the importance of early detection as people who was diagnosed with stage 1 of breast cancer will have a highest chance to survive. In addition, a research of breast cancer in Malaysia by The Global Cancer Observatory (Figure 1), shows that breast cancer in Malaysia in year of 2018, have the second most highest number of new cases compare to other type of cancer which need immediate intention and solution.

Malaysian women have poor survival from breast cancer and it is estimated that half of the deaths due to breast cancer could be prevented (Yip, Pathy and Teo, 2014). In the past year, the technology used to detect are rely only on expertise of pathologist who will diagnose the patient with little amount of assistance from technologies sides. Today, a lot of breakthrough technologies aid in early detection of cancer such as oncology imaging and diagnosis workflow but none of the technologies are really affordable and involve a lot of hardware. It also need a lot of staff to handle and learn it can be quite complex. In addition, the trial proof accessible for most new bosom malignant growth recognition advancements was not sufficiently able to help authoritative decisions about their definitive clinical worth and use.

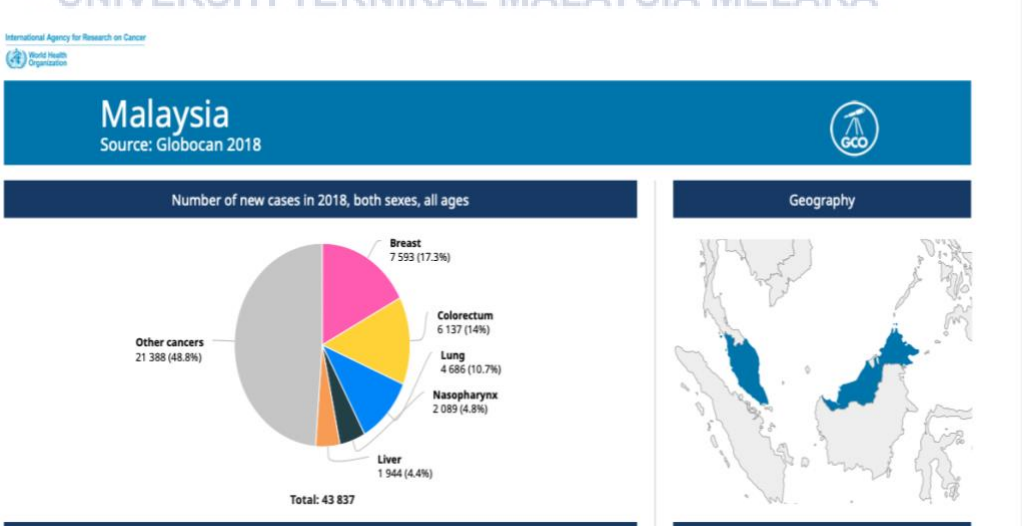


Figure 1.1 : Number of New cases 2018

In this figure, we can see that breast cancer represents second highest number of new cases in 2018 among cancer disease. These indicates that early detection can be importance instrument to combat the rising of new cases in Malaysia.

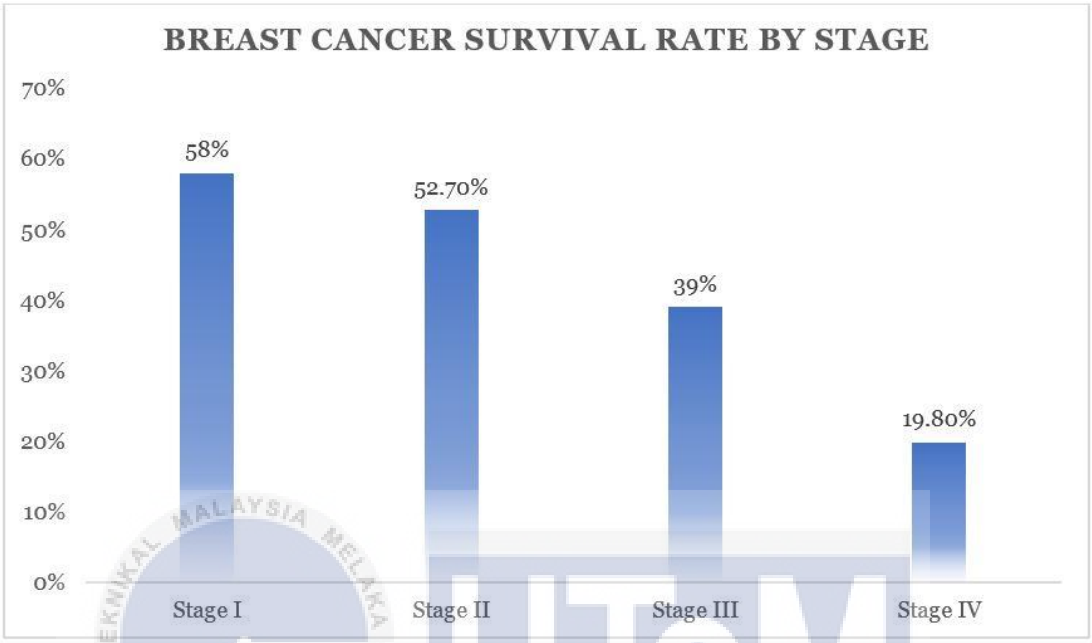


Figure 1.2 : Breast Cancer Survival Rate By Stage

Most of this advent technologies comes with several of issues in terms of cost, flexibility, complexity, reliability and functionality. The invention of fast, reliable, affordable, flexible, non-complex would solve the issues that always encounter a technology when it comes to functionality and practicality. Majority of technologies these days whether software or hardware lack of practicality in order to perform its functionality. Maybe as many as 70% of all medical decisions are based on laboratory tests. Also, the tremendous amount of information in the electronic clinical record is from the clinical research facility. Test results from pathology and the clinical research facility every now and again fill in as the best quality level for clinical results contemplates, clinical preliminaries, and quality improvement. This monstrous measure of information requires colossal limit with regards to capacity and refined strategies for dealing with and recovery of data, requiring the use of specific information science teaches, for example, Machine Learning (Rashidi et al., 2019).

The areas of anatomical and clinical pathology have brought substantial contributions to enhance patient safety, including the establishment of standardised and informal standards for the identification and monitoring of quality lapses. Public protection advocates promote immediate and total disclosure of all significant medical mistakes in order to ensure a patient-physician relationship and minimise the likelihood of injury to patients. Unfortunately, certain medical accidents are unavoidable. Garland's study on the accuracy of radiological and other diagnostic procedures showed a "surprising degree of inaccuracy" including "34% error margin in the diagnosis of myocardial infarction," "agreement rate of only 7% among five experienced paediatricians" in the diagnosis of malnutrition in infants, and "a 28% error rate among clinical laboratories in the calculation of erythcocyte count" (Schwartz *et al.*, 2007). The most common way to spot a mistake is indirect inspection with another pathologist which they also examining at a slide and the problems is that there are no standardized method as other pathologist does not always have same way. Using machine learning model system, it helps the doctor or pathologist to minimize the error able to produce a consistent diagnosis result without affected by any external factor which can harms the patient.

Moreover, this project which involves a machine learning model can help the doctor to make a cross-check process easily without needed to gather a large amount of other pathologists to do this process. In conventional way, after observed the physical charateristic of the tumour cell on microscope, the pathologists will do cross-check process with each other to make a final result of diagnosis. This process can be shorten with the existence of our system which help them to make a better diagnosis. As stated by (Annuar, 2018), Health Ministry of Malaysia said that they were 350 pathologists serving at 145 healthcare facilities run by ministry nationwide and it cannot fulfill the needs of demand that consist of 70% of all disease which relied on pathology lab tests. Pathologists are highly trained specialists, many with years of experience, surrounded by qualified technicians trying to get the right diagnosis. And how are accidents going to happen? Industry-derived error detection systems allow for an analysis of all phases of the specimen management process and help to determine where errors exist. Such scrutiny detected pre-analytical (poor specimen marking, delayed delivery to the laboratory, poor clinical information), analytical (incorrectly applied experiments, errors in interpretation) and post-analytical (incorrect description of the patient on the sample, errors in analysis of the findings of the report) errors.

1.3 Objectives

The main objective for this project is to design and develop machine learning for breast cancer diagnosis based on data classification. This project carried out on the following objectives:

1. To study sklearn machine learning and deep learning platform for classification and regresssion

2. To plan and build up the framework to anticipate finding of breast cancer prognisis by utilizing Artificial Intelligence.

3. To evaluate the performance and efficiency of each algorithm used for breast cancer diagnosing using supervised methods

1.4 Scope of work

This project will explore how the diagnosis of breast cancer can be predicted using a trained machine learning model based on samples from the University of Wisconsin. The goal of this prediction system is to act as assistance for pathologists by using different supervised algorithms. The key features of this project are the prediction of the diagnosis of breast cancer on the basis of data classification and regression. There are several limitations that come across during these projects. Firstly, there are not enough data to actually ensure that this machine learning model can be trained with zero percent loss of accuracy. The higher amount data that been put into this model to train, the more accuracy can be achieved which will produce of top quality output without any doubts. Next, this project is based on the data of international patients which conduct by University of Wisconsin. When it comes to breast cancer, there are other external effects such as family health history, genetic and lifestyle. If we have datasets sample from Malaysian patients, it will be more relatable and can increase this project hypothesis. This project is to focus on the ability of the algorithm of the model that been chosen to predict the diagnosis of breast cancer of each patients.

1.5 Organization

This chapter addresses the context of the project, the declaration of issues and the complexity of the project. On the other hand, the current method implemented in the past-related project is examined in chapter two by analysing the authorized article, journal and website. The advantages and disadvantages of the project will be compared and presented in the table. Furthermore, the flow of the project and the method that been use as to complete this project will be discussed in chapter three. After that, the project analysis will be analysed in chapter four. The project conclusion and future recommendation will be discussed in chapter five.

1.6 Summary

This chapter also covered the implementation of this initiative. It starts with the project background, where the background system was explained. The problem statement that the system and environment normally occurs is briefing. The aims of this project and the research scope also included in this chapter.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

A literature review is an important part before starting any project. This chapter will be concentrated on past and current project works. Many of the application and related reviews have been studied and all the analysis was identified with the point. This chapter will show the summary for each of the research that has done especially the system operation that been implemented in previous project.

2.1 Past Related Research

This chapter will be discussed more in the previous and current research to build up this project. The source of these inquiries must be satisfactory in the system format which comes from articles, journals, books and websites that are authorized.

2.2 Machine Learning Analysis of Inflammatory Bowel Disease

(H. Hacılar, O. U. Nalbantoğlu and B. Bakir-Güngör, 2018) stated that the use of state-of-the art machine learning approaches became popular to address a variety of questions like early diagnosis of certain diseases using human microbiota. In this regard, we can classify a machine learning is an instrument to achieve a breakthrough in healthcare, especially in early detection. In this project, it used datasets of human gut metagenomic tests from people with inflammatory bowel diseases (IBD) as the principle object is to endeavor whether it can separate IBD patients from sound individuals utilizing the degrees of 3302 microorganisms as highlights. Crohn's disease and ulcerative colitis are so akin that they often are mistaken for each other. Making an accurate diagnosis is critical so that an individual can receive the most effective treatment for his or her disease.