THE APPLICATION OF ARTIFICIAL INTELLIGENT TECHNIQUES IN ORAL CANCER PROGNOSIS BASED ON CLINICOPATHOLOGIC AND GENOMIC MARKERS

CHANG SIOW WEE

FACULTY OF COMPUTER SCIENCE &
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR
2013

THE APPLICATION OF ARTIFICIAL INTELLIGENT TECHNIQUES IN ORAL CANCER PROGNOSIS BASED ON CLINICOPATHOLOGIC AND GENOMIC MARKERS

CHANG SIOW WEE

THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE &
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR
2013

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **CHANG SIOW WEE** (I.C/Passport No: **770320-01-6038**)

Registration/Matric No: **WHA070024**

Name of Degree: **Doctor of Philosophy (PhD)**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

THE APPLICATION OF ARTIFICIAL INTELLIGENT TECHNIQUES IN ORAL CANCER PROGNOSIS BASED ON CLINICOPATHOLOGIC AND GENOMIC MARKERS

Field of Study: **ARTIFICIAL INTELLIGENCE**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

ABSTRACT

Artificial intelligent (AI) techniques are becoming useful as an alternative approach to conventional medical diagnosis or prognosis. AI techniques are good for handling noisy and incomplete data, and significant results can be attained despite small sample size. Various AI techniques have been applied in medical research such as artificial neural networks, fuzzy logic, genetic algorithm and other hybrid methods. AI techniques have been proved to generate more accurate predictions than statistical methods and the predictions are based on the individual patient's conditions as opposed to the statistical methods which made predictions based on a cohort of patients.

Traditionally, clinicians make prognostic decisions based on clinicopathologic markers. However, it is not easy for the most skilful clinician to come out with an accurate prognosis by using these markers alone. In order to make a more accurate prognosis, one needs to include both clinicopathologic markers and genomic markers. Currently, there are very few published articles on researches that combine both clinicopathologic and genomic data. Thus, there is a need to use both of the clinicopathologic and genomic markers to improve the accuracy of cancer prognosis.

In addition, the mortality rate for oral cancer is high (at approximately 50%) and almost two-thirds of oral cancer occurs in developing countries such as Asian countries, yet there are very few studies using AI techniques in the prognosis of oral cancer. Furthermore, there is no Malaysian study yet on the application of AI techniques in the prognosis of oral cancer. Therefore, there is a need to investigate how AI techniques can be used in the prognosis of oral cancer.

The main aim of this research is to apply AI techniques in the prognosis of oral cancer based on the parameters of correlation of clinicopathologic and genomic markers. To this end, a hybrid AI model, namely ReliefF-GA-ANFIS was proposed. The proposed model consists of two stages, where in the first stage, ReliefF-GA is used as feature selection method and in the second stage ANFIS with *k*-fold cross-validation is used as classifier. The proposed prognostic model was experimented on the oral cancer dataset with optimum feature subsets and validated against three other models which are artificial neural networks, support vector machine and logistic regression. The results for the proposed model of ReliefF-GA-ANFIS outperformed the other three models and the results revealed that the prognosis is superior with the presence of genomic markers.

This research provides an insight to apply AI techniques in oral cancer prognosis based on both clinicopathologic and genomic markers. It is hoped that this research is capable of setting a basis for embarking more Malaysians in medical informatics research, particularly in the field of genomic markers.

ABSTRAK

Teknik pembuatan pintar (AI) semakin berguna sebagai pendekatan alternatif untuk diagnosis atau prognosis perubatan konvensional. Teknik AI berguna dalam mengendalikan data yang bising dan tidak lengkap, dan keputusan yang signifikan boleh dicapai dengan saiz sampel yang kecil. Pelbagai teknik AI yang telah digunakan dalam penyelidikan perubatan adalah seperti *artificial neural network, fuzzy logic, genetic algorithm* dan kaedah hibrid yang lain. AI teknik telah terbukti dapat membuat ramalan yang lebih tepat daripada kaedah statistik dan ramalan AI adalah berdasarkan kepada keadaan-keadaan dalam pesakit individu manakala kaedah statistik pula membuat ramalan berdasarkan kepada kohort pesakit.

Secara tradisinya, pakar perubatan membuat keputusan prognosis berdasarkan kepada penanda klinikopathologik. Walau bagaimanapun, ia tidak mudah untuk pakar perubatan untuk mengeluarkan keputusan prognosis yang tepat dengan menggunakan penanda ini sahaja. Oleh yang demikian, terdapat keperluan untuk menggunakan penanda genomik untuk meningkatkan ketepatan prognosis. Dalam usaha untuk membuat ramalan yang lebih tepat, kedua-dua penanda klinikopathologik dan penanda genomik diperlukan. Kini, tidak banyak artikel yang memaparkan penyelidikan yang menggabungkan kedua-dua data klinikopathologik dan genomik. Oleh itu, terdapat keperluan untuk menggunakan kedua-dua penanda clinicopathologic dan genomik untuk meningkatkan ketepatan prognosis kanser.

Tambahan pula, kadar kematian bagi kanser mulut adalah tinggi (pada kira-kira 50%) dan hampir dua pertiga daripada kanser mulut berlaku di negara-negara membangun seperti negara-negara di Asia. Akan tetapi, hanya terdapat sedikit kajian yang menggunakan teknik AI dalam prognosis kanser mulut. Di samping itu, tidak ada kajian

di Malaysia yang mengaplikasikan teknik AI dalam prognosis kanser mulut. Oleh itu, terdapat keperluan untuk menyiasat bagaimana teknik AI boleh digunakan dalam prognosis kanser mulut.

Tujuan utama kajian ini adalah untuk menggunakan teknik-teknik AI dalam prognosis kanser mulut berdasarkan kepada parameter korelasi penanda klinikopathologik dan genomik. Untuk mencapai matlamat ini, model hibrid AI, iaitu ReliefF-GA-ANFIS telah dicadangkan. Model yang dicadangkan ini terdiri daripada dua peringkat, di mana peringkat pertama terdiri daripada ReliefF-GA yang digunakan sebagai kaedah pemilihan penanda-penanda utama (feature selection) dan peringkat kedua ANFIS dengan *k-fold cross-validation* digunakan sebagai pengelas (classifier). Model yang dicadangkan telah diaplikasikan ke atas dataset kanser mulut dengan subset ciri optimum dan dibandingkan dengan tiga model yang lain iaitu *artificial neural network, support vector machine* dan *logistic regression*. Keputusan yang didapati telah menunjukkan bahawa model yang dicadangkan iaitu ReliefF-GA-ANFIS adalah lebih baik berbanding dengan ketiga-tiga model yang lain dan keputusan juga menunjukkan bahawa prognosis yang lebih jitu boleh diperolehi dengan kehadiran penanda genomik.

Penyelidikan ini telah membuktikan potensi yang baik untuk menaplikasikan teknik AI dalam prognosis kanser mulut berdasarkan kepada penanda-penanda klinikopathologik dan genomik. Maka dengan ini, diharapkan bahawa kajian ini mampu menetapkan asas untuk menggalakan lebih ramai rakyat Malaysia terlibat dalam penyelidikan informatik perubatan, terutamanya dalam bidang petanda genomik.

ACKNOWLEDGMENTS

I wish to record my indebtedness and appreciation to everyone that has been so helpful and supportive in this research and brought it to success.

First and foremost, I would like to express my deepest appreciation and gratitude to my supervisors, Assoc. Prof. Datin Dr Sameem Abdul Kareem, Assoc. Prof. Dr Amir Feisal Merican Aljunid Merican and Prof Rosnah Binti Zain, for their invaluable guidance, assistance and criticism during the course of preparing this research.

Special thanks to Dr. Thomas George Kallarakkal and staff of Oral & Maxillofacial Surgery department, Oral Pathology Diagnostic Laboratory staff, OCRCC staff, Faculty of Dentistry, and staff of ENT department, Faculty of Medicine, University of Malaya for their help and contributions towards the accomplishment of this research project.

To my late father, Chang Sow Chiang, without your guidance, I am just nobody. Next, my highest gratitude to my mother and my siblings, thank you for your continuous encouragement and support. Not forgetting my friends whom I have seek advice and being there when I needed them, thank you!

Last but not least, I delicate my work to my loving husband, my son and my daughter, who are the true supporters of my research, thank you for being together with me and your support is the main source of my strengths in completing this research successfully.

THANK YOU VERY MUCH.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVATIONS AND ACROMYMS

AI	Artificial Intelligent
ANFIS	Adaptive Network based Fuzzy Inference System
ANN	Artificial Neural Network
AUC	Area under ROC Curve
CC	Pearson's Correlation Coefficient
CC-GA	Pearson's Correlation Coefficient and Genetic Algorithm
CV	Cross-validation
FIS	Fuzzy Inference Systems
FL	Fuzzy Logic
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
IHC	Immunohistochemistry
IRPA	Intensification for Research in Priority Areas
KNN	k-nearest neighbours
LR	Logistic Regression
MLP	Multi Layer Perceptron
NPC	Nasopharyngeal Carcinoma
OCDTBS	Malaysian Oral Cancer Database and Tissue Bank System
OCRCC	Oral Cancer Research and Coordinating Centre
PSO	Particle Swarm Optimization
ROC	Receiver Operating Characteristic
SCC	Squamous Cell Carcinomas
SVM	Support Vector Machine
TMaA	Tissue Macroarray
TNM	Tumour-Node-Metastasis
TN	True Negative
TP	True Positive
UMMC	University Malaya Medical Centre

CHAPTER 1

INTRODUCTION

1.1 Background

Various artificial intelligent (AI) methods have been applied in the diagnosis or prognosis of cancer research such as, artificial neural networks, fuzzy logic, genetic algorithm, support vector machine and other hybrid techniques (Baker and Abdul-Kareem, 2007; Abdul-Kareem et al., 2002; Dom et al., 2007; Futschik et al., 2003; Gevaert et al., 2006; Hassan et al., 2010; Kawazu et al., 2003; Li et al., 2007a; Passaro et al., 2005; Rao et al. 2011; Saritas et al. 2010; Seker et al., 2003; Thongkam et al., 2008; Xu et al., 2005; Zhong et al., 2011). From the medical perspective, diagnosis is to identify a disease by its signs and symptoms while prognosis is to predict the outcome of the disease and the status of the patient, whether the patient can survive or recover from the disease or vice versa. Researchers have proved that AI methods could generate more accurate diagnosis or prognosis results as compared to traditional statistical methods (Dom et al., 2007; Kawazu et al., 2003; Li et al., 2007a; Passaro et al., 2005; Rao, et al., 2011; Seker et al., 2003; Thongkam et al., 2008).

Normally, clinical data, pathological data or genomic data/microarray data together with socio-demographic data are used in researches either involving diagnosis or that with respect to prognosis. Clinical data refers to the signs and symptoms directly observable by the physicians, examples are the size of primary lesion, clinical neck node, clinical staging, metastasis, and so on. While, pathological data relates to the results obtained from the laboratory examination and the parameters are pathological staging, number of neck nodes, tumour size and thickness and other post surgical pathologic parameters. In

some researches, both clinical and pathological data are used, and are referred as the term clinicopathologic data.

On the other hand, genomic marker is the alterations in the DNA that may indicate an increased risk of developing a specific disease or disorder (Institute, 2010). A genomic marker may be used to see how well the body responds to a treatment for a disease or condition (Institute, 2010). Different types of cancers might have different genomic markers, the most common genomic marker that is currently being investigated by the researchers is *p53*.

Currently, there are very few published articles on researches that combine both clinicopathologic and genomic data. Research has shown that prognosis results are more accurate when using both clinicopathologic and genomic data, the examples are Futschik et al., (2003) in diffuse large B-cell lymphoma (DLBCL) cancer, Gevaert et al., (2006) and Sun et al., (2007) in breast cancer, Exarchos et al. (2011), Oliveira et al., (2008), and Passaro et al., (2005) in oral cancer, and Catto et al., (2006) in bladder cancer.

Oral cancer starts in the mouth, also called the oral cavity. The oral cavity includes the lips, the inside lining of the lips and cheeks (buccal mucosa), the teeth, the gums, the front two-thirds of the tongue, the floor of the mouth below the tongue, the bony roof of the mouth (hard palate), and the area behind the wisdom teeth (retromolar trigone) (Society, 2010).

The mortality rate for oral cancer is high (at approximately 50%) because the cancer is usually discovered late in its development. Well known risks associated with this cancer

include smoking, alcohol consumption, tobacco use, and betel quid chewing. The World Health Organization (WHO) expects a worldwide rise in oral cancer incidence in the next few decades due to high smoking prevalence and increasing cases of unhealthy diet. Almost two-thirds of oral cancer occurs in developing countries such as African and Asian countries, and this geographic variation probably reflects the prevalence of specific environmental influences (Oliveira et al., 2008). Besides socio-demographic and habits factors, there are still other factors associated with oral cancer such as viral infection, genetic factors, diet, and poor oral hygiene (Jefferies and Foulkes, 2001; Mehrotra and Yadav, 2006; Oliveira et al., 2008; Reichart, 2001; Sunitha and Gabriel, 2004).

According to the Malaysian Cancer Statistics, Peninsular Malaysia 2006, oral cancer can be divided into five main categories based on the cancer sites, namely, tongue, mouth, salivary glands, lip and other sites. In Malaysia, Indians are more susceptible to oral cancer and Indian women face the greatest risk, this might be related to their habits of betel quid chewing (Omar et al., 2006). Tongue cancer is listed as the sixth top most frequent cancer (4.6%) in Indian male (after colorectal cancer, prostate gland cancer, lung cancer, stomach cancer and bladder cancer), and mouth cancer is listed as the fourth top most frequent cancer (7.3%) in Indian female (after breast cancer, cervix uteri cancer and colorectal cancer).

A common problem associated with medical dataset is small sample size. It is time consuming and costly to obtain large amount of samples in medical research and the samples are usually inconsistent, incomplete or noisy in nature. Moreover, high accuracy and reliable estimation is needed in medical diagnosis and prognosis where the subsequent decisions have serious consequences on patients. Thus, identifying the high

risk diagnostic/prognostic markers will aid the clinicians in improving the accuracy of prediction of an individual patient's diagnosis/prognosis. The small sample size problem is more visible in the oral cancer research since oral cancer is not one of the top ten most common cancers in Malaysia, hence there are not many cases. For example, in Peninsular Malaysia, there are only 1,921 new oral cancer cases from 2003 to 2005 (Gerard et al., 2005) and 592 new oral cancer cases in the year 2006 (Omar et al., 2006) as compared to breast cancer, where the incidence between 2003 and 2005 is 12,209 and the incidence for 2006 is 3,591. Out of these oral cancer cases, some patients are lost to follow-up, some patients seek treatments in other private hospitals and thus, their data are not available for this research. Another reason for small sample size is caused by the medical confidentiality problems. This can be viewed from two aspects, namely, patients and clinicians. Some patients do not wish to reveal any information about their diseases to others, and are not willing to donate their tissues for research/educational purposes. As for clinicians, some may not want to share patients' data with others especially those from the non-medical fields, while some do not keep their medical records in the correct medical form. From those available cases, some patients' clinicopathologic data are incomplete, some tissues are missing due to improper management and some are duplicated cases. Due to that, the number of cases that can actually be used for this research is very limited.

In this research, an oral cancer prognostic model is developed. This research used real-world oral cancer dataset which has been collected locally in the Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, University of Malaya, Malaysia. Clinicopathologic data is available from the OCRCC while the genomic data is obtained through the process of immunohistochemistry (IHC) staining on selected oral cancer tissues. IHC is a method of localizing the antigens or proteins in cells or

tissues by the use of primary antibody as specific reagents through antigen-antibody interactions that are visualized by a marker such as fluorescent dye or enzyme. The prediction model is designed for small datasets where high accuracy can be achieved using only a small sample size. The model takes both clinicopathologic and genomic data that have been determined in order to investigate the relationship of each marker or combination of markers to the accuracy of the prognosis of oral cancer.

1.2 Problem Statement

The mortality rate of oral cancer is high yet there are very few studies using AI techniques in the prognosis of oral cancer. The application of AI techniques in oral cancer susceptibility prediction was done by Arulchinnappan et al. (2011); Dom et al. (2007, 2008 & 2010); Passaro et al. (2005) and Baronti & Starita (2007). Screening prediction was done by Speight & Hammond (2001). The prediction of lymph nodes metastasis in oral cancer was done by Kawazu et al. (2003) and oral cancer diagnosis prediction was done by Kent (1996). The prediction of oral cancer reoccurrence was done by Exarchos et al. (2011). Furthermore, there is no Malaysian study yet on the application of AI techniques in the prognosis of oral cancer. A previous study that utilized AI techniques in oral cancer was done by Dom et al. (2007, 2008 & 2010) and it was on the susceptibility prediction. Therefore, there is a need to investigate how AI techniques can be used in the prognosis of oral cancer. We believe the research will result in the development of a tool that is adaptable to the multi-ethnic society in Malaysia, and hence, benefit the Malaysian people.

Second, in order to make an accurate prognosis/survival prediction, one needs to include both clinicopathologic markers and genomic markers. Currently, many studies use only clinicopathologic factors without taking into consideration the tumor biology

and molecular information, while some studies use genomic markers or microarray information only without the clinicopathologic parameters. Thus, these studies may not be able to predict the diagnosis/prognosis of a patient effectively. It has been proven by Catto et al., (2006) in bladder cancer, Futschik et al., (2003) in DBLCL cancer, Gevaert et al., (2006) and Sun et al., (2007) in breast cancer, Exarchos et al. (2011), Oliveira et al., (2008) and Passaro et al., (2005) in oral cancer, Seker et al., (2003) in breast and prostate cancer, that prognosis results are more accurate when using both clinicopathologic and genomic data.

Third, traditional statistical methods such as Kaplan-Meier method, logistic regression, Cox regression and decision trees are usually used in the prediction of cancer survival. However, in Dom et al., (2007 & 2008), Jerez et al, (2010), Hayward et al. (2010), Kawazu et al., (2003), Li et al., (2007), Lin & Chuang, (2010), Passaro et al., (2005), Rao, et al., (2011), Regnier-Coudert et al. (2011), Seker et al., (2000 & 2003), and Thongkam et al., (2008) had proved that AI techniques can generate more accurate predictions than statistical methods. AI techniques are good for handling noisy and incomplete data, and significant results can be attained with small sample size. Thus, there is a need to develop an AI model which is able to improve prognosis based on the individual patient's conditions.

1.3 Research Aim

The main aim for this research is to apply AI techniques in the prognosis of oral cancer based on the parameters of the correlation of clinicopathologic and genomic factors. This research is highly influenced by the works of Catto et al., (2006) in bladder cancer, Futschik et al., (2003) in DBLCL cancer, Gevaert et al., (2006) and Sun et al., (2007) in breast cancer, Exarchos et al. (2011), Oliveira et al., (2008) and Passaro et al., (2005) in

oral cancer, Seker et al., (2003) in breast and prostate cancer, who have used both factors in the prognosis of cancer studies.

Passaro et al., (2005) used AI techniques in the oral cancer susceptibility studies. They proposed a hybrid adaptive system inspired from learning classifier system, decision trees and statistical hypothesis testing. The algorithm can work with different data types and is robust to missing data. The dataset includes both demographic data and 11 types of genes. Their results showed that the proposed algorithm outperformed the other algorithms of Naive Bayes, C4.5, neural network and XCS (Evolution of Holland's Learning Classifier). However, they validated the algorithm on the Winconsin Breast Cancer dataset (WBC), it will be more appropriate if the benchmark dataset is chosen from the same type of cancer.

Oliveira et al. (2008) focused on the 5-year overall survival in a group of oral squamous cell carcinoma (OSCC) patients and investigated the effects of demographic data, clinical data and genomic data, and human papillomavirus on the prognostic outcome. They used the statistical method for the prediction and their results showed that the 5-year overall survival was 28.6% and highlighted the influence of *p53* immunoexpression, age and anatomic localization on OSCC prognosis. In this research, no AI methods were used and compared.

Another oral cancer research that was done by Exarchos et al. (2011) was in the oral cancer reoccurrence. Bayesian network was used and compared with ANN, SVM, decision tree, and random forests. They used multitude of heterogeneous data which included clinical, imaging and genomic data. They build a separate classifier for different types of data and combined the best performing classification schemes. They

claimed that they had achieved an accuracy of 100% with the combinations of all types of data and proved that the prediction accuracy is the best when using all types of data. However, more than 70 markers are required for their final classifier.

This work differs from that of the researchers named above is that we are working in the domain of oral cancer prognosis using AI techniques, which based on our literature review, is the first study in Malaysia. Furthermore we tested the system by using data collected locally, here in the OCRCC, Faculty of Dentistry, University of Malaya, Malaysia. We used the same classifier for both clinicopahologic and genomic data and we compared the results generated with and without the inclusion of genomic data. In addition, we also compared our results with the results generated by other AI methods and statistical method. Lastly, we validated our results with the human experts' (oral cancer clinicians) prediction.

Since the mortality rate of oral cancer is high, there is a need to develop a computerized tool that can aid clinicians in the decision support stage and to identify the high risk markers in order to better predict the survival rate for each oral cancer patient and to extend the tool to other cancer/disease prognosis prediction.

1.4 Research Questions

In this research, we hypothesize that by using feature selection method, neuro-fuzzy and cross-validation techniques, we can predict the prognosis/survival of oral cancer more accurately with a few promising markers and coupled with the problem of small sample size. With this hypothesis, some questions are formulated:

1. What are the clinicopathologic and genomic markers that are most commonly used in the prognosis of oral cancer?

2. How to ensure the accuracy of the immunohistochemistry (IHC) staining for locating the genomic markers?
3. Which feature selection method is most suitable for the oral cancer prognosis dataset?
4. What is the optimum number of markers to use in oral cancer prognosis using the proposed model?
5. Is the proposed model more accurate than the traditional statistical methods, and other AI methods?
6. How to evaluate the performance of the proposed model?

1.5 Research Objectives

The objectives of this study are:

1. To identify the most common clinicopathologic markers associated with oral cancer prognosis.
2. To analyse the genomic markers from the results of immunohistochemistry (IHC) staining.
3. To determine the optimum subset of markers for oral cancer prognosis using feature selection methods.
4. To develop a prognostic model for oral cancer prognosis using ANFIS techniques and to prove that the proposed model is the optimum tool for oral cancer prognosis.
5. To prove that the prognosis of oral cancer is more accurate when both the clinicopathologic and genomic markers are considered.

1.6 Significance of Study

Based on our literature review, we found out that this is the first study in Malaysia which applies AI techniques in oral cancer prognosis prediction using both clinicopathologic and genomic markers. The study is based on real world data, namely, the Malaysian oral cancer dataset provided by the OCRCC, Faculty of Dentistry, University of Malaya.

As for genomic data, immunohistochemistry staining will be performed on the selected oral cancer tissues and the results will be analysed. We believe this is a novel study for oral cancer prognosis involving clinicopathologic and genomic data as we indicated in our early literature review.

An optimum subset of markers for oral cancer prognosis will be obtained using the combination of feature selection method and the classification method. We believe that the model with fewer markers will help to predict oral cancer survival with higher accuracy and thus to avoid the over-fitting problems. Over-fitting occurs when there are too many parameters relative to the number of samples.

1.7 Scope and Limitation

This research focuses on the identification of optimum markers for oral cancer prognosis by using feature selection and AI methods for comparative analysis. There is no clinical testing/evaluation involve in this study as the developed model is not ready for the use of clinician yet. More tests and experiments are needed to further verify the results obtained in this research.

This research considers 17 variables which include 15 clinicopathologic variables and 2 genomic variables. There are a number of genes that can be considered as genomic markers in the prognosis of oral cancer as discussed in Chapter 2. Due to time and cost limitations, only two genes are chosen based on the recommendations of oral pathologists and clinicians as well as the literature. Testing for one particular gene involves the cost of the testing materials (i.e. reagent, antibody, etc.), time taken for the tests, the efforts and time of the laboratory technician and oral pathologists as discussed in Chapter 5. Therefore, other genes will be included in future works and will not be considered in this study.

1.8 Thesis Overview

This thesis is organized as follows:

- Chapter 1 provides the introduction of the proposed study including the problem statement, research aim, research questions, research objectives, significance of the study and the scope and limitation of the research.
- Chapter 2 discusses various AI techniques in medical research, introduction to artificial neural network, fuzzy logic, neuro-fuzzy, support vector machine, and logistic regression techniques, re-sampling techniques, feature selection techniques, and the model mesurements.
- Chapter 3 discusses the oral cancer, overview of oral cancer in Malaysia, risk factors of oral cancer, and clinicopathologic and genomic markers of oral cancer, cancer management, and survival analysis.
- Chapter 4 discusses the general methodology used in this research.
- Chapter 5 presents a more detailed discussion of methodology concerning on the preparations and procedures for acquiring oral cancer prognosis data involving both clinicopathologic and genomic data.

- Chapter 6 discusses the feature selection methods used in this research in order to reduce the number of inputs and to obtain an optimum subset of the markers and also discusses the classifier used in this research.
- Chapter 7 discusses the results, discussions, comparisons and validation of the developed model with other AI and statistical models.
- Chapter 8 concludes the presented works and proposes some future works.

CHAPTER 2

ARTIFICIAL INTELLIGENCE IN CANCER RESEARCH

2.1 Introduction

There are three important areas in the application of artificial intelligent (AI) techniques in cancer prediction which are: the prediction of cancer susceptibility, the prediction of cancer recurrence and the prediction of cancer survival. In the cancer susceptibility prediction, one is trying to predict the likelihood of developing a type of cancer prior to the occurrence of the disease based on the selected risk factors. While in the cancer recurrence prediction, one is trying to predict the likelihood of redeveloping cancer after treatment and after a period of time in which no cancer could be detected. In the prediction of cancer survival, one is trying to predict an outcome after the diagnosis of the disease, that is, the chance that a patient will survive or die (Cruz et al., 2006).

Typically, cancer prognosis involve multiple physicians from different specialties using different subsets of genomic markers and multiple clinicopathologic factors, including the socio-demographic data (age, gender, ethnic) of the patient, risks factors (smoking, alcohol drinking, betel quid chewing), the location and type of cancer, size of the tumour, metastasis of lymph nodes, staging classification (stage 1 to 4) and types of treatment (surgery, radiotherapy, chemotherapy or a combinations of these methods). It is not easy for even the most skilful physicians to come up with an accurate and reasonable prognosis, and it is not 100% accurate (Fielding et al., 1992; Catto et al., 2006; Reichart, 2001).

Unfortunately these conventional clinicopathologic parameters generally do not provide enough information to make robust prognoses. Ideally what is needed is some very specific molecular details about either the tumour or the patient's own genetic make-up which are the genomic markers (Colozza et al., 2005).

With the rapid development of genomic (DNA sequencing, microarrays), proteomic (protein chips, microarrays, immunohistology) and imaging (CT scan, PET scan, MRI) technologies, this kind of molecular-scale information about patients or tumours can now be readily acquired. If these molecular patterns are combined with clinicopathologic data, the robustness and accuracy of cancer prognoses can be improved (Cruz et al., 2006).

The prognostic models are complex tools in the decision making that combine two or more items of patient data to predict the clinical outcomes. These models are intended to help the clinicians in making difficult clinical decisions such as ordering invasive test or choose patients for certain clinical trials. However, most of the published prognostic models are rejected by the clinicians due to lack of clinical credibility (no clinical testing and evaluation, data reliability and model simplicity) and lack of clinical accuracy, evidence and effectiveness (Wyatt & Altman, 1995). A way to improve the clinical acceptability of the prognostic models is combining the prognosis generated by the model with the doctor's own estimate of prognosis and clinical validation (Goddard et al., 2011; Liu et al., 2006; Wyatt & Altman, 1995).

The diagnosis/prognosis models are developed based on the clinical prediction rules. The purpose of clinical prediction rules is to reduce the uncertainty inherent in medical practice by defining how to use clinical findings to make predictions. The clinical

prediction rules derived from the clinical observations done by the clinicians. These models can help clinicians identify patients who require diagnostic tests, treatment or to predict the survival rate of patients (Wasson et al., 1985). The scientific methods and testing procedures of the prediction models were discussed in the state of the art papers such as Wasson et al. (1985), Spiegelhalter et al. (1983), and Wyatt & Altman (1995). These papers discussed the evaluation and validation of the clinical predictive models by using mathematical/statistical techniques. The evaluation of clinical prediction models has long been recognized as an important part of the overall field of medical computing. However, in this research, we focused on the identification of optimum prognosis markers for oral cancer by using feature selection and AI techniques. As this is a preliminary study of the research, there is no evaluation/testing/implementation of the developed model into clinical use. The purpose of this research is to prove that the prognosis is better with both clinicopathologic and genomic data if compared to only clinicopathologic data.

There is a growing interest in the application of AI techniques in medical research. This is due to the nature of AI approaches that perform well in domains where the sample size is small, as opposed to the statistical methods which require a "big enough" sample size in order to achieve statistically significant results (Mitchell, 1997). It is hard to get a large amount of samples in medical research as it takes a long time and is very costly, and the samples are usually either incomplete or noisy. This is where AI techniques are needed in making the diagnosis or prognosis more accurate.

Since the introduction of AI to this field, numerous algorithms have been designed and applied to medical datasets. Various methods have been applied in either the diagnosis or prognosis of cancer such as artificial neural network, Bayesian network, fuzzy logic,

support vector machine, genetic algorithm and other hybrid methods. Most of these researches compare a new method with the traditional ones, affirming the effectiveness and efficiencies of their methods in particular datasets which will be further discussed in section 2.2.

2.2 Artificial Intelligent Techniques In Cancer Research

This section reviews some major artificial intelligent (AI) techniques which have been applied in cancer research. Artificial neural network, fuzzy logic, genetic algorithm, and Bayesian methods are amongst the most common AI techniques used in cancer research. In addition, hybrid methods will be discussed too as these methods are getting more attention recently. Most researches focus on breast cancer study (Akay, 2008; Bellaachia & Guven, 2006; Delen et al., 2005; Gevaert et al., 2006; Jerez et al. 2011; Hassan et al., 2010; Sivaraksa et al., 2008; Seker et al., 2003; Song et al., 2005; Sun et al., 2007; Thongkam et al., 2008; Xu et al., 2005), diffuse large B-cell lymphoma (DLBCL) (Futschik et al., 2003; Xu et al., 2005), nasopharyngeal carcinoma (Abdul-Kareem et al., 2002; Baker & Abdul-Kareem, 2007; Wang et al., 2009), bladder cancer (Li et al., 2007; Almal et al., 2006; Catto et al., 2006), laryngeal cancer (Jones, 2006), prostate cancer (Regnier-Coudert et al., 201;, Seker et al., 2003; Castanho et al., 2008) and pancreatic cancer (Hayward et al., 2010). Whereas, the application of AI techniques in oral cancer susceptibility and diagnosis was done by Arulchinnappan et al. (2011), Dom et al. (2007, 2008 & 2010), Passaro et al. (2005) and Baronti & Starita (2007). Oral cancer screening prediction was done by Speight & Hammond (2001). The prediction of lymph nodes metastasis in oral cancer was done by Kawazu et al. (2003), oral cancer diagnosis prediction by Kent (1996) and the prediction of oral cancer reoccurrence was done by Exarchos et al. (2011).

Table 2.1: Summary of cancer research using AI techniques

Type of cancer	Type of prediction	AI technique	Benchmark	Improve-ment (%)	Training data	Reference
Breast	Diagnosis	SVM with feature selection	N/A	N/A	clinical	Akay F.M., 2009
Breast	Prognostic	Naïve Bayes, ANN, Decision tree	N/A	N/A	Clinical	Bellaachia, 2006
Breast	Prognostic	ANN, decision trees	LR	4	Clinical	Delen et al., 2005
Breast	Prognostic	Bayesian network	70 genes	No significant difference	Clinical & genomic	Gevaert et al., 2006
Breast	Diagnosis	Hybrid hidden Markov model (HMM)-fuzzy	Denfis, SVM, NEFCLASS	Yes	Clinical	Hassan et al., 2010
Breast	Prognostic	MLP, KNN, SOM	Statistical methods	Yes	Clinical	Jerez et al., 2011
Breast	Prognostic	ANN	N/A	N/A	Genomic	Sivaraksa et al., 2008
Breast	Diagnosis	ANFIS	Different feature selection algorithm	Yes	Clinical	Song et al., 2005
Breast	Prognostic	I-RELIEF	70-gene Clinical	20	Clinical & genomic	Sun et al., 2007
Breast	Prognostic	AdaBoost	Bagging, C4.5, C-SVC, Random forest	1 - 4	Clinical	Thongkam et al., 2008
Breast	Prognostic	SVM (GA-CG-SVM)	C5.6 decision tree, KNN	12.5	Genomic	Zhong et al., (2011)
Breast & Prostate	Prognostic	Fuzzy k-nearest neighbour	Logistic regression, ANN	2 - 16	Clinical	Seker et al., 2003
Bladder	Prognostic	Genetic programming	N/A	N/A	Genomic	Almal et al., 2006
Bladder	Prognostic	Neuro-fuzzy	ANN, LR	Yes	Clinical & genomic	Cotto et al., 2006
Bladder	Diagnosis	ANN (Mega-trend-diffusion)	ANN, DT	12 - 40	Genomic	Li et al., 2007
DLBCL	Prognostic	ANN & Bayesian network	Compare with single predictor module	9-14	Clinical & genomic	Futschik et al., 2003

Type of cancer	Type of prediction	AI technique	Benchmark	Improve-ment (%)	Training data	Reference
DLBCL	Prognostic	Particle swarm optimization (PSO)	N/A	N/A	Genomic	Xu et al., 2005
Laryngeal	Prognostic	ANN	Cox regression	No, but ANN is more sensitive	Clinical	Jones et al., 2006
NPC	Prognostic	ANN (MLP, recurrence)	Statistical methods	ANN performed better	Clinical	Abdul-Kareem et al., 2002
NPC	Prognostic	Genetic algorithm	N/A	N/A	Clinical	Baker & Abdul-Kareem, 2007
NPC, Leukaemia, colon, breast	Cancer classificati on	ANN	FDA, kNN, Bayesian network, SVM	Yes	Genomic	Wang et al., 2009
Oral	Susceptibil ity	Fuzzy correlation	N/A	N/A	Clinical	Arulchinna ppan et al., 2011
Oral	Susceptibil ity	Learning Classifier, decision tree, statistical methods	Naïve Bayes, C4.5, ANN	6 - 20	Smoking & genomic	Baronti & Starita, 2007
Oral	Susceptibil ity	Fuzzy regression	Statistical, Logistic regression	No significant difference	Clinical & genomic	Dom et al., 2007, 2008 & 2010
Oral	Reoccuren ce	Bayesian network	ANN, SVM, DT, random forests	Bayesian network outperform ed the others	Clinical, genomic & imaging	Exarchos et al., 2011
Oral	Lymph node metastasis	ANN	Radiologists prediction	No significant difference	Clinical	Kawazu et al., 2003
Oral	Diagnosis	Genetic programming	N/A	N/A	Clinical	Kent, 1996
Oral	Prognostic	ANFIS	Statistical methods	ANFIS is more effective tool.	Genomic	Muzio et al., 2005
Oral	Susceptibil ity	*XCS (evolution of Holland's Learning Classifier)*	DT	4 - 20	Clinical & genomic	Passaro et al., 2005
Oral	Screening	ANN	C4.5	No significant difference	Clinical	Speight & Hammond, 2001
Pancreatic	Prognostic	Bayesian network	DT, LR, ANN	AI techniques performed better than statistical methods	Clinical	Hayward et al., 2010

Type of cancer	Type of prediction	AI technique	Benchmark	Improve-ment (%)	Training data	Reference
Prostate	Diagnosis	Fuzzy rule-based	N/A	N/A	Clinical	Castanho et al., 2008
Prostate	Prognostic	Bayesian network, ANN	LR	Bayesian network outperformed the others	Clinical	Regnier-Coudert et al., 2011

*ANN-Artificial neural network, ANFIS-Adaptive Network Based Fuzzy Inference System, DT-Decision Trees, FDA-Fisher Discriminant Analysis, kNN-k-Nearest Neighbour, LR-Logistic Regression, MLP-Multilayer perceptrons, SVM-Support Vector Machine

2.2.1 Artificial Neural Network

The use of artificial neural networks (ANNs) in cancer research has vastly proliferated during the last few decades. Neural network analysis has been shown to be particularly useful in those cases where the problem to be solved is ill defined, and development of an algorithmic solution is difficult. This is exactly the situation with cancer data where a highly nonlinear, almost brain-like, approach is required.

An ANN is a computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. An artificial neuron is a computational model inspired by biological neurons. Biological neurons receive signals through synapses located on the dendrites or membrane of the neuron (Figure 2.1). When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal though the axon. This signal might be sent to another synapse, and might activate other neurons. In ANN, these basically consist of inputs (like synapses), which are multiplied by weights, and then computed by an activation function which determines the output of the neuron, as shown in Figure 2.2 (Gerhenson, 2003).

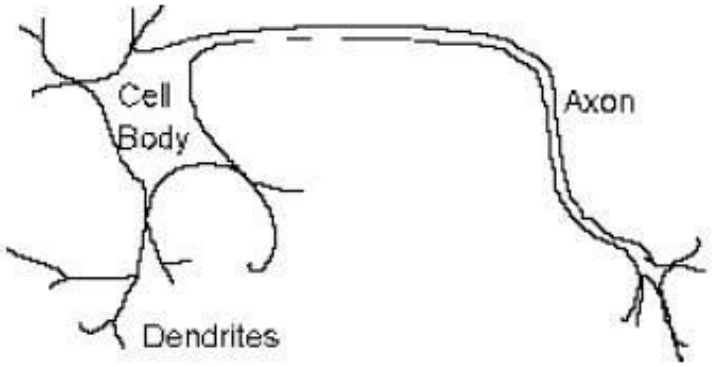


Figure 2.1: The biological neuron

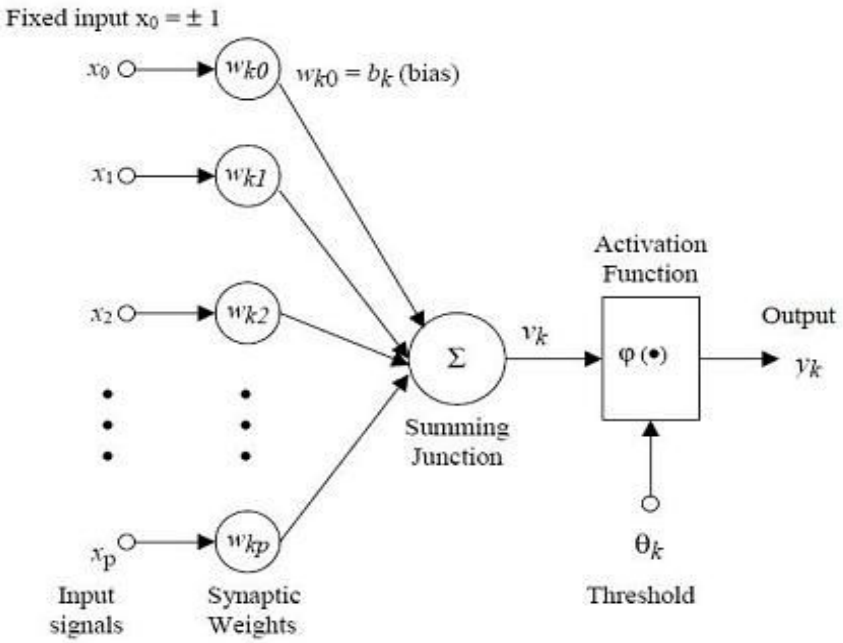


Figure 2.2: An ANN Model

The architecture of an ANN is concerned with the way the neurons are divided into layers in a network. The ANN has at least 2 layers, which are the input layer and the output layer. Most neural networks have one or more middle layers known as the hidden layer (Abdul-Kareem, 2001).

ANN can be divided into two main groups based on the pattern of connections, which are feed forward neural networks and recurrent neural networks. In the feed forward neural networks, the signal flows from input neuron to output neuron in a forward direction. The data processing can extend over multiple layers but there is no feedback connection (Abdul-Kareem, 2001). The examples of feed forward neural network are single-layer perceptron and multi-layer perceptron (MLP). An example of an MLP is shown as in Figure 2.3.

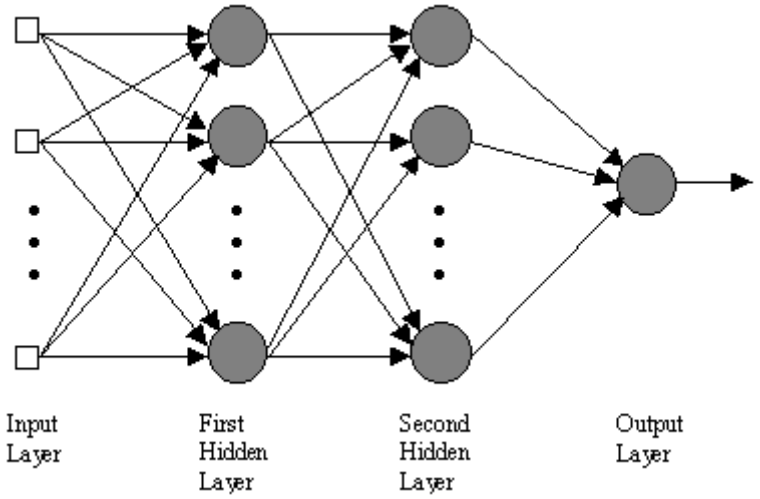


Figure 2.3: An example of MLP

The recurrent neural networks provide feedback connections, so that the networks can incorporate context or temporal information. The examples of recurrent neural networks are Hopfield network and Elman network. Figure 2.4 shows an example of recurrent neural network. Most ANNs are structures using multi-layered feed-forward architecture, meaning they have no feedback, or no connections that loop (Cruz and Wishart, 2006). The design and structure of an ANN must be customized or optimized for each application.

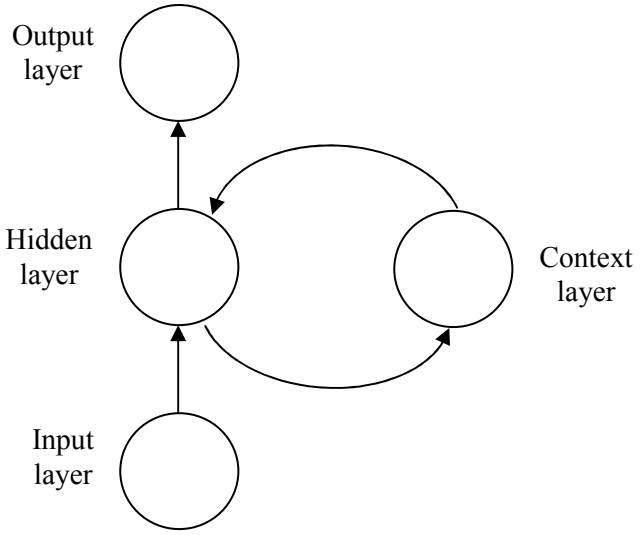


Figure 2.4: An example of recurrent neural network

An ANN needs to be trained in order to learn the patterns and change the weights according to the rules. The training methods can be categorised into supervised learning and unsupervised learning. In supervised learning, the network is trained by providing it with input and output patterns. In unsupervised learning, an output is trained to respond to the pattern of inputs. The network discovers the similarity between the inputs, and the similar inputs are clustered to the same output (Rios, 2010).

Kawazu et al. (2003) proposed a three-layer feed-forward network with a back-propagation algorithm in the prediction of lymph node metastasis of patients with oral cancer. They constructed numerous different architectures with different number of hidden layers and units. The diagnosis was most accurate (= 93.6%) when the network consisted of two hidden layers, namely with 6 and 4 units for each layer. Their results showed that the network performance was equivalent to the analysis made by radiologists and was better than statistical analysis (Quantification theory type II).

Another example that utilised ANNs in the prognosis of cancer was done by Abdul-Kareem et al. (2002) to predict the prognosis of nasopharyngeal carcinoma (NPC). Two neural network models were designed i.e. multi-layered feed-forward network and the Elman recurrent network. Both networks consisted of 22 hidden nodes in the middle layer. Their results showed that the predictive performance of the multi-layered feed-forward network was better than the recurrent network and statistical method.

2.2.2 Genetic Algorithm

Genetic algorithms (GA) were formally introduced in the United States in the 1970s by John Holland at the University of Michigan. Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms (also known as evolutionary computation) that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover.

The algorithm starts with a set of solutions (represented by chromosomes) called the population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions that form new offsprings are selected according to their fitness - the more suitable they are the more chances they have to reproduce. This is repeated until some condition (for example the number of populations or improvement of the best solution) is satisfied (Obitko, 1998).

Pappalardo et al. (2006) proposed the use of genetic algorithm to find out effective therapies for protecting virtual mice from mammary carcinoma. An accurate model of the immune system responses to vaccination was developed and *in silico* experiments consisting of a large population of individual mice were performed. The genetic

algorithm model was used as a fitness evaluator to find a schedule which controlled the growth of cancer cells by a minimal number of vaccine injections. Their results showed that the genetic algorithm model found complete immunoprevention with a much lighter vaccination schedule, and the number of injections was roughly one third of those used in conventional schedules (Chronic).

Baker et al. (2007) used genetic algorithm (GA) in the prognosis of nasopharygeal carcinoma (NPC). Two models were developed i.e. GA with algebraic rule-based classifier and GA with a hybrid function. The survival time was provided to the nearest one year, up to ten years. A series of sub-classifiers were generated to predict a specific time range for one-year interval-based classifier, and then chained together to operate in unison and resolved efficiently the prognosis outcome for a given patient.

2.2.3 Fuzzy Logic

The concept of Fuzzy Logic (FL) was conceived by Lotfi Zadeh, a professor at the University of California at Berkley. Fuzzy logic is a form of multi-valued logic derived from the fuzzy set theory to deal with reasoning that is approximate rather than precise. Fuzzy logic is used in system control and analysis design, because it shortens the time for engineering development and sometimes, in the case of highly complex systems, is the only way to solve the problem (Jang, 1993).

Fuzzy logic starts with the concept of a fuzzy set. A fuzzy set is a set without a crisp, clearly defined boundary. It can contain elements with only a partial degree of membership. A membership function is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. Fuzzy logic models, called fuzzy inference systems (FIS), consist of a number of

conditional "if-then" rules, membership functions and logical operations. There are two types of fuzzy models namely the Mamdani model and the Takagi-Sugeno model. Mamdani's fuzzy inference method is the most commonly seen fuzzy methodology. The output membership functions of Mamdani's model are fuzzy sets. The main difference between Mamdani and Takagi-Sugeno is that the output membersip functions of Takagi-Sugeno's model are either linear or constant (Kaehler, 1993).

Fuentes-Uriarte et al. (2008) presented two fuzzy logic methods for breast cancer diagnosis namely, fuzzy clustering with the Fuzzy C-Means (FCM) algorithm and fuzzy inference system (FIS). They applied their algorithms to the Winsconsin Breast Cancer Diagnosis (WBCD) database. The FCM was used to find the similarities between different variables while the FIS was implemented with a genetic algorithm for creating and activating the optimal rules. Their simulated results showed that the FCM performed better by classifying 99.3% of the data in WBCD correctly as compared to 80.136% in FIS.

Dom et al. (2008) proposed a fuzzy regression model for the prediction of oral cancer susceptibility. The prediction of oral cancer susceptibility as a function of demographic profiles (age, gender, ethnicity), risk habits (smoking, alcohol drinking, tobacco chewing) and genetic markers (GSTM, GSTT1) were done using statistical logistic regression and fuzzy regression models. The models were tested on a sample of 84 oral cancer patients and 87 controls data. The results show that there was no significant difference in the prediction performance of fuzzy regression model (AUC = 0.888) and the logistic regression model (AUC = 0.851). Thus, it is feasible to use fuzzy regression model for oral cancer susceptibility studies.

Another example that utilised fuzzy logic in cancer research was done by Seker et al. (2000). They developed a model for breast cancer prognosis by using fuzzy k-nearest neighbour algorithm (FK-NN). The dataset consisting of 100 cases with seven inputs and two outputs were predicted, namely, nodal involvement assessment and 5-year survival analysis. Their highest accuracy for nodal involvement assessment was 78% for the three-marker subset and survival analysis was 88% for the five-marker subset. They claimed that fewer inputs were sufficient to yield more accurate results if compared to all inputs.

2.2.4 Bayesian Network

A Bayesian network is a probabilistic graphical model that represents a set of variables and their probabilistic independencies. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Bayesian networks are directed acyclic graphs, where the nodes represent the variables and the edges represent the conditional independencies between the variables. Bayesian networks that model sequences of variables are called dynamic Bayesian networks (Sebastiani et al., 2003).

An example of a research that utilized Bayesian networks in cancer research was done by Gevaert et al. (2006). They predicted the prognosis of breast cancer by integrating clinicopathologic and microarray data with Bayesian networks. The main advantage of this model was that the model was able to integrate these data sources in several ways and aided in the investigation and understanding of the structure and parameters of the

model. Their results showed that the partial integration method was the most promising method and the integrated use of clinicopathologic and microarray data outperformed the indices based on clinicopathologic data and has comparable performance with the 70 genes prognosis profile.

2.2.5 Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning algorithm that can be used for classification or regression problems. The SVM was developed by Corinna Cortes and Professor Vladimir Vapnik in 1995 (Cortes & Vapnik 1995). SVM performs the classification by constructing an *N*-dimensional hyperplane that optimally separates the data into two categories. An example of a 2-dimensional hyperplane is shown in Figure 2.5.

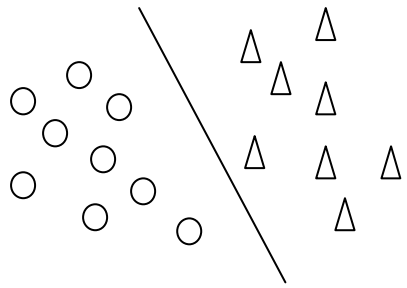


Figure 2.5: An example of a 2-dimensional hyperplane

When given a training set of (x_i , y_i), $i = 1, ..., l$ where $x_i \in R^n$, the optimization solution for the support vector machine is as below:

$$\min_{w,b,\varepsilon} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \varepsilon_i \qquad (2.1)$$

subject to $y_i (w^T \emptyset(x_i) + b) \geq 1 - \varepsilon_i , \ \varepsilon_i \geq 0$

w - Orthogonal vector

b - Bias , ε_i - slack variables

The training vector x_i are mapped into a higher dimensional space by the function ∅. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term (Chih-Wei, H. et al., 2010).

The $K(x_i , x_j) \equiv \emptyset(x_i)^T \emptyset(x_j)$ is the kernel function, there are four basic kernel functions used in the SVM, which are (Chih-Wei, H. et al., 2010):

i. linear: $K(x_i , x_j) = x_i^T x_j$

ii. polynomial: $K(x_i , x_j) = (\gamma x_i^T x_j + r)^d , \gamma > 0$

iii. radial basis function (RBF): $K(x_i , x_j) = exp(-\gamma || x_i - x_j ||^2), \gamma > 0$

iv. sigmoid: $K(x_i , x_j) = tanh(\gamma x_i^T x_j + r)$

2.2.5.1 LIBSVM

LIBSVM is a SVM tool for classification, regression and distribution estimation. It was first released in the year 2000, the current version is version 3.11. Currently, LIBSVM is one of the most widely used SVM software, and it has been used in many areas such as computer vision, bioinformatics and natural language processings. LIBSVM supports the following learning tasks (Chih-Chung, C. & Chih-Jen, L., 2011):

1) Support vector classification (SVC) for two-class and multi-class
2) Support vector regression (SVR)
3) One-class SVM

There are 2 steps involve in the LIBSVM, which are, (1) the dataset is trained to obtain a model and (2) the model is used to predict the information for the testing dataset. For the SVC and SVR, can also output probability estimates (Chih-Chung, C. & Chih-Jen, L., 2011).

2.2.6 Hybrid Artificial Intelligent Methods

Currently, there are an increasing number of researchers (Cotto, 2006; Dom, 2007; Futshik, 2003; Hassan, 2010; Muzio, 2005; Zhong, 2011) that use hybrid artificial intelligent methods. A hybrid method involves the use of a combination of two or more artificial intelligent techniques for example fuzzy neural networks, genetic fuzzy system, fuzzy Bayesian system, evolutional neural networks and etc.

Futshik et al. (2003) proposed a hybrid prognostic model for outcome prediction of diffuse large B-cell lymphoma (DLBCL) using the integration of microarray data and clinical parameters. They constructed separate modules for microarray and clinical data. The Bayesian classifier was used for clinical data and the fuzzy neural network classifier was used for genomic data. A prediction accuracy of 87.5% was achieved. Their study demonstrated that the integration of microarray data with clinical data improves disease outcome prediction.

Muzio et al. (2005) evaluated the relationship between the expression of three cell cycle markers (surviving, MIB-1 and PCNA) and human papillomavirus (HPV) infection in oral cancer by using fuzzy neural networks (FNN) namely, ANFIS and traditional statistics. Their findings showed that the FNN is able to differentiate cell cycle pattern for HPV-positive vs. HPV-negative in oral cancer and HPV may have a protective role in the expression level of survival; especially in tobacco smokers.

2.3 Neuro-Fuzzy Systems

Neuro-fuzzy refers to the combination of artificial neural networks (ANN) and fuzzy logic. Fuzzy systems lack in their learning ability; they are not robust to the topological changes of the system, and require *a priori* rules and the disadvantages of ANN are,

high computational cost is required to minimize the over-fitting problems, difficulty in justifying the relations between the input and output variables which are the black box problem, and large sample size. The neuro-fuzzy systems solve the black box problem in neural network and the prior knowledge essential problem in fuzzy logic, and it provides the learning ability to the systems.

Fuzzy systems can be categorized into two families. The first includes linguistic models based on collections of IF-THEN rules, whose antecedents and consequents utilize fuzzy values. It uses fuzzy reasoning and the system behaviour can be described in natural terms. The Mamdani model is one example of this group. The knowledge is represented as:

R^i : If x_1 is A_1^i and x_2 is A_2^i ⋯ and x_n is A_n^i,

then y^i is B^i (2.2)

where,

R^i (i=1,2, ⋯, l) – ith fuzzy rule

x_j (j=1,2 ⋯,n) - input

y^i - output of the fuzzy rule R^i

A_1^i, A_2^i, ⋯, A_n^i, B^i(i=1,2, ⋯,l) - fuzzy membership functions

The second category, based on the Takagi-Sugeno type systems, uses a rule structure that has fuzzy antecedent and functional consequent parts,

R^i : If x_1 is A_1^i and x_2 is A_2^i ⋯ and x_n is A_n^i

then $y^i = a_0^i + a_1^i x_1 + \cdots + a_n^i x_n$ (2.3)

If one needs a more precise solution, then the Takagi-Sugeno-type is the choice, otherwise, the Mamdani-type maybe more suitable (Mitra & Hayashi, 2000). In medical research, accuracy is an important criterion when making a diagnosis or a prognosis, thus the Takagi-Sugeno type is more suitable.

There are two ways in which neuro-fuzzy hybridization could be done. First, a neural network is equipped with the capability of fuzzy information i.e. fuzzy neural network and second, a fuzzy system augmented by neural networks to enhance some of its characteristics like flexibility, speed, and adaptability i.e. neuro-fuzzy system (Mitra & Hayashi, 2000).

There are different types of neuro-fuzzy systems presented in the literature, for example, NEFCON (Nauck & Kruse, 1993), NEFCLASS (Nauck & Kruse, 1995), ANFIS (Jang, 1993 & 1996), GARIC (Berenji, 1992), and FALCON (Lin & Lee, 1991). NEFCLASS and NEFCON (Nauck & Kruse, 1993 & 1995) use a generic fuzzy perceptron to model Mamdani-type neuro-fuzzy systems, thus they are not concentrating on generating the exact solution. Both methods use reinforcement learning rather than the supervised learning. GARIC (Berenji, 1992) uses a differentiable soft minimum function to implement a fuzzy controller. It has a self-tuning ability in the fuzzy logic controller and it utilises the reinforcement learning as well. FALCON's (Lin & Lee, 1991) learning ability is based on the use of the Kohonen learning rule and supervised learning algorithm. There is no refinement of rules after the Kohenen learning using back-propagation method.

In this research, we focused on ANFIS (Jang, 1993 & 1996) as ANFIS implements a Sugeno-type fuzzy system and uses the back-propagation method to learn the

antecedent membership functions. The details about ANFIS are further discussed in the section 2.3.1.

2.3.1 ANFIS

ANFIS stands for Adaptive Neuro-Fuzzy Inference System. ANFIS was first proposed by Jyh-Shing Roger Jang from the Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan. ANFIS implements the Takagi-Sugeno fuzzy inference system. Consider a first order Sugeno fuzzy inference system which contains two rules (Jang, 1993 & 1996):

Rule 1: If X is A_1 and Y is B_1, then $f_1 = p_1x + q_1y + r_1$

Rule 2: If X is A_2 and Y is B_2, then $f_2 = p_2x + q_2y + r_2$

The corresponding fuzzy reasoning mechanism is shown in Figure 2.6, where the firing strength w is the rule weight of all membership functions and output f is the weighted average of each rule's output. Figure 2.7 shows the architecture for ANFIS. All the nodes in the same layer will perform the same type of functions.

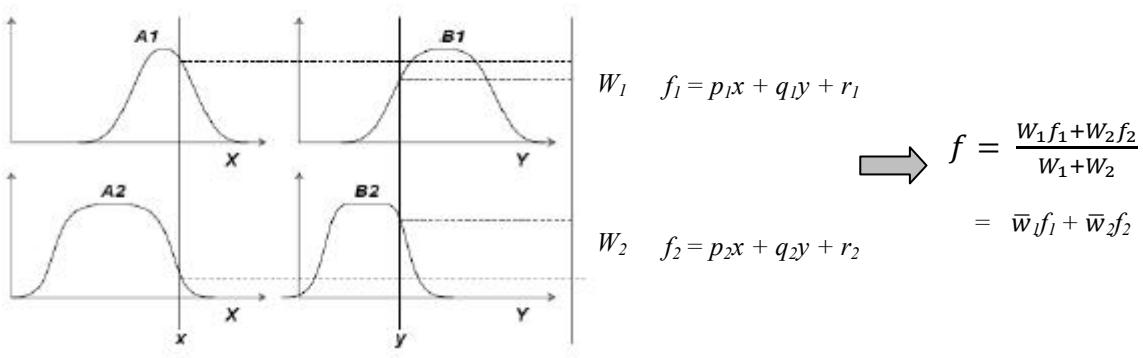


Figure 2.6: First order Takagi-Sugeno fuzzy model

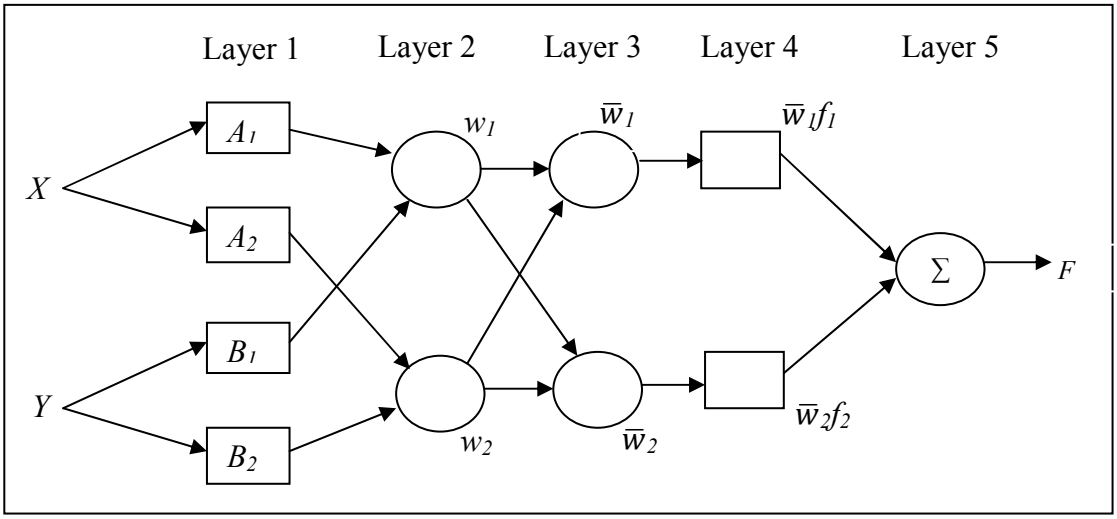


Figure 2.7: An example of ANFIS architecture

ANFIS architecture has five layers, and the functions for each layer are described as below:

Layer 1: Each node in this layer generates a membership function for each input variables.

$$O_i^1 = \mu_{Ai}(x) = \frac{1}{1 + \left|\frac{x - c_i}{a_i}\right|^{2b_i}} \quad , \qquad O_i^1 = \mu_{Bi}(y) = \frac{1}{1 + \left|\frac{y - c_i}{a_i}\right|^{2b_i}} \tag{2.4}$$

x, y - Inputs to node i;

Ai, B_i - Linguistic labels to this node;

$\{a_i, b_i, c_i\}$ - Premise parameters that will change the shapes of membership functions.

Layer 2: Each node in this layer calculates the firing strength, w, for a rule:

$$O_i^2 = w_i = \mu_{Ai}(x)\, \mu_{Bi}(y), \; i = 1, 2. \tag{2.5}$$

Layer 3: Normalises the rule strength:

$$O_i^3 = \overline{w}_i = \frac{w_i}{w_1 + w_2}, \; i = 1, 2. \tag{2.6}$$

Layer 4: Computes the consequents of the rules toward the overall output:

$$O_i^4 = \bar{w}_i f_i = \bar{w}_i (p_i x + q_{i+} y + r_i), \quad (2.7)$$

$\{p_i, q_{i,} r_i\}$ - Consequent parameters.

Layer 5: Computes the overall output as the summation of contribution from each rule:

$$O_i^5 = \text{overall output} = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (2.8)$$

ANFIS uses backpropagation gradient learning to calculate the error signals (squared error) from the output layer backward to the input nodes, and uses the least mean square method to determine the consequents parameters.

2.3.2 Advantages and Limitations of ANFIS

The advantages of ANFIS are, fast convergence due to hybrid learning, it can construct good input membership functions and is suitable for small sample size. Whereas, the limitations of ANFIS are, it has only a single output, no rule-sharing is allowed, each rule has a unity weight and only is allowed for feed forward type networks. Another limitation of ANFIS is that it cannot cater for systems with too many inputs, as overfitting will occur due to too many rules are being generated.

2.4 Statistical Methods

Traditionally, statistical methods are usually used to estimate diagnosis and prognosis. Some of the most common statistical methods are logistic regression, *t*-test, Chi-Square test, ANOVA, linear regression, and correlation.

2.4.1 Logistic Regression

Logistic regression (LR) is the most commonly used statistical method for the prediction of diagnosis and prognosis in medical research. LR is the prediction of a relationship between the response variable y and the input variables x_i. Basically, there are two types of logistic regression, namely, simple logistic regression and multiple logistic regression.

2.4.1.1 Simple Logistic Regression

In the simple logistic regression, there is only 1 response variable y and 1 input variable x. It is a straight-line relationship between the response variable y and the input variable x. The expression of simple logistic regression is given by:

$$y = \alpha + \beta x \tag{2.9}$$

The parameters of α and β are unknown and have to estimate from the data (Ross, 2010).

2.4.1.2 Multiple Logistic Regression

The multiple logistic regression is used when there is more than 1 input variables. The multiple logistic regression model supposes that the response y is related to the input values x_i, $i = 1, 2, ..., k$, through the relationship:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + e \tag{2.10}$$

In which, $\beta_o, \beta_1, ... \beta_k$ are regression parameters which are unknown and must be estimated from the dataset and e is an error random variable that has mean 0 (Ross, 2010).

2.5 Re-sampling Techniques

Re-sampling methods are becoming increasingly popular as they are very robust, simple, and the computing intensive is no longer an issue with high computational power nowadays. Re-sampling is the method used to draw repeated samples from the original data sample. The section below describes four major types of re-sampling techniques, namely, permutation test, cross-validation, Jackknife, and bootstrapping.

2.5.1 Permutation Test

A permutation test is a test procedure in which data are randomly re-assigned so that an exact *p*-value is calculated based on the permutated data. A typical problem involves testing the hypothesis that two or more samples might belong to the same population. The permutation test proceeds as follows (Good, 2004):

(i) Combine the observations from all the samples

(ii) Shuffle them and redistribute them as re-samples of the same sizes as the original samples.

(iii) Record the statistic of interest (the outcome of a statistic test that you are looking for, for example mean, variance, standard deviation, mean square error, *p-value* and so on).

(iv) Repeat (ii) and (iii) for many times

(v) Determine how often the re-sampled statistic of interest is as extreme as the observed value of the same statistic.

2.5.2 Cross-validation

In cross-validation, a sample is randomly divided into two or more subsets and test results are validated by comparing across sub-samples. Cross-validation technique is the world-wide acceptance technique to overcome the problem of small sample size (Braga-

Neto & Dougherty, 2004; Fu et al., 2005; Molinaro et al., 2005). The most common types of cross-validation are *k*-fold cross-validation and leave-one-out cross-validation. In *k*-fold cross-validation, the data is divided into *k* subsets of equal size, train the data for *k* times (the folds), each time leaving out a single subset for validation data. 10-fold cross-validation is the commonly used methods. If *k* equals to the sample size, this is called leave-one-out cross-validation (Sarle, 2010).

Cross-validation suffers from the same weakness as spilt-half reliability when the sample size is small. By dividing the sample into two halves, each analysis is limited by a smaller number of observations (Yu, 2003).

2.5.3 Jackknife

In Jackknife, the same test is repeated by leaving one subject out each time. It is used in statistical inferencing to estimate the bias and standard error in a statistic, when a random sample of observations is used to calculate it, but this process is more complicated than leave-one-out cross-validation (Sarle, 2010).

2.5.4 Bootstrapping

The re-sampling strategy of Bootstrap is more thorough in terms of the magnitude of replication if compared with Jackknife. In Jackknife, the number of re-samples is confined by the number of observations (*n-1*). However, in bootstrap, the original sample could be duplicated as many times as the computing resources allow, and then this expanded sample is treated as a virtual population. Then samples are drawn from this population to verify the estimators. Obviously the "source" for re-sampling in bootstrap could be much larger than that in the other two (Siow-Wee et al., 2010; Yu, 2003).

In addition, unlike cross-validation and Jackknife, the bootstrap employs sampling with replacement. Indeed, sampling with replacement in a bootstrap is more accurate than sampling without replacement in terms of simulating chance. Further, in cross-validation and Jackknife, the *n* in the subsample is smaller than that in the original sample, but in bootstrap every resample has the same number of observations as the original sample. Thus, the bootstrap method has the advantage of modelling the impacts of the actual sample size (Yu, 2003). For estimating generalization error in classification problems, the *0.632*+bootstrap is one of the methods that has the advantage of performing well even when there is severe over-fitting (Sarle, 2010).

2.6 Introduction to Feature Selection

Feature selection is used to select the inputs which are most significant in the modelling process, in order to produce more accurate outputs. The purpose of feature selection is to reduce the number of inputs in the modelling process, but retain the accuracy of the outputs if compared to the full-input model. Thus, this can have a good predictive and less computationally intensive model. This is important especially in medical research where fewer inputs means lower test and diagnosis/prognosis costs.

Feature selection can be classified into three main groups, which are filter, wrapper and embedded methods. Filter methods rank the variables by some chosen criterion, and select the variables with the highest criteria. This method, however, is independent of any algorithm. The examples of filter selection are Pearson's correlation coefficient, linear discriminant analysis (LDA), independent component analysis (ICA) and Relief-F method.

Wrapper methods evaluate the variables in subsets and use the heuristic search methods for an optimal subset. The embedded method is built into a classifier to search for a subset and it is specific to the learning algorithm (Saeys et al., 2007, Song et al., 2005). Genetic algorithm is one example of the wrapper approach.

There are various feature selection techniques. For example, in Song et al. (2005)'s study, a couple of feature selection methods i.e. genetic algorithm, decision tree and correlation coefficient computation are proposed with ANFIS and Adaboost in order to reduce the computational overhead and enhance the system performance. Their results showed that ANFIS with the feature selection system performed better than the ANFIS full-input system with ANFIS-decision tree achieving the highest positive predicted value (97.95%).

Zhang et al. (2007) proposed principal component analysis (PCA) as a feature selection tool for clinical pattern recognition analysis for thyroid cancer and cervical cancer. The PCA was applied on the multiple layer perceptron artificial neural networks (MLP ANN). They proved that the accuracy rate of the MLP ANN based on PCA input selector was improved if compared to the leave-one-out cross-validation method. They claimed that they achieved 100% classification rate with the proposed method.

Sun et al. (2007) proposed a new feature selection algorithm named as I-RELIEF. I-RELIEF combines the advantages of both filter and wrapper methods. It approximates the leave-one-out accuracy of a nearest-neighbour classifier, thus, it addresses the issues of feature correlation and the removal of redundant features. It is used to identify a hybrid signature through the combination of both genetic and clinical markers. The

results showed that the hybrid signature model outperformed other models for breast cancer prognosis.

2.6.1 Genetic Algorithm (GA)

The GA has been discussed in section 2.2.2. In the feature selection problem the main interest is in representing the space of all possible subsets of the given feature set. Then, the simplest form of representation is the binary representation where, each feature in the candidate feature set is considered as a binary gene and each individual consists of a fixed-length binary string representing some subset of the given feature set. Generally, there are seven steps involved in the GA feature selection method, which are:

(i) Solution Encoding

In the feature subset selection problem, a solution is a specific feature subset that can be encoded as a string of *n* binary digits (bits). Each feature is represented by binary digits of *1* or *0*. If a bit is equal to *1*, the feature is selected; consequently, if a bit is equal to *0*, the feature is not selected (Marinakis et al. 2009).

(ii) Initial population

The initial population is generated randomly to select a subset of variables (solutions). For the feature selection problems, the variables selected must be different. If the variables are all different, the subset is included in the initial population. If not, it generates again until an initial population with desired size has been created.

(iii) Fitness function

The fitness function will determine which solutions can precede to the next generation. The fitness function will make the assessment and return a fitness value for a solution. Only the solutions with good/highest fitness value will be accepted.

(iv) Selection

Selection is a process to select the parent chromosome from the population to produce the next generation. There are many types of selection, some examples are listed as below (MathWorks, 2010; NeuroDimension, 2011):

- Roulette wheel - A selection operator in which the probability of being selected for a chromosome is directly proportionate to the fitness. This is the most common type of selection.
- Tournament - A selection operator which uses roulette wheel selection for *N* times in order to produce a tournament subset of chromosomes. The number of *N* is specified by the user.
- Top Percent - A selection operator which randomly selects a chromosome from the top N percent of the population. The number of *N* is specified by the user.
- Best - This operator selects the best chromosome as determined by the fitness.
- Random - This operator selects the chromosome randomly.

(v) Crossover

The crossover function is used to combine two chromosomes and produce an offspring. The most common types of crossover are (MathWorks, 2010; NeuroDimension, 2011):

- One-point - A crossover that randomly selects a point within a chromosome, followed by interchanging the two parent chromosomes to produce two new offsping.

- Two-point - A crossover that randomly selects two points within a chromosome, followed by interchanging the two parent chromosomes to produce two new offsping.
- Scattered - A crossover that creates a random binary vector to select the chromosomes. When the vector is 1, it selects from the first parent, and when the vector is 0, it selects from the second parent, and these combine to form the offspring.
- Arithmetic - A crossover that creates children that are the weighted arithmetic mean of two parents.
- Heuristics - A crossover operator that uses the fitness values of the two parent chromosomes to determine the direction of the search.

(vi) Mutation

Mutation functions specify how the genetic algorithm makes small random changes in the chromosomes to create mutated children. The types of mutation include (MathWorks, 2010; NeuroDimension, 2011):

- Uniform - A mutation that replaces the value of a chosen gene with a uniform random value selected between the user-specified upper and lower bounds for that gene.
- Non-uniform - A mutation that increases the probability so that the amount of the mutation will be close to 0 as the generation number increases.
- Gaussian - A mutation that adds a random Gaussian distribution value to the chosen gene.
- Flip bit - A mutation that inverts the value of the chosen gene (0 becomes 1 and 1 becomes 0).

(vii) Stopping criteria

The stopping criteria is used to determine when to stop the GA when to stop. The algorithm stops as soon as any of the stopping criteria is met. The examples of some common stopping criteria are, number of generations, time limit, and fitness limit.

2.6.2 Pearson's Correlation Coefficient

Pearson's correlation coefficient, *r*, is used to see if the values of two variables are associated. It measures the strength and the direction of a linear relationship between two variables. It was developed by Karl Pearson and is sometimes referred to as Pearson's product moment correlation coefficient. The mathematical formula for computing *r* between two variables of *x* and *y*, with *n* sample size, is denoted as (Rosner, 2006):

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \tag{2.11}$$

The correlation coefficient is a number between -1 and 1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively. A positive correlation indicates a direct relationship, and a negative correlation indicates an inverse relationship between two variables. If there is no relationship between the predicted values and the actual values, the correlation coefficient is 0 or very low. Thus, the higher the correlation coefficient, the better the input variable is.

2.6.3 Relief-F

Relief-F is the extension to the original Relief algorithm, in which it is able to deal with noisy and incomplete datasets and deal with multi-class problems. Figure 2.8 shows the pseudo-code for the Relief-F algorithm. The key idea of Relief is to estimate attributes according to how well their values distinguish among instances that are near to each other. For that purpose, Relief-F will find the *k* nearest neighbours from the same class (nearest hit, H_j) and the *k* nearest neighbours from the different class (nearest miss, $M_j(c)$). It updates the quality estimation $W[A]$ for all attributes *A* depending on their values for R_i, H_j, and $M_j(c)$. If instances R_i and H_i have different values of the attribute *A* then the attribute *A* separates two instances with the same class in which *W[A]* will decrease. Whereas, if instances R_i and *M* have different values of the attribute *A,* then the attribute *A* separates two instances with different class values and $W[A]$ will increase and the contribution of all the hits and all the misses are averaged. The contribution for each class of the misses is weighted with the prior probability of that class $P(C)$ (estimated from the training set).

To ensure the contributions of hits and misses in each step to be in [0,1] and also symmetric, the misses' probability weights is sum to 1. As the class of hits is missing in the sum, each probability weight is divided with factor 1-$P(class(Ri))$ (sum of probabilities for the misses' classes). The process is repeated for *m* times (Kononenko, 1994).

Input: for each training instance a vector of attribute values and the class value
Output: the vector W of estimations of the qualities of attributes

set all weights $W[A]$:= 0.0;
for i := 1 to m do begin
randomly select an instance R_i;
find k nearest hits H_j;
for each class $C \neq class(R_i)$ do
from class C find k nearest misses $M_j(C)$;
for A := 1 to a do
$W[A] := W[A] - \sum_{j=1}^{k} diff(A, R_i, H_i)/(m \cdot k)$

$\sum_{C \neq class(R_i)} [\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} diff(A, R_i, M_j(C))]/(m \cdot k)$;
end;

Figure 2.8: Pseudo-code for the Relief-F algorithm (Kononenko, 1994)

The estimated weight, $W[A]$ for attribute A is an approximation of the following difference of probabilities as shown below (Kononenko, 1994):

W_x = P (different value of A|different class)

−P (different value of A|same class) (2.12)

The rationale is that good attribute should differentiate between instances from different classes and should have the same value for instances from the same class.

2.7 Model Performance Measurements

Several performance measures have been used to evaluate and validate the performance of the proposed model. The measures are accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) curve. The true classification performance of the model is defined as the area under the ROC curve (AUC) (Dom et al., 2008).

A person with positive condition (alive) who is predicted as alive is termed a true positive (TP), whereas a person with positive condition (alive) who is predicted as

negative is termed a false negative (FN). On the other hand, a person with negative condition (dead) who is predicted as positive is termed as false positive (FP), while a person with negative condition (dead) who is predicted as negative is termed as true negative (TN). Table 2.2 shows the above confusion matrix for Oral Cancer Prognosis.

Sensitivity is the true positive conditions divided by all the living patients. This is the probability that a patient will be classified as alive when he is alive.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \text{ x } 100\% \qquad (2.13)$$

Table 2.2: Confusion Matrix for Oral Cancer Prognosis

		Actual conditions	
		Alive (Positive)	Dead (Negative)
Predicted outcomes	Alive (Positive)	True positive (TP)	False positive (FP)
	Dead (Negative)	False negative (FN)	True negative (TN)

The specificity is the true negative conditions divided by all the dead patients. This is the probability that a patient will be classified as dead when he is dead.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \text{ x } 100\% \qquad (2.14)$$

1-specificity is the probability that a patient will be classified as alive when he is dead.

Accuracy is the proportion of true results in the samples, the higher the accuracy, the better the model is.

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TP+TN+FP+FN}} \text{ x } 100\% \quad (2.15)$$

A theoretical, optimal prediction can achieve 100% sensitivity (i.e. predict all people from the surviving group as alive) and 100% specificity (i.e. not predict anyone from the dead group as alive). Positive predictive value is the proportion of patients with positive results (surviving) who are correctly classified as having survived and negative predictive value is the proportion of patients with negative results (dead) who are correctly classified as having dead.

The ROC curve is a plot of sensitivity versus (1 - specificity) for different test results. To generate the ROC curve it is first necessary to determine the sensitivity and specificity for each test result. The *X*-axis (1-specificity) ranges from 0 to 1, or 0% to 100% and is the false positive rate. The *Y*-axis (sensitivity) ranges from 0 to 1, or 0% to 100% and is the true positive rate. The endpoints of the curve will run to these points and an area of the resulting trapezoids can therefore be calculated. The larger the area under the curve the better is the prediction (Adbul-Kareem, 2001). The area calculated under the ROC curve is termed as the area under curve (AUC).

An ideal test will have an AUC of 1 because it achieves both 100% sensitivity and 100% specificity. A good prognostic test is one that has small false positive and false negative rates across a reasonable range of cut off values (Dom, 2008). Table 2.3 shows the formulae for the measures.

Table 2.3: Formulae for measures

Accuracy (%) = $\frac{TN+TP}{TP+TN+FP+FN}$ x 100%
Sensitivity = $\frac{TP}{TP+FN}$ x 100%
Specificity = $\frac{TN}{TN+FP}$ x 100%
Positive Predictive Value (PPV) = $\frac{TP}{TP+FP}$ x 100%
Negative Predictive Value (NPV) = $\frac{TN}{TN+FN}$ x 100%
ROC curve: plot (1 - Specificity) vs. Sensitivity
AUC: Area under ROC curve

2.8 Summary

This literature review gives an overview on the application of AI techniques in cancer research, a brief discussion on the artificial neural network, genetic algorithm, fuzzy logic, Bayesian networks, support vector machine and hybrid AI techniques, and a detail explanation on the neuro-fuzzy techniques, especially on the ANFIS technique. Statistical method of logistic regression has also been discussed. In addition, the major types of re-sampling techniques and feature selection methods are also discussed. Lastly, the performance measurements used in medical research to assess the proposed model are discussed.

CHAPTER 3

ORAL CANCER

3.1 Definition of Oral Cancer

Oral cancer is part of a cancer group called the head and neck cancers, and is defined as an uncontrollable growth of cancerous cells that invades the mouth (oral cavity) and the part of the throat behind the mouth (oropharynx). There are two types of oral cancer, oral cavity cancer and oropharyngeal cancer:

(i) Oral cavity cancer

The oral cavity includes the followings:

- The front two thirds of the tongue.
- The gingiva (gums).
- The buccal mucosa (the lining of the inside of the cheeks).
- The floor (bottom) of the mouth under the tongue.
- The hard palate (the roof of the mouth).
- The retromolar trigone (the small area behind the wisdom teeth).

(Institute, 2009)

(ii) Oropharyngeal cancer

The cancer that starts in the oropharynx, which includes the soft palates (the back of the mouth), the base of the tongue, uvula (The small piece of soft tissue that dangling down from the soft palate at the back of the tongue), and tonsils (one of two small masses of lymphoid tissue located on either side of the throat) (Morrow, 2007).

Around two-thirds of the oral cancers are found in the mouth, while one-third are found in the pharynx (Morrow, 2007). Our study is based on the oral cavity cancer only.

There are two types of tumors: benign and malignant. Benign tumors refer to non-cancerous mass or growth which are not life threatening, because benign tumors do not spread and damage adjacent tissues, structures, and organs. Generally, benign tumors can be removed, and they usually do not grow back. Malignant tumors refer to the cancerous mass or growth which can invade and destroy adjacent tissues and organs inside the body causing death. Malignant tumors often can be removed, but sometimes they metastasize (Institute, 2009).

More than 90% of oral cancers are squamous cell in origin and are called squamous cell carcinomas (SCC). The squamous cells are the thin, flat cells that line the lips and oral cavity. Cancer cells may spread into deeper tissue as the cancer grows. Oral SCC may develop in areas of leukoplakia which are predominantly white patch that cannot be categorized (National Cancer Institute, 2004). A variant of SCC is the verrucous carcinoma. This is a low-grade cancer that rarely metastasises, and has a good prognosis. This type of oral cancer is common among patients that chew tobacco or use snuff. It represents less than 5 percent of all diagnosed oral cancers world widely (Morrow, 2007).

3.2 Oral Cancer Statistics

According to the Malaysian Cancer Statistics, Peninsular Malaysia, 2006, tongue cancer is listed as the sixth top most cancer (4.6%) in Indian male, and mouth cancer is listed as fourth top most cancer (7.3%) in Indian female (Omar, 2006). In Malaysia, Indians are more susceptible to oral cancer and Indian women face the greatest risk, this might

be related to their oral habits of betel quid chewing. Although oral cancer is not listed as the top ten most occurring cancer in Malaysia, the high mortality rate related to this cancer has resulted in the need to improve its survival rate.

There are over 400,000 new oral cancer cases reported worldwide each year. The incident rate of oral cancer differs from region to region. The annual age-adjusted incident rates per 100 000 in several European countries vary from 2.0 (UK, south Thames Region) to 9.4 in France. In the Americas the incident rates vary from 4.4 (Cali, Colombia) to 13.4 in Canada. In Asia, it ranges from 1.6 (Japan) to 13.5 (India). In Australia and New Zealand, it varies from 2.6 (New Zealand - Maori) to 7.5 in South Australia (Sunitha & Gabriel, 2004).

The death rate associated with this cancer is particularly high not because it is hard to discover or diagnose, but because the cancer is routinely discovered late in its development. Oral cancer is particularly dangerous because in its early stages it may not be noticed by the patient, as it can frequently prosper without producing pain or symptoms that might readily be recognized, and because it has a high risk of producing second, primary tumors. This means that a patient who survives a first encounter with the disease, has up to 20 times the risk of developing a second cancer. This heightened risk factor can last for 5 to 10 years after the first occurrence. There are several types of oral cancers, but, around 90% are squamous cell carcinomas (Foundation, 2010).

Oral cancers have an 80 to 90 % survival rate if diagnosed early (Foundation, 2010). Unfortunately at this time, the majority of oral cancer is found as late stage cancers, and this account for the very high death rate of about 50% at five years from diagnosis. According to the statistics from the World Health Organization (WHO), almost two-

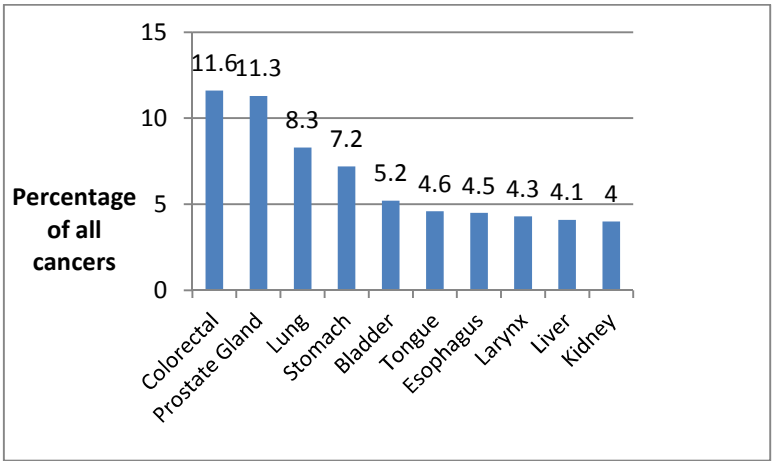
thirds of oral cancers occur in developing countries and it expects a worldwide rise in oral cancer incidence in the next few decades (Oliveira et al., 2008). Thus, it is important to have an accurate survival prediction tool in order to find out the best prognosis methods for individual oral cancer patients.

3.3 Risks Factors of Oral Cancer

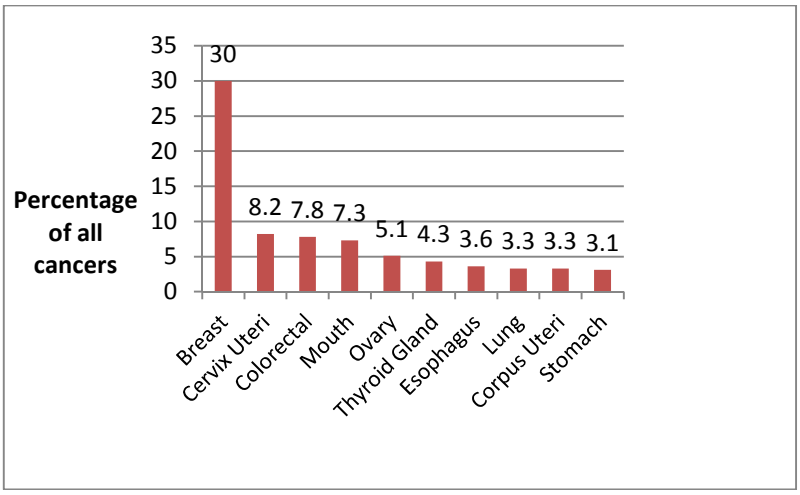
Oral cancer has a relatively low incidence but is potentially very serious if not identified early. Thus, identifying the risk factors associated with oral cancer is very important in the early diagnosis of a patient. There are various factors which have been identified as risks to oral cancer. This include age, gender & ethnicity, smoking, tobacco & betel quid chewing, alcohol consumption, diet, virus infection, specific genes, and oral hygiene (Jefferies & Foulkes, 2001; Mehrotra & Yadav, 2006; Reichart, 2001; Sunitha & Gabriel, 2004).

3.3.1 Age, Gender and Ethnicity

Studies (Oliveira et al., 2008; Chen et al., 2007; Razak et al., 2010) have shown that the incidence of oral cancer increases with age. In Western countries, 98% of oral cancer cases occur in individuals over 40 years of age (Reichart, 2001). From a gender perspective, for decades this has been a cancer which affected 6 men for every woman. That ratio has now become 2 men to each woman (Foundation, 2010). In Malaysia, Indians are more susceptible to oral cancer and Indian women face the greatest risk, this might be related to their oral habits of betel quid chewing. According to Figure 3.1, adapted from the Malaysian Cancer Statistics, Peninsular Malaysia 2006, tongue cancer is listed as the sixth top most cancer (4.6%) in Indian male, and mouth cancer is listed as the fourth top most cancer (7.3%) in Indian female.



(a) Indian Male



(b) Indian Female

Figure 3.1: Ten Most Frequent Cancers in Indians, Peninsular Malaysia 2006
* Modified from Malaysia Cancer Statistics, Peninsular Malaysia 2006

In Malaysia, the oral cancer incident rate is the highest (71.6%) for individual above 50 years old. Tongue cancer has the highest incidence rate when compared to cancers in other parts of the mouth. Table 3.1 shows the oral cancer frequency by age, gender and site for Peninsular Malaysia 2006.

Table 3.1: Oral Cancer Frequency by age, gender and site for Peninsular Malaysia 2006

		Sites											
		Tongue		**Mouth**		**Salivary Glands**		**Lip**		**Others**		**Total**	
Gender	**Age**	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Male	0-14	2	0.3	0	0.0	1	0.2	0	0.0	0	0.0	3	0.5
	15-49	27	4.6	12	2.0	31	5.2	3	0.5	3	0.5	76	12.8
	50-69	63	10.6	28	4.7	24	4.1	2	0.3	6	1.0	123	20.8
	>70	32	5.4	23	3.9	9	1.5	3	0.5	3	0.5	70	11.8
Female	0-14	1	0.2	0	0.0	2	0.3	0	0.0	1	0.2	4	0.7
	15-49	24	4.1	12	2.0	42	7.1	5	0.8	2	0.3	85	14.4
	50-69	43	7.3	72	12.2	25	4.2	9	1.5	6	1.0	155	26.2
	>70	18	3.0	43	7.3	8	1.4	6	1.0	1	0.2	76	12.8
	Total	**210**	**35.5**	**190**	**32.1**	**142**	**24.0**	**28**	**4.7**	**22**	**3.7**	**592**	**100.0**

Almost two-third of oral cancers occurred in developing countries, especially in South East Asia, India, and Taiwan. This geographical variation probably reflects the prevalence of specific environmental influences such as the traditional oral habits in these countries like betel quid chewing and tobacco smoking. Also, oral cancer occurs twice as often in the black population as in whites, and survival statistics for blacks over five years are also poorer at 33%, versus 55% for whites (Foundation, 2010). This might be related to the lifestyle choices among different ethnics.

3.3.2 Tobacco, Smoking, Betel Quid Chewing

Tobacco use account for most oral cancer. About 95% of cases of oral and pharyngeal cancer in the United States have been attributed to smoking (Reichart, 2001). Smoking cigarettes, cigars, or pipes; using chewing tobacco; and dipping snuff are all linked to oral cancer. The most common form of tobacco use is cigarette smoking which demonstrates a very high relative risk for oral cancer. The mortality risk for oral cancer in cigarette smokers is substantially greater than that observed among lifelong "never smokers." Although estimates vary, most studies have reported mortality ratios for smokers versus non-smokers of about 28:5, with several reporting ratios in excess of

10:1. Furthermore, the risk for death from oral cancer is consumption related; the more cigarettes consumed daily and the more years one has smoked, the greater the risk (Foundation, 2010).

There are many different preparations of smokeless tobacco i.e. moist or dry snuff, chewing tobacco and etc. In South and South East Asia, smokeless tobacco encompasses betel quid, and many others like nass, naswar, khaini, mawa, mishri, and gudakhu. In Northern Africa chewing habits of shammah are also prevalent. In India, betel quid chewing and pan masala are popular habits among Indians and this had led to major health problems of oral cancer (Reichart, 2001).

The preliminary results from the Malaysian Oral Cancer Database and Tissue Bank System (MCDTBS) indicated that among 156 oral cancer patients, the risk habits that was most commonly practiced was betel quid chewing (59.9%), followed by smoking (36.1%) and alcohol consumption (35.2%) (Mustafa et al., 2007).

3.3.3 Alcohol Consumption

Alcohol is the second most important risk factor of oral cancer. All three forms of alcohol (beer, hard liquor, and wine) have been associated with oral cancer, although hard liquor and beer have a higher associated risk. Studies that found alcohol use to be a factor for oral carcinogenesis have usually concluded that the level of consumption was important, the more you consume the higher the risk. The risk increased by 6 to 15 times compared to non-smokers and non-drinkers if the person both drinks alcohol and uses tobacco (Reichart, 2001).

3.3.4 Diet

Diet plays an important role in the prevention of oral cancer. Previous studies (Reichart, 2001; Sunitha & Gabriel, 2004) have shown that a healthy diet may protect an individual from cancer. Increasing the intake of fibre-rich food, vegetables, fruits, and vitamin C can help to reduce the risk of cancer. Dietary deficiencies may cause epithelial atrophy, which renders the epithelium vulnerable to the action of carcinogens.

A study done by the MCDTBS in comparing oral cancer patients against healthy, non-cancer patients found that frequent intake of vegetables were higher among those who did not have cancer (83%) as opposed to those who have (70%). The scenario was also the same for fruits consumption where it was found that more non-cancer patients frequently consumed fruits (50.9%) as compared to cancer patients (45%). It was thought that the high antioxidant content in these types of foods is responsible for the reduction in risk (Tan et al., 2005).

3.3.5 Virus Infection

The role of oncogenic viruses in certain human cancers is well known. Viruses are believed to induce cancers by altering the DNA and the chromosomal structures of the cells and by inducing proliferative changes of the cells. The Human papilloma viruses (HPV) types HPV16 and 18 which are well known for their oncogenic potential in cervix cancer, are also present in 80% of oral squamous cell carcinomas (Reichart 2001). In Bouda et al.'s (2000) study, HPV 16, 18 & 33 were detected in oral precancer and cancer, but not in normal oral mucosa. According to Saunders Comprehensive Veterinary Dictionary, precancer is the precancerous condition that tends to become malignant but does not necessarily do so (Blood & Studdert, 2007).

HPV E6 protein is known to bind to and inactivate the p53 tumour suppressor gene, possibly allowing chromosomal instability and subsequent neoplastic growth. HPV-16 has also been shown to produce obviously dysplastic epithelial cells in differentiating tissue cultures, which are otherwise sterile. HPV-31, HPV-33 and HPV-35 have also been associated with oral precancers and cancers. HPVs are found in up to 10% of normal oral mucosa (mucous membrane that covers all structures inside the oral cavity except the teeth), 15-42% of leukoplakias (white patches in the oral cavity), and in 50% of erythroplakias (red patches in the oral cavity) and in 50-100% of oral squamous cell carcinomas. The prognostic significance of HPV presence in oral precancers is yet to be determined by large follow up investigations. Survival from oral carcinoma does not appear to be associated with the presence or lack of HPV (Sunitha & Gabriel, 2004).

3.3.6 Specific Genes

A number of oncogenes and tumor suppressor genes have been identified by previous studies associated with oral cancer (Anantharaman et al., 2007; Chen et al., 2007; Hamid et al., 2008; Jefferies & Foulkes, 2001; Mehrotra & Yadav, 2006; Oliveira et al., 2008; Reichart, 2001). The tumor suppressor genes most frequently altered in carcinomas of the upper aerodigestive tract is the p53 gene, located on chromosome 17p. p53 mutation or over-expression has been demonstrated in 43% - 93% of cases of oral carcinoma cells than in any other human cancer. Its occurrence in oral dysplasia (pre-cancerous condition) and microscopically normal mucosa adjacent to head and neck carcinomas suggest that its alteration is an event, which occurs early in carcinogenesis (Sunitha & Gabriel, 2004). Another possible tumour suppressor gene is the p63. Some studies have suggested a better prognosis for tumours with p63 immunoexpression (Oliveira et al., 2008).

Cheng et al.'s (1999) study demonstrated a significantly elevated risk of disease in patients with GSTM1 and GSTT1 null genotypes (odds ratio of 3.67%; 95% CI: 1.94-6.84). The study by Anantharaman et al. (2007) showed that the GSTM1 null genotype is a risk factor in oral cancer among Indian tobacco habits but GSTT1 null genotype emerged as a protective factor. A study by Marques et al. (2006) suggested that the NAT2 polymorphism, alone or combined with GSTM3, may modulate susceptibility to oral cancer in Rio de Janeiro.

3.4 Clinicopathologic and Genomic Markers

There are two types of markers that can be used for the prognosis of cancer, these are, namely, clinicopathologic markers and genomic markers. Traditionally, clinicopathologic markers are used by the clinicians to determine the best prognosis approach for the individual patient. It is not easy for the most skilful clinician to come out with an accurate prognosis by using clinicopathologic factors alone. Thus, there is a need to use genomic markers or biomarkers to improve the accuracy of the prognosis.

Tumor markers, such as oncogene and tumour suppressor mutations, have been investigated to determine the relationship of such molecular alterations to clinicopathologic outcome. The development of such markers would allow treatment to be more properly tailored to the individual tumour. To date, however, there is no specific marker which has been identified that correlates with a specific cancer or response to treatment.

3.4.1 Clinicopathologic Markers of Oral Cancer

The clinical staging of oral cancer is of paramount importance as it helps the clinician to plan treatment, to evaluate various treatment modalities and to make international

comparisons on various aspects of this disease. The system of staging suggested has three parameters (UICC, 1974): T, the extent of the primary tumour; N, the condition of regional lymph nodes; and M, the absence or presence of distant metastasis. Two more parameters - S, site and P, pathology of tumour have been added subsequently. Site refers to the primary site of the cancer which includes tongue (excluding base of tongue), floor of mouth, upper gingiva, lower gingival, lips, and cheeks. The lesion evolution time, recurrences, and histological classification are also checked.

Beside clinical factors, socio-demographic factors also need to be considered in the prognosis of oral cancer. The social demographic factors include age of the patient at the time of diagnosis, gender, ethnicity, risk habits (smoking, alcohol intake, tobacco and betel quid chewing), and family history of cancer.

The pathological data relates to the results obtained from the laboratory examination and the parameters are pathological staging, number of neck nodes, tumour size and thickness and other post surgical pathologic parameters. The clinical factors, socio-demographic factors and pathological factors are combined to become the clinicopathologic markers of oral cancer.

3.4.2 Genomic Markers of Oral Cancer

The main problem with the TNM system is that it does not take the tumour biology and molecular characteristics into consideration, thus, it may not predict patient outcomes accurately (Oliveira et al., 2008). Cancer occurs through multiple steps, each characterized by the sequential stimulation of additional genetic defects, followed by clonal expansion (Mehrotra & Yadav, 2006). The best way to identify genetic changes to oral cancer is to compare between the progressing and non-progressing oral lesions,

in which the progressing lesions are genetically different from those non-progressing lesions (Mehrotra & Yadav, 2006).

The genetic alterations observed in oral cancer are mainly due to oncogene activation and tumour suppressor gene inactivation, leading to de-regulation of cell proliferation and death and polymorphisms in some carcinogen metabolizing enzyme genes. These genetic alterations include gene amplification and overexpression of oncogenes such as *myc*, *erbB-2*, *Epidermal Growth Factor Receptor* (*EGFR*), *cyclin D1* and mutations, deletions and hypermethylation leading to *p16*, *p53* and *p63* tumour suppressor gene inactivation (Mehrotra & Yadav, 2006). Polymorphisms occur most in the *GSTM1, GSTT1, GSTP1, CYP1A1, NAT1* and *NAT2* genes.

p53 is the most frequently associated marker in the head and neck cancers (Mehrota & Yadav, 2006; Oliveira, 2008). *p53* is called the "Guardian of the genome", having a role in maintaining genomic stability, cell cycle progression, cellular differentiation, DNA repair and apoptosis. Apoptosis is a programmed cell death process, which refers to the death of a cell resulting from a normal series of genetically programmed events, when a cell is no longer needed (Editors, 2010). Due to its high catabolic rate, it is not usually possible to demonstrate *p53* protein in normal tissues using immunohistochemistry procedures, whereas mutated *p53* exhibits a much lower catabolic rate and accumulates in the cells (Mehrota & Yadav, 2006).

In addition, *p63* gene, a homolog (homologous protein) of the *p53* is located in chromosome *3q21-29*, and its amplification has been associated with prognostic outcome in oral cancer (Thurfjell et al., 2005; Muzio et al., 2007). The *p63* gene is highly expressed in the basal (deepest layer of the epidermis) or progenitor (stem cells) layers of many epithelial tissues (the cellular covering of internal and external body

surfaces). *p63* shows remarkable structural similarity to *p53* in the exon/intron organization (NCBI, 2010).

Genetic alterations involving the tumor suppressor genes *p16* and *p53*, are frequently observed in the head and neck tumours. Genetic abnormalities inactivating the *p16* gene might confer cell growth defects, contributing to the tumorigenic process. *p53* is important in maintaining genomic stability, cell cycle progression, cellular differentiation, DNA repair and apoptosis (programmed cell death). The study done by Oliveira et al. (2008) indicated that the *p53* immunoexpression, age, and primary anatomic localization are important survival factors for oral squamous cell carcinoma.

Over expression of the *EGFR* and *Transforming growth Factor* (*TGF*) has been found to be associated with oral squamous cell carcinomas (Mehrotra & Yadav, 2006). The study done by Teri et al. (2002) observed that frequent overexpression of apoptosis regulators *p53*, *bcl-2* and *bax*, was observed in oral cancers and in a subset of oral lesions by immunohistochemistry. This indicated that evasion of apotosis via abnormal expression of *bcl-2, bclxL, MCL-1* and *p53* may contribute to oral cancer pathogenesis (Mehrotra & Yadav, 2006).

3.4.3 Current Research that Used Clinicopathologic and Genomic Markers

Most current researches (Baker and Abdul-Kareem, 2007; Abdul-Kareem et al., 2002; Hassan et al., 2010; Kawazu et al., 2003; Li et al., 2007a; Rao et al. 2011; Saritas et al. 2010; Thongkam et al., 2008; Xu et al., 2005; Zhong et al., 2011) use clinicopathologic markers or genomic markers/microarray for the diagnosis or prognosis of oral cancer. There is very little published work that utilizes both clinicopathologic and genomic markers either for prognosis or diagnosis. However, the results from the works that combined both of the clinicopathologic and genomic markers (Catto et al., 2006;

Exarchos et al., 2011; Futschik et al., 2003; Gevaert et al., 2006; Oliveira et al., 2008; Passaro et al.,2005; Seker et al., 2003; Sun et al., 2007) have confirmed that the prognosis of cancer is more accurate when using both clinicopathologic and genomic markers.

Oliveira et al. (2008) focused on the 5-year overall survival in a group of oral squamous cell carcinoma (OSCC) patients and investigated the effects of age, gender, anatomic localization, tumor evolution time, smoking and alcohol intake, nodal status, tumor recurrences, histologic classification, *p53* and *p63* immunoexpression, human papillomavirus, DNA presence, and treatment on the prognostic outcome. The survival curves were generated using the Kaplan-Meier method, and univariate and multivariate analyses were done using the log rank test and Cox regression. Their results showed that the 5-year overall survival was 28.6% and highlighted the influence of *p53* immunoexpression, age and anatomic localization on OCSS prognosis. Oliveira's statistical method can better predict the outcome of OSCC patients with this specific subset of clinicopathologic variables.

Another research that applied both clinical and genomic data is the application of a learning classifier system into the head and neck squamous cell carcinoma (HNSCC). The system was proposed by Passaro et al. (2005) and further enhanced by Baronti et al. (2007). The system named as Hypothesis testing with Classifier Systems (HCS), is a hybrid adaptive system inspired from learning classifier system, decision trees and statistical hypothesis testing. The algorithm can work with different data types and is robust to missing data. The HNSCC dataset that was used comprised of demographic data (age, gender, smoke and alcohol) and 9 genes involved with carcinogen-metabolizing (*CCND1, NQ01, EPHX1, CYP2A6, CYP2D6T, CYP2E1, NAT1, NAT2,*

GSTP1) and 2 genes in DNA repair systems (*OGG1, XPD*). Passaro et al. applied the proposed algorithm in the HNSCC dataset and the Winconsin Breast Cancer dataset (WBC) and compared their results with Naive Bayes, C4.5, neural network and XCS (Evolution of Holland's Learning Classifier). The results showed that the proposed algorithm outperformed the other algorithms in the HNSCC dataset. On the benchmark WBC dataset, all the tested algorithms had comparable results. Passaro et al.'s long-term goal is to identify the genes that are actually involved in the susceptibility to oral cancer. However, it will be more appropriate if the benchmark dataset is chosen from the same type of cancer.

Dom et al. (2007) proposed a fuzzy regression model for the prediction of oral cancer susceptibility. The prediction tool is based on demographic data (age, gender, ethnicity), risk habits (smoking, alcohol drinking, tobacco chewing) and genetic markers (*GSTM, GSTT1*). The model was tested on a sample of 84 oral cancer patients and 87 control data. The results show that the highest prediction set consists of key markers of chewing habits, ethnic group, age, and alcohol drinking.

Another study that utilised both clinical and genetic markers was carried out by Sun et al. (2006) to improve breast cancer prognosis. Sun et al. used both clinical (tumour grade, angio-invation) and genetic markers (*AL080059, CEGP1 & PRAME*). The dataset used was the van't Veer breast cancer dataset. The dataset consists of 70-gene prognostic signature. A computational algorithm, I-RELIEF, was used to identify a hybrid signature through the combination of both genetic and clinical markers. I-RELIEF is the first feature selection algorithm that utilizes the performance of a nonlinear classifier when searching for informative features and can be implemented efficiently through optimization and numerical analysis techniques, instead of

combinatorial searching. The results showed that the hybrid model outperformed those using the clinical markers or genetic markers alone. This proved that combination markers are more accurate in the prognosis of breast cancer.

3.5 Immunohistochemistry Staining

Immunohistochemistry (IHC) is a method for demonstrating the presence and location of proteins in tissue sections. This is very useful for assessing the progression and treatment of diseases such as cancer. In general, the information gained from IHC combined with microscopy literally provides a "big picture" that can help make sense of data obtained using other methods (Abcam, 2010).

IHC staining is accomplished with antibodies that recognize the target protein. Since antibodies are highly specific, the antibody will bind only to the protein of interest in the tissue section. The antibody-antigen interaction can be visualized using two techniques, namely, chromogenic detection or fluorescent detection. In chromogenic detection, the enzyme will be conjugated to the antibody and cleaves a substrate to produce a coloured precipitate at the location of the protein. Meanwhile, for the fluorescent detection, a fluorophore is conjugated to the antibody and can be visualised using fluorescence microscopy (Abcam, 2010). In this research, we are using the chromogenic detection.

3.6 Management of Cancer

3.6.1 Diagnosis

Cancer management can be classified into diagnosis, prognosis and follow-up. Diagnosis is the process of identifying a cancer from its signs and symptoms, which is done through a series of medical check-up by the physicians, which may include laboratory tests (urine test, blood test), imaging procedures (X-rays, CT scan,

ultrasound), and biopsy. In most cases, physicians need to do a biopsy to make a diagnosis of cancer. Normally, a sample tissue will be taken with a needle/endoscope or through a surgery, and it is sent to the pathologists for examination under a microscope. Upon confirmation of cancer, the patient will undergo the treatment prognosis procedures (Foundation, 2010).

3.6.2 Treatment

Cancer can be treated by using surgery, radiotherapy or chemotherapy. The choice of treatment depends on the type of cancer, stage of the cancer, location and grade of the tumour, as well as patient's age and general health.

(i) Surgery

Surgery is the most common treatment for cancer patients. Normally, surgery is performed on less invasive tumours which tries to preserve the normal oral cavity and function. When a tumour is localized, a surgery may be able to remove the entire tumour, however, if the cancer has metastasized to other sites, complete surgical excision is quite impossible. Surgery can be used in conjunction with radiotherapy or chemotherapy as treatment for the cancer patients (Foundation, 2010).

(ii) Radiotherapy

Radiotherapy is the treatment of cancer using ionizing radiation. Ionizing radiation destroys both healthy and cancer cells in the area being treated by damaging the DNA in those cells, making it impossible to continue growing. However, healthy cells are able to self-repair back to normal function. Radiotherapy is suitable to use for the treatment of localized tumours (oral

cancer) and also to treat those blood-forming and lymphatic cancers (leukimia and lymphoma cancer) (Foundation, 2010).

(iii) Chemotherapy

Chemotherapy is the treatment of cancer by using drugs/chemicals to destroy cells. The drugs/chemicals used in chemotherapy are targeted to all rapidly diving cells, both healthy and cancer cells, but healthy cells usually can self-repair themselves. Chemotherapy is good in treating widespread cancer, which is cancer with more than one location in the body.

3.6.3 Prognosis

Prognosis is a prediction of the outcome of a disease and the survival status of the patient, either in the absence of presence of treatment. Most physicians predict prognosis based on several clinical factors, such as, type of cancer, stage of disease at diagnosis, age of patient at diagnosis, treatment type, general health of the patient and so on.

Normally, a 5-year survival rate will be used to measure the survival rate of the cancer over a 5-year period of time. Survival rate for 5-year means a patient can survive or not after 5 years of diagnosis (Abdul-Kareem, 2001). In this research, the survival rate of up to 3-year was used; this is due to lack of survival information for the 5-year survival rate.

3.6.3.1 Follow Up / Survival Analysis

Cancer survival analysis involves studying the length of time a cancer patient lives after a treatment or after the onset of a disease. It involves studying the time from diagnosis until the event occurs, usually death. The survival time can be measured in years, months, weeks, or days from the beginning of follow-up until an event occurs. While, the event can be death, disease incidence, relapse from remission, recovery or any designated experience of interest that may happen to an individual (Kleninbaum, 1995). It is important to state what the event is and when the period of observation starts and finishes (Clark, 2003).

3.6.3.2 Censored Data

The specific difficulty relating to survival analysis in cancer is that, is arises from the fact that different patients cannot be observed for the same length of time. This may be due to the fact that some patients are diagnosed at the beginning of the period under study; some near the end and others may be diagnosed at any other time of the study period. In survival analysis terminology, patients who are observed until they reach the end point (e.g. death) are called uncensored cases while those who survive beyond the end, but the exact survival time is unknown are called censored cases (Chap, 1997, Kareem, 2001).

There are generally three reasons why censoring may occur:

(1) A person does not experience the event by the time the study ends;

(2) A person is lost to follow-up during the study period;

(3) A person withdraws from the study because of other reasons or experiences a different event that makes further follow-up impossible.

When the actual survival time of a patient is beyond the end of the study time, or lost to follow-up or is withdrawn, the type of censoring is referred to as right-censoring. In Figure 3.2, patients A, C, D and E are examples of right censoring.

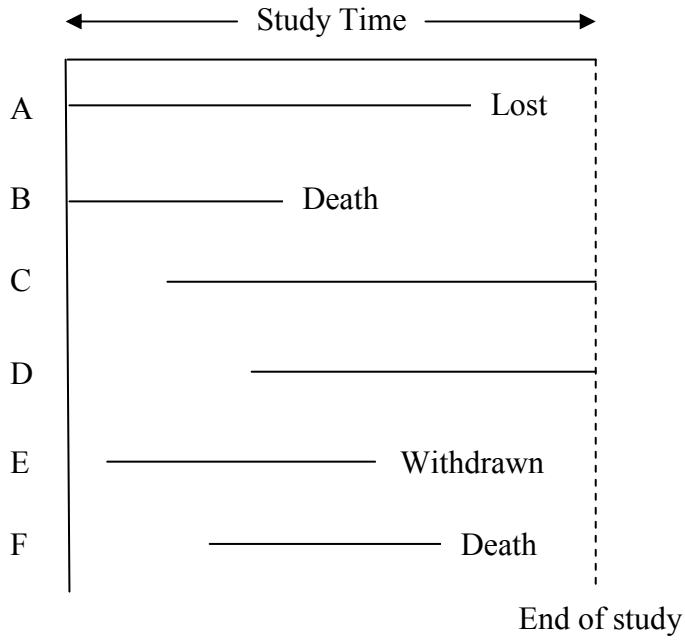


Figure 3.2: Right Censoring

Left censoring is when an event occurs before the study ends. For example, in the study to investigate the time recurrence of a cancer for 3 months after surgery, a patient had a recurrence before 3 months.

Interval censoring occurs when the event happen between two observations. Using the previous example, if the patients are disease free at 3 months and lost to follow-up between 3 to 6 months, they are consider as interval censored.

Most survival data is right-censored. In this research, we used right censored data only.

3.7 Summary

This chapter discusses the basic of oral cancer, oral cancer statistics worldwide and in the local scenario of Malaysia. In addition, the risk factors of oral cancer were also discussed. The clinical and genomic markers of oral cancer were identified and some existing researches that utilized both markers were discussed. This chapter also briefly discussed about the immunohistochemistry staining. Lastly, cancer management procedures namely, diagnosis, prognosis, survival analysis and the data censored which were used in survival analysis were also discussed.

CHAPTER 4

RESEARCH METHODOLOGY

4.1 Introduction

In the previous chapter, various AI techniques, resampling techniques and feature selection algorithms which apply to cancer research have been reviewed, and the promising results from using both clinicopathologic and genomic data were presented. Therefore, in this research, an oral cancer prognosis model will be developed based on both clinicopathologic and genomic data. The methodology adopted in this research will be discussed in the following sections.

Basically, there are four components in this research, which are explained in detail in this chapter. The findings from each component are discussed in the following chapters. The four components are:

(i) Acquisition of oral cancer prognosis data for both clinicopathologic and genomic variables.

(ii) The application of feature selection method on the oral cancer prognosis dataset.

(iii) Development of oral cancer prognostic model using cross-validation ANFIS techniques.

(iv) Model measurements, validation and comparison using other methods.

4.2 Acquisition of Oral Cancer Prognosis Data

Two types of data are needed for developing the oral cancer prognostic model; these are, clinicopathologic data and genomic data. Both types of data are collected from the Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, University of Malaya and are taken from the same patients. The data consist of oral squamous cell

carcinoma (SCC) samples from Malaysia. 31 oral cancer cases have been selected for the purpose of this research based on the completeness and availability of the data (some data are not available due to medical confidentiality problems).

4.2.1 Clinicopathologic Data

Clinicopathologic data was obtained from the Malaysian Oral Cancer Database and Tissue Bank System (MOCDTB) maintained by the OCRCC. The database consists of oral cancer cases collected from participating hospitals from all over Malaysia. From this database, 31 cases were selected based on the completeness of the clinicopathologic data and the availability of the oral cancer tissue samples. The clinicopathologic data consists of information for social-demographic data, primary sites, clinical and pathological Tumour-Node-Metastasis (TNM) stage, nodal status, tumour size, invasion status, types of treatments, survival information and etc. The key clinicopathologic variables used for this research will be identified by the oral pathologists. The details of clinicopathologic data are discussed further in Section 5.2.

4.2.2 Genomic Data

Two genomic variables have been identified through the literature study and discussions with oral cancer experts as important to the prediction of oral cancer survival. The variables are *p53* and *p63*. Immunohistochemistry (IHC) staining is performed on the selected formalin-fixed paraffin embedded oral cancer tissues. The oral cancer tissues are taken from the same 31 cases as in the clinicopathologic data. The procedures of IHC staining, analysis and scoring of the staining results are discussed in Section 5.3.

4.3 Development of Oral Cancer Prognostic Model

Based on the literature review done, it has been identified that ANFIS is suitable for use on small sample dataset and this suits our case which has only 31 samples. Furthermore, ANFIS applies the Takagi-Sugeno model that has a more precise solution, which is very important in medical research. The framework for oral cancer prognostic model is shown in Figure 4.1. Clinicopathologic data from the OCRCC database and genomic data from IHC staining are fed into the model. Basically, there are three execution parts in this research for the oral cancer prognostic model which are wet-lab testing for genomic variables, feature selection methods and ANFIS classification model.

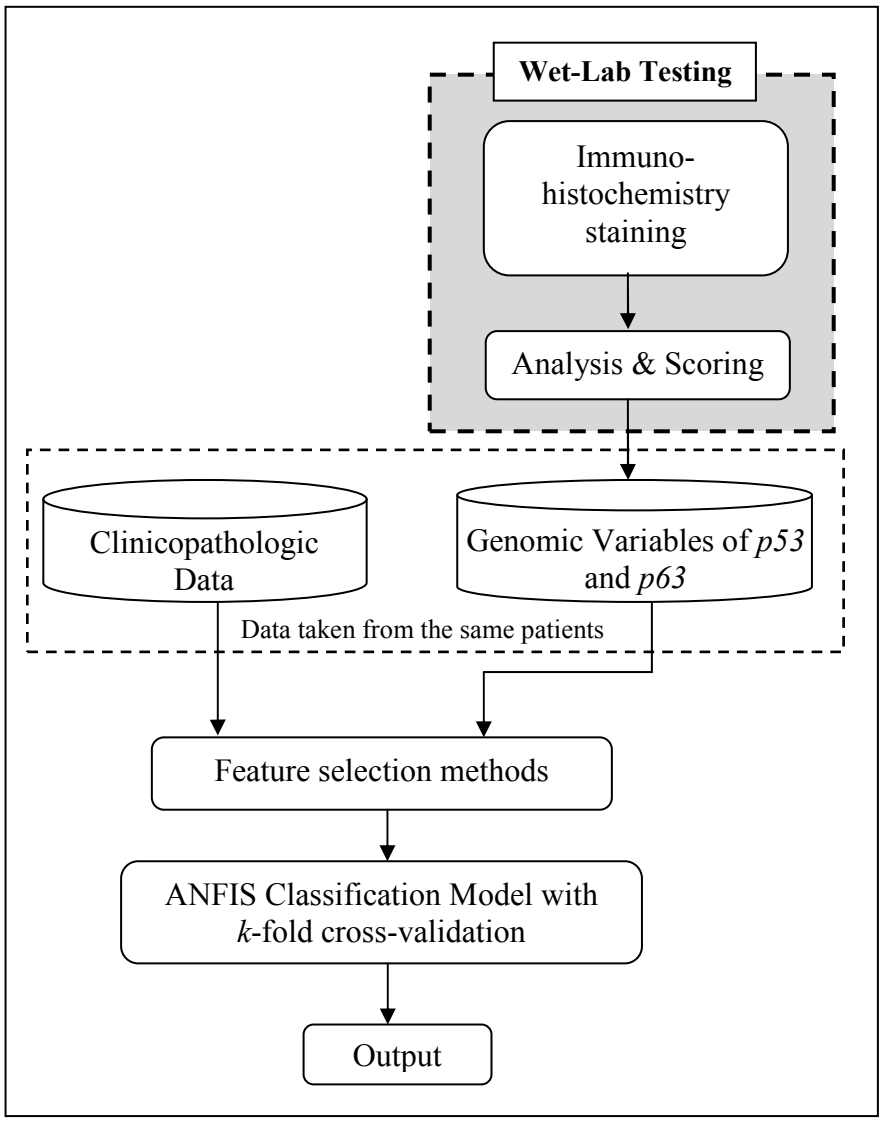


Figure 4.1: Framework for oral cancer prognostic model

4.3.1 Wet-lab Testing for Genomic Variables

In the wet-lab testing part, the immunohistochemistry staining was performed on the oral cancer tissues, which taken from the same 31 oral cancer patients as indicated in the clinicopathologic data, in order to obtain the staining results for *p53* and *p63* genomic data. The detail procedures are described in Chapter 5.

4.3.2 Feature selection methods

Due to the vast numbers of clinicopathologic variables and the small sample size, it is important to have a feature selection algorithm in the proposed model to avoid over-fitting. It is beneficial to limit the number of inputs in a classifier in order to have a good predictive and less computationally intensive model. In medical research, a small input subset means lower test and diagnostic/prognostic costs.

Five feature selection methods are proposed for use in this research to find out the most optimum feature subset for oral cancer prognosis. The aim is to minimize the number of input variables and thus to reduce the time and costs needed for oral cancer prognosis. The proposed feature selection methods are genetic algorithm (GA) as the wrapper approach, correlation coefficient (CC) and Relief-F as the filter approach, and CC-GA and ReliefF-GA as the hybrid approach (filter and wrapper). The methods, procedures and results for feature selection are discussed in Chapter 6 and Chapter 7 respectively.

4.3.3 ANFIS Classification Model

Lastly, the data with *n* selected features are fed into the ANFIS classification model, with *n* varying from three to seven. The final output is the classification accuracy for oral cancer prognosis, which classifies the patients as alive or dead after subsequent

years of diagnosis with the optimum feature of subset. The ANFIS was reviewed in section 2.3.1.

Figure 4.2 shows an example of the ANFIS structure for the 3-input model and Figure 4.3 shows the input membership functions for the same model. In the input layer, the number of input is defined by *n,* with $n = 3, 4, 5, 6, 7$. In the input membership (inputmf) layer, the number of membership function is defined by m_i, with $i = 2, 3, 4$. The rules generated are based on the number of input and the number of input membership functions, and it is represented as $(m_2^{n1} \times m_3^{n2} \times m_4^{n3})$ rules, in which n_1, n_2, and n_3 represent the number of input with m_i membership functions respectively, and $n_1+n_2+n_3=n$. For example, in the ANFIS with 3-input, *x, y, and z,* in which input *x* has 2 membership functions, input *y* has 2 membership functions, and input *z* has 4 membership functions, hence the number of rules generated is $(2^2 \times 3^0 \times 4^1) = 16$ rules (as in Figure 4.2).

The rules generated are the output membership functions which will be computed as the summation of contribution from each rule towards the overall output. The output is the survival condition, either alive or dead after 1-year to 3-year of diagnosis. The output is set as 1 for dead and -1 for alive; the pseudo-code is as below:

```
if output ≥ 0
    then set output = 1, classify as dead
else output < 0,
    then set output = -1, classify as alive
```

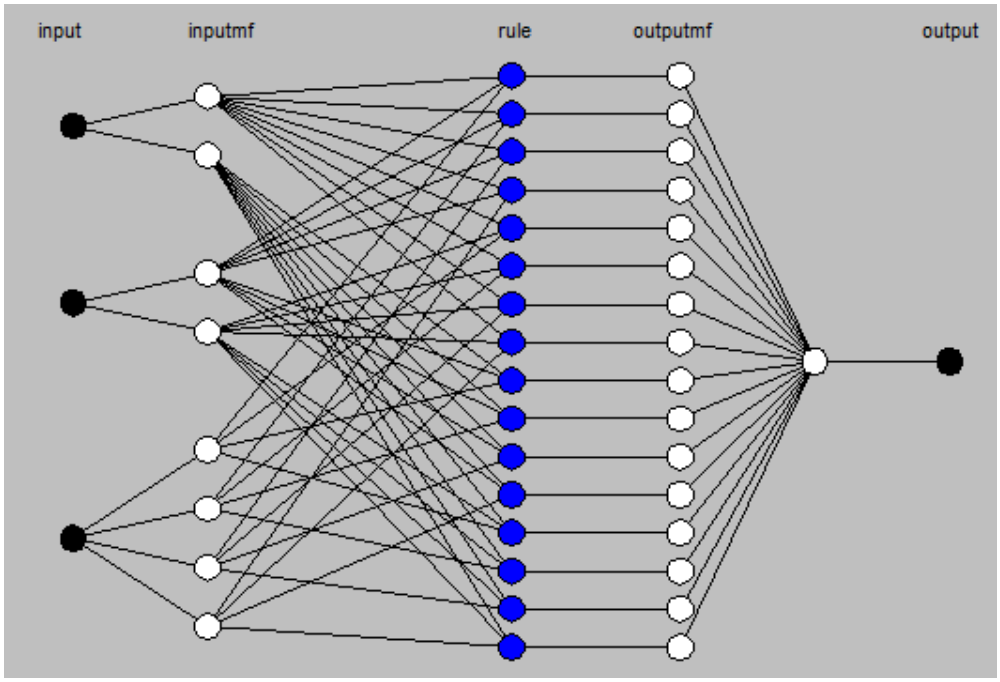


Figure 4.2: ANFIS model structure for a 3-input model

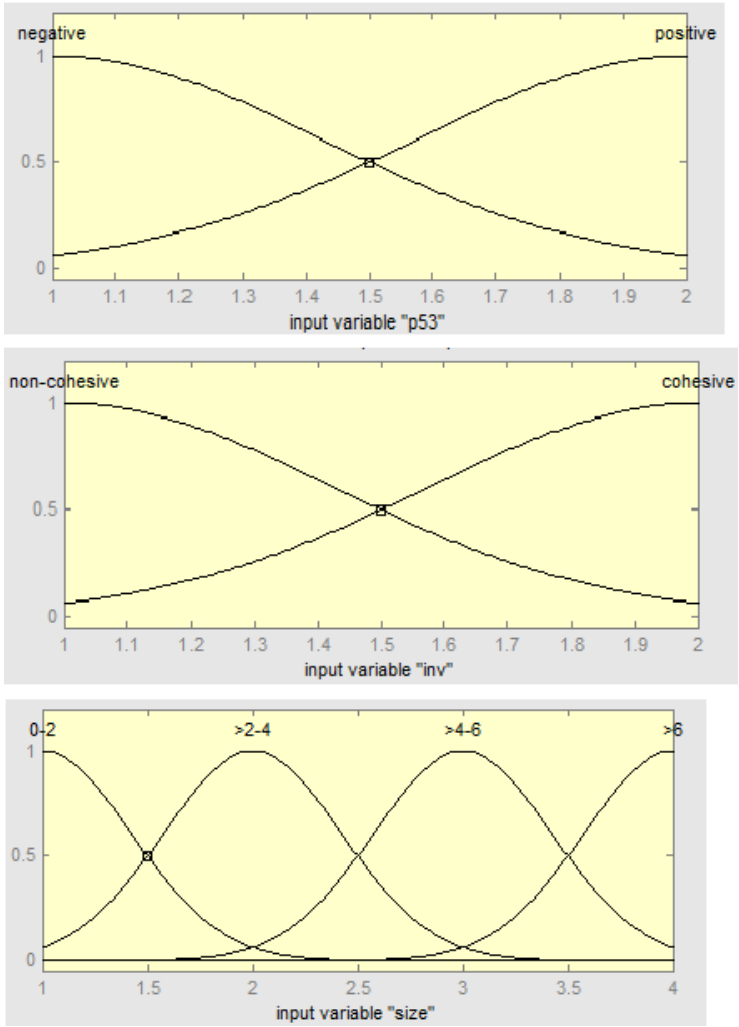


Figure 4.3: An example of membership functions for a 3-input model

Due to the small sample size, a re-sampling technique is needed to create more "versions" of samples in order to get statistically significant results. Bootstrapping and cross-validation (CV) are two methods which are commonly used for small dataset classification problems and the results are promising. In this research, CV is selected as the re-sampling technique. CV is normally used when a validation set is not available or when the data set is too small to split into training and validation sets (Hubert and Engelen, 2004). CV provides unbiased estimation, however, in some researches, CV presents high variance with small samples (Fu et al., 2005).

In this research, *k*-fold cross-validation is used, in which $k = 5$. The 31 samples of oral cancer prognosis data are divided into 5 subsets of equal size and train for 5 times, each time leaving out a sample as validation data.

4.4 Implementation and Testing of the Developed Model on Oral Cancer Prognosis Dataset

The proposed model is tested using real-world oral cancer prognosis dataset from the OCRCC. The samples comprise of patients with oral squamous carcinoma and the dataset consists of clinicopathologic data. The dataset will be combined with the output of the genomic variables obtained from the IHC staining. Both of the clinicopathologic and genomic data are taken from the same set of patients, meaning clinicopathologic data of patient A are combined with the genomic data of patient A. These combined data are tested with the proposed feature selection and ANFIS classification model.

The oral cancer prognosis dataset are divided into two groups, which are Group 1 with clinicopathologic data only and Group 2 with clinicopathologic and genomic data. The objective is to investigate if the classification accuracy from the combination of both

types of data is more accurate. In addition, the proposed model is tested with full-input model of oral cancer prognosis, in which the full-input model is the model with all the 17 variables (15 clinicopathologic variables and 2 genomic variables). This is to investigate if the proposed model with fewer inputs (reduced model) is able to produce compatible results. The prediction ability of the proposed model is measured using the techniques as described in the next section. Besides that, the proposed model is compared with other AI methods such as artificial neural network (ANN) and support vector machine (SVM), and also statistical method of logistic regression (LR).

4.5 Model Performance Measurements, Validation and Comparison

The performance measures used to assess the oral cancer prognostic model are accuracy, sensitivity, specificity, ROC and AUC as discussed in section 2.7. For the purpose of validation and comparison, two commonly used AI methods, which are ANN and SVM are used as the benchmark. As the comparison between the proposed model and the statistical model, a LR model is used. The results and discussions are shown in Chapter 7.

4.6 Summary

This chapter describes the methodology proposed for this research. Two types of variables are used as the input variables, namely, clinicopathologic variables and genomic variables. An oral cancer prognosis model which is based on the ANFIS techniques is proposed. The model is specifically designed for small dataset. The model consists of three parts, first the wet-lab testing for obtaining the genomic data, followed by the feature selection parts where five feature selection methods are proposed, namely, genetic algorithm, correlation coefficient, ReliefF, CC-GA, and ReliefF-GA with the aim to find out the optimum feature subset for oral cancer. The proposed model is tested

using real-world oral cancer prognosis dataset, measured for its performance using the formulae for measures as described in the section 2.7, validated and compared with two other AI methods which are ANN and SVM and a statistical method of LR.

CHAPTER 5

THE ORAL CANCER PROGNOSIS DATASET

5.1 Introduction

Traditionally, the diagnosis and prognosis of cancer have been based on the assessment of clinicopathological features from the patient. This method, however, depends strongly on the expertise and training of the pathologist examining the tissue samples, so the final diagnosis or prognosis may be subjective. Recent studies have demonstrated that genomic markers provide a powerful new approach for determining disease outcome (Muzio et al., 2005; Sun et al., 2007; Oliveira et al., 2008; Gaveart et al., 2006; Cotto et al., 2006; Futschik et al., 2003; Passaro et al., 2005).

In this chapter, methods, preparations and procedures for acquiring oral cancer prognosis data are discussed. First, clinicopathologic data for oral cancer prognosis are discussed as in section 5.2. Next, the genomic markers for oral cancer are identified as in section 5.3. The selection of oral cancer cases and tissue preparations are described in section 5.4, followed by procedures of immunohistochemistry (IHC) staining in section 5.5, and finally results analysis and scoring for IHC staining is discussed in section 5.6.

5.2 Clinicopathologic Data

A total of 31 oral cancer cases were selected from the Malaysian Oral Cancer Database and Tissue Bank System (MOCDTBS) coordinated by the Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, University of Malaya. The selection was based on the completeness of clinicopathologic data, the availability of tissues and the availability of data (some data were not available for use due to medical confidentiality problems). Most of the cases selected for this research were obtained

from the Intensification for Research in Priority Areas (IRPA) project, funded by the Ministry of Science, Technology and the Environment, Malaysia. The MOCDTBS reused from the IRPA project, in which all patients' records were well kept and followed up. Thus, this provides us with the completeness of the data needed for our research.

The selected cases are based on the oral cancer cases seen in Faculty of Dentistry, University of Malaya and Hospital Tunku Ampuan Rahimah, Klang, a Malaysian government hospital, from the year 2003 to 2007. These cases were diagnosed and followed up and the data were recorded in the standardised forms prepared by the MOCDTBS. Later, MOCDTBS transcribed all the data from paper to electronic version and stored in the database. The role of OCRCC is to house the MOCDTBS which maintains and coordinates all the data collected from different hospitals for record and research purposes. All the cases selected are diagnosed as squamous cell carcinomas (SCC).

Basically, three types of data are available for each oral cancer case, namely, social demographic data (risk factors, ethnicity, age, occupation, marital status and others), clinical data (type of lesion, size of lesion, primary site, clinical neck node and etc.), and pathological data (pathological TNM, neck node metastasis, bone invasion, tumour thickness and etc.). Pathological data were obtained from the biopsy reports before and after surgical procedures. In this research, we refer to the clinical and pathological data as clinicopathologic data. Based on the discussions with oral cancer experts, 15 key variables have been identified as important prognostic factors of oral cancer. These variables were commonly used in oral cancer prognosis as discussed in the literature

study of Chapter 2. Table 5.1 lists out the selected 15 key variables and Table 5.2 shows the descriptive statistics for the clinicopathologic variables.

Table 5.1: The Selected 15 Clinicopathologic Variables

	Name	**Description**
1	Age	Age at diagnosis
2	Eth	Ethnicity
3	Gen	Gender
4	Smoke	Smoking habit
5	Drink	Alcohol drinking habit
6	Chew	Quid chewing habit
7	Site	Primary site of tumor
8	Subtype	Subtype and differentiation for SCC
9	Inv	Depth of Invasion front
10	Node	Neck nodes
11	PT	Pathological tumor staging
12	PN	Pathological lymph nodes
13	Stage	Overall stage
14	Size	Size of tumor
15	Treat	Type of treatment

Based on Table 5.2, we can see that there was a much higher case of female at 77.4% while male was 22.6%, of which, the majority of the cases were Indian (74.2%). All 31 cases selected in our research were within the age of 40-80 years old. In terms of risk factors, the most common practice was betel quid chewing (71%), followed by drinking (19.4%) and smoking (16.1%). As for the clinical variables, the most common site of oral cancer was buccal mucosa (41.9%) and most patients had tumour with the size between 4-6cm (35.5%). 64.5% of the cases belonged to the moderately differentiated squamous cell carcinoma while 35.5% belonged to the subtype of well differentiated. Most of the cases had non-cohesive invasive front (83.9%) and the neck node status were almost equal, with 45.2% positive node status and 54.8% negative status. For the pathological TNM stages, 45.2% were in the final stage of T4, 61.3% had nodal status

of N0 (No regional lymph node metastasis). Most of the cases were diagnosed only at stage 4 (67.7%) and the most common treatment was surgery and radiotherapy (54.8%).

Table 5.2: Descriptive statistics of clinicopathologic variables for 31 cases

Variable	Description	No	%
Gender	Male	7	22.6
	Female	24	77.4
Ethnicity	Malay	5	16.1
	Chinese	3	9.7
	Indian	23	74.2
Age	40-50	6	19.4
	>50-60	9	29.0
	>60-70	12	38.7
	>70	4	12.9
Smoke	Yes	5	16.1
	No	23	74.2
	No info	3	9.7
Drink	Yes	6	19.4
	No	22	71.0
	No info	3	9.7
Chew	Yes	22	71.0
	No	6	19.4
	No info	3	9.7
Site	Buccal mucosa	13	41.9
	Tongue	3	9.7
	Floor	3	9.7
	Others	9	29.0
	No info	3	9.7
Size	0-2cm	6	19.4
	>2-4cm	6	19.4
	>4-6cm	11	35.5
	>6cm	6	19.4
	No info	2	6.5

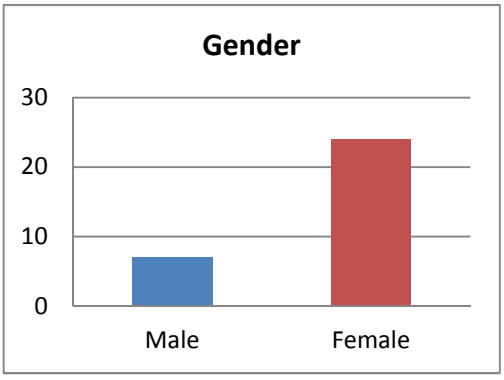
Variable	Description	No	%
Invasion	Cohesive	5	16.1
	Non-cohesive	26	83.9
Nodes	Positive	14	45.2
	Negative	17	54.8
Subtype	Well differentiated	11	35.5
	Moderate differentiated	20	64.5
	Poorly differentiated	0	0.0
PT	T0	0	0.0
	Tis	0	0.0
	T1	4	12.9
	T2	7	22.6
	T3	6	19.4
	T4	14	45.2
PN	N0	19	61.3
	N1	3	9.7
	N2A	1	3.2
	N2B	8	25.8
	N2C	0	0.0
	N3	0	0.0
Stage	Stage I	3	9.7
	Stage II	4	12.9
	Stage III	3	9.7
	Stage IV	21	67.7
Treatment	Surgery only	8	25.8
	Surgery + Radiotherapy	17	54.8
	Surgery + Chemotherapy	4	12.9
	No info	2	6.5

For these 31 cases, based on a 1-year follow-up, 27 had survived and 4 were dead; for a 2-year follow-up, 19 had survived, 10 were dead and 2 were lost to follow-up; while for a 3-year follow-up, 17 had survived, 11 were dead and 3 cases were lost to follow-up, as shown in Table 5.3. Figure 5.1 shows the bar charts for each of the clinicopathologic

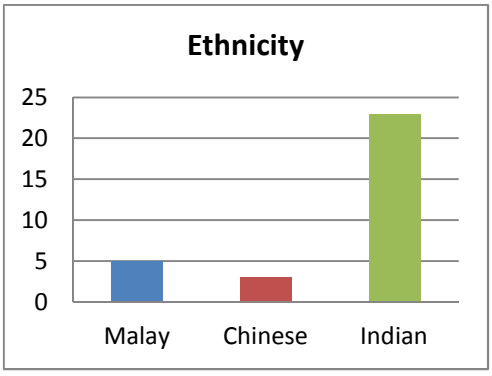
variables. There are some missing data and the methods of handling these data are discussed in Chapter 6.

Table 5.3: 1-year, 2-year and 3-year survival

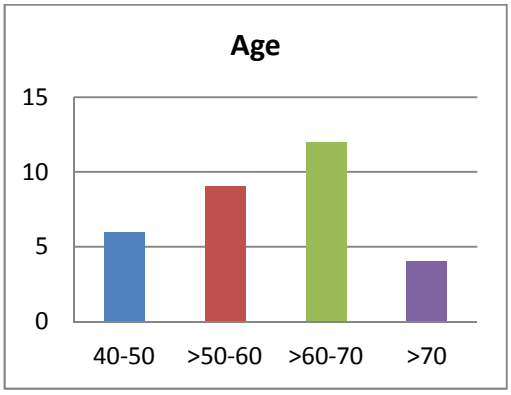
Duration of follow-up	**Survival**	**No**	**%**
1-year	Survive	27	87.1
	Dead	4	12.9
	Lost of follow-up	0	0.0
2-year	Survive	19	61.3
	Dead	10	32.3
	Lost of follow-up	2	6.5
3-year	Survive	17	54.8
	Dead	11	38.7
	Lost of follow-up	3	9.7



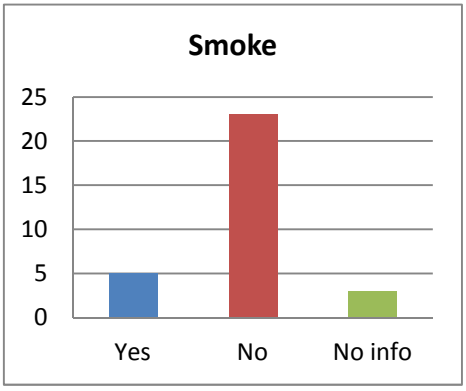
(a)



(b)

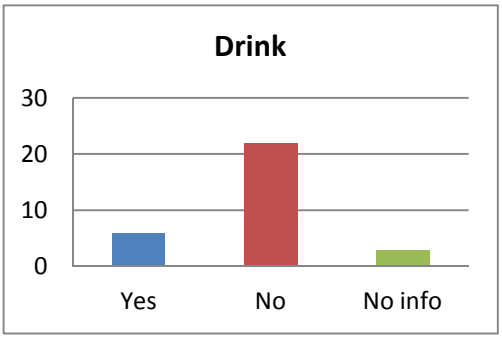


(c)

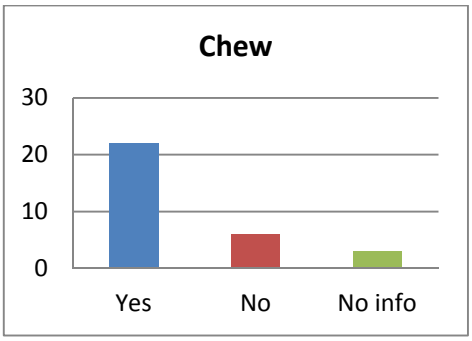


(d)

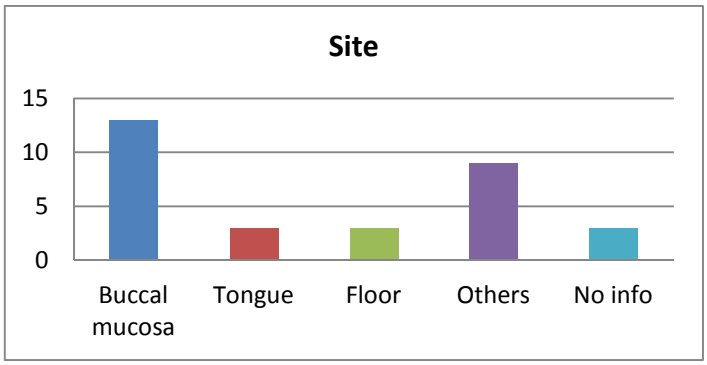
Figure 5.1: Bar Charts for clinicopathologic variables



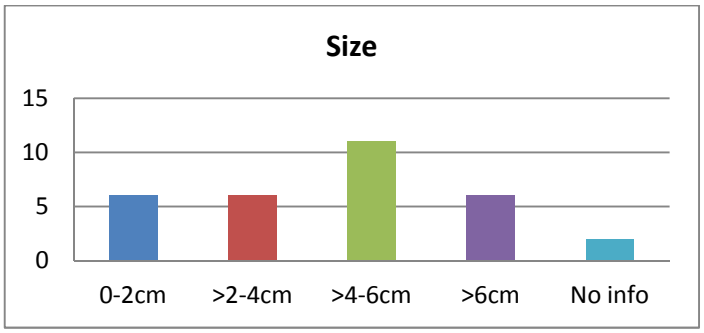
(e)



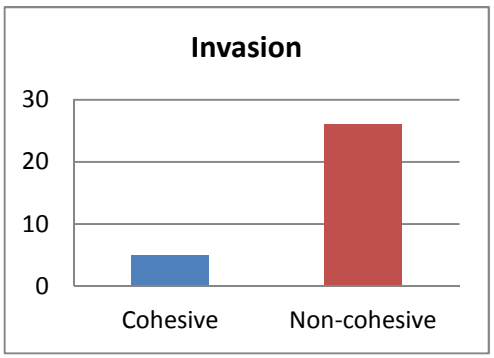
(f)

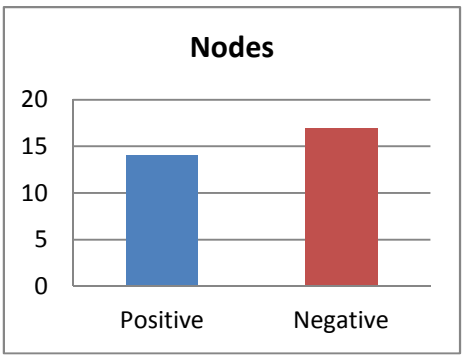


(g)



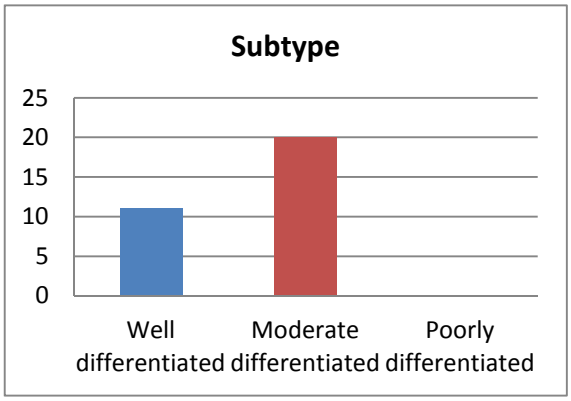
(h)



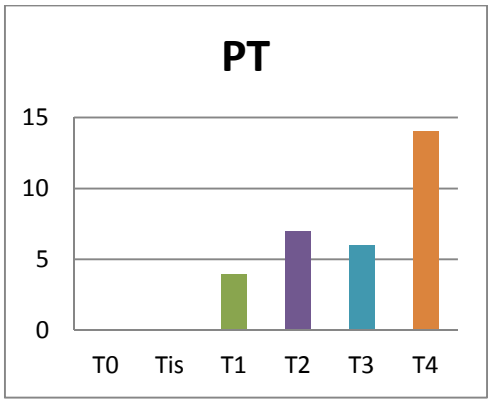


(i) (j)

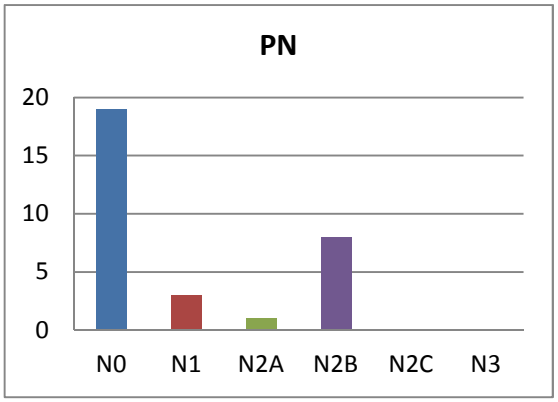
Figure 5.1: Bar Charts for clinicopathologic variables (continued)



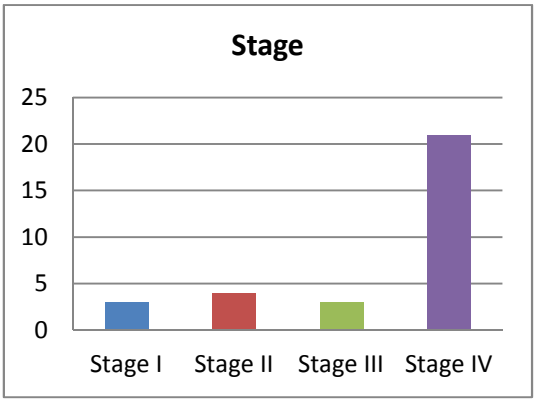
(k)



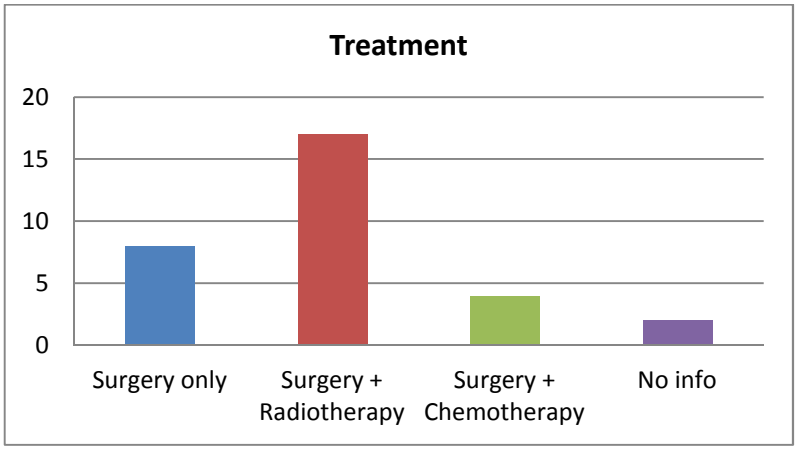
(l)



(m)



(n)



(o)

Figure 5.1: Bar Charts for clinicopathologic variables (continued)

5.3 Identification of Genomic Markers for Oral Cancer

Previous literature studies have revealed a number of genomic factors which are associated with oral cancer prognosis, as discussed in section 3.3.6 and 3.4.2. Genetic polymorphisms in *GSTM1, GSTT1, GSTP1, NAT1* and *NAT2*, (Jefferies & Foulkes, 2001; Anantahraman et al. 2007), tumour suppressor genes such as *p53, p16* and *p63* (Mehrota & Yadav, 2006; Muzio et al. 2005, 2007), overexpression of the epidermal growth factor receptor (Brinkman & Wong 2006) are amongst the most popular genetic factors in oral cancer prognosis.

In this research, due to time, cost and tissue constraints, only two genomic markers are selected, both are tumour suppressor genes, namely *p53* and *p63*. The review for p53 and p63 is in section 3.4.2. The selection is done based on the literature studies and discussions with two oral pathologists, Professor Rosnah Binti Zain and Dr Thomas George from Department of Oral Pathology and Oral Medicine and Periodontology, Faculty of Dentistry, University of Malaya.

5.4 Selection of Oral Cancer Cases and Tissue Preparations

31 oral cancer cases had been selected from the OCRCC database based on the completeness of the clinicopathological data. The cases selected were the same as in the clinicopathologic data. The archival formalin-fixed paraffin embedded tissues were obtained from the Oral Pathology Diagnostic Laboratory, Faculty of Dentistry, University of Malaya. Tissue containing tumour were cored and re-embedded and made into Tissue Macroarray blocks (TMaA). A total of 4-μm-thick sections of the resulting TMaA blocks were cut and placed on the poly-L-lysine-coated glass slides for immunohistochemistry staining. Figure 5.2 shows the macroarray (TMaA) slides prepared for this research.

Figure 5.2: Macroarray (TMaA) slides prepared for this research

5.5 Immunohistochemistry Staining

IHC staining was performed on 4-μm-thick sections cut from the TMaA blocks. The samples were mounted on the glass slides and ready for IHC staining. In this research, Dako REAL^{TM} EnVision^{TM} Detection Kit was used. Figure 5.3 lists the procedures for immuno peroxidise EnVision^{TM} Technique.

The IHC staining was done in the Oral Pathology Diagnostic Laboratory, Faculty of Dentistry, University of Malaya on 10th March 2010 with the help from the laboratory technicians, oral pathologists, and staff from the OCRCC. In total, 15 TMaA slides with 31 oral cancer cases were stained where some cases were repeated. Two types of antibodies were used namely Monoclonal Mouse Anti-Human *p53* protein, clone *318-6-11* for *p53* and Monoclonal Mouse Anti-Human *p63* protein, clone *4A4* for *p63*. Figure 5.4 shows the slides being stained with selected antibody and incubated at room temperature.

1. Deparaffinize and bring sections to water:
    a. Xylene I - 5 min
    b. Xylene II - 4 min
    c. Absolute alcohol - 3 min
    d. 95% alcohol - 3 min
    e. 70% alcohol - 3 min
2. Place the slides in a microwave-resistant plastic staining jar containing Tris-EDTA pH9.0 antigen retrieval buffer (10mM Tris, 1mM EDTA). Make sure the slides are fully covered with buffer.
    a. Operate the microwave oven at 99 degrees.
    b. Let the slides cool at room temperature.
    c. Wash in running water for 5 min.
3. Endogenous peroxidise blocking:
    a. Place the slides in a trough filled with 3% hydrogen peroxide in methanol.
    b. Ensure the slides are fully covered with solution.
    c. Incubate 10 min at room temperature
    d. Wash in 2 baths of TRIS buffer saline (TBS), pH7.6
    e. Drain off excess fluid and wipe the slide carefully
4. Primary antibody** for 30- 40 min in room temperature:
    a. Place the slides on a flat lever surface. Do not allow slides to touch to each other.
    b. Add enough antibody to cover the whole section.
    c. Incubate the sections for 30 – 40 min in room temperature.
    d. Rinse in 2 baths of TBS
    e. Allow each slide to drain off excess fluid and wipe the slides.
5. Second antibody** for 30 min in room temperature:
    a. Place the slides on a flat lever surface. Do not allow slides to touch to each other.
    b. Add enough antibody to cover the whole section.
    c. Incubate the sections for 30 min in room temperature.
    d. Rinse in 2 baths of TBS.
    e. Allow each slide to drain off excess fluid and wipe the slides.
6. Incubate in Diaminobenzidine (DAB) for 5 min:
    a. Apply enough drops of freshly prepared substrate mixture to cover the tissue section.
    b. Incubate for 5 min.
    c. Washing gently with running water for 5 min.
7. Counter stain, dehydrate and coverslip with the followings:
    a. Harris haemotoxylin, 1 min, wash in running water, 3 min,
    b. Acid alcohol, 10 seconds, wash in running water, 3 min,
    c. Potassium acetate, 4 dips, wash in running water,
    d. 95% alcohol for 2 min,
    e. Absolute alcohol, 2 min for 2 times,
    f. Xylene, 2 min for 3 times, and
    g. Mount with Depex

** Monoclonal Mouse Anti-Human *p53* protein, clone *318-6-11* for *p53*
Monoclonal Mouse Anti-Human *p63* protein, clone *4A4* for *p63*

Figure 5.3: Procedures for Immuno Peroxidase EnVision™ Techniques

Figure 5.4: Slides stained with antibody and incubated at room temperature

5.6 Results Analysis and Scoring

The results of staining are analyzed and the images are captured by using an image analyzer system which consists of the Nikon Eclipse E400 Microscope with CFI plan Fluor 40X objective for measurements, QImaging Evolution digital colour cooled camera with 5.0 megapixels, a personal computer (Pentium 4, 2.5Ghz, 2GB RAM) and MediaCybernatics Image Pro Plus version 6.3 image analysis software. Figure 5.5 shows the above mentioned image analyzer system.

Each slide was first examined under the microscope with lower objective, that is, the 4X objective. Cases were considered sufficient for evaluation if there were tumour cells presented in the sections. Next, the slide was divided into 20 grid cells and numbered accordingly from left to right. A simple randomization program was used to generate random numbers. For each case, five tumour representative areas were selected. If the number falls on the non-tumour representative area, the next number (cell) was chosen

until all five areas were selected. Next, five selective areas were examined under the microscope using a higher objective, that is, the 40X objective and the images were captured. The percentage of positive nuclear cells for each area was counted and the average for five areas was calculated. The staining result is considered positive if there is more than 10% positive nuclear stained, in accordance with the practice used in the previous studies (Oliveira et al., 2008; Ziguener et al. 2004). The procedures of the results analysis and scoring as illustrated in Figure 5.6.

Figure 5.5: Image analyzer system

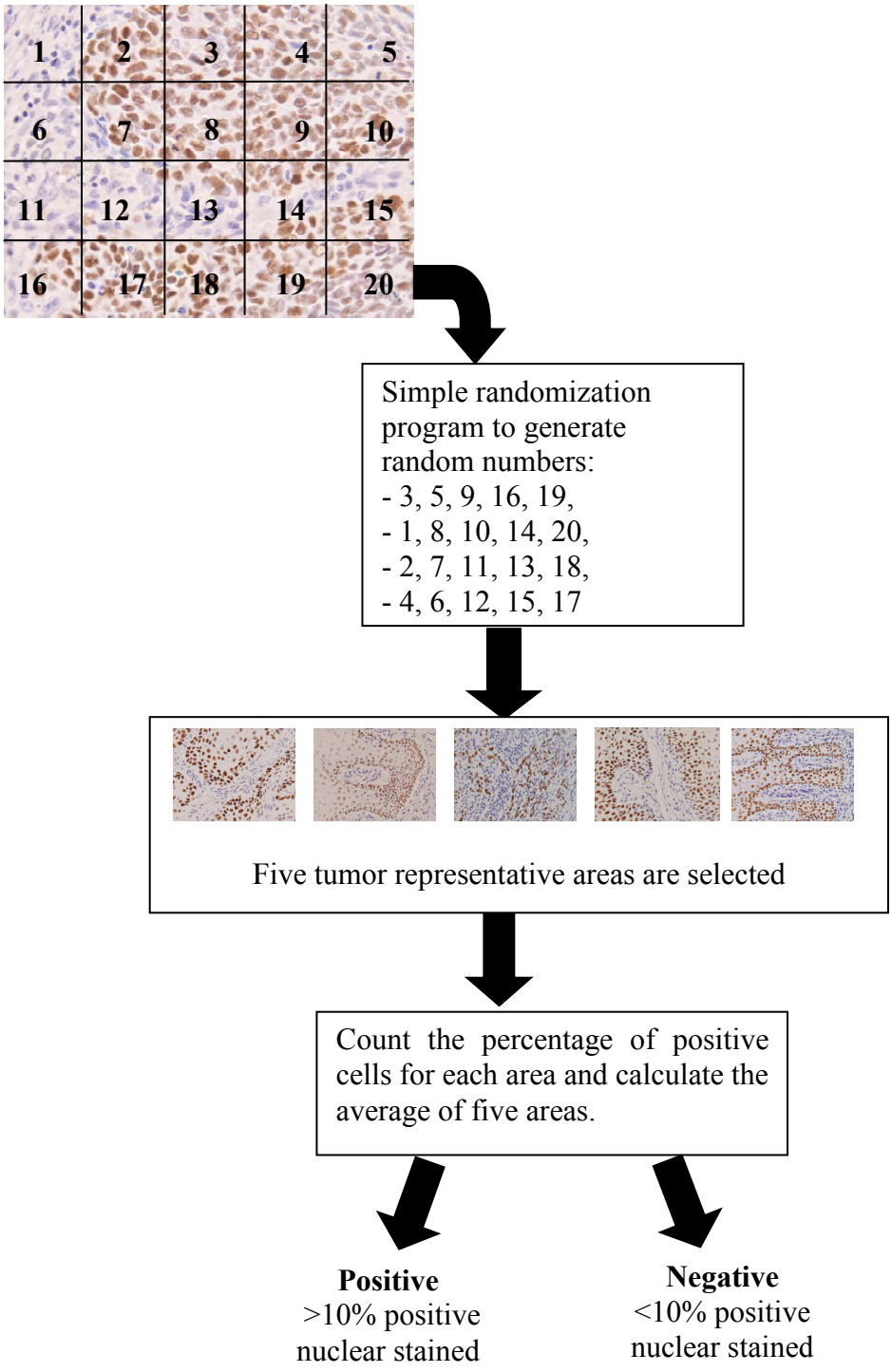


Figure 5.6: Procedures for IHC results analysis and scoring

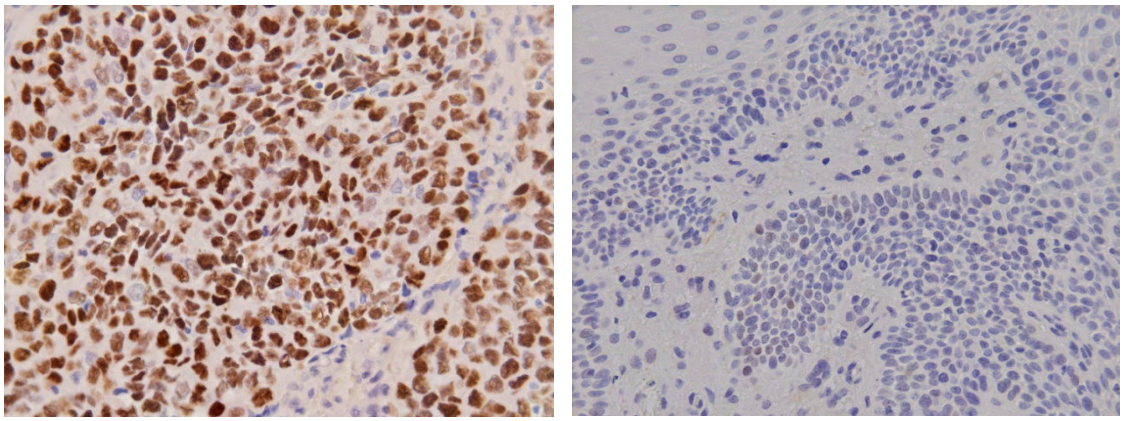
For positively stained slides, nuclear slide will have the nuclear stained in brown colour while negatively stained slide will have the colour of blue/purple. Figure 5.7(a) shows an example of positive stained slide and (b) is an example for negative stained slide. The staining results and descriptive statistics for the results are summarised in Table 5.4 and Table 5.5 respectively.

Table 5.4: Results for IHC staining

No.	**Macroarray Slides**	**p53**	**p63**
1	TMaA 1-1,4-3	neg	neg
2	TMaA 1-1,2-7	neg	pos
3	TMaA 1-2,2-9	pos	neg
4	TMaA 1-2,2-8	neg	pos
5	TMaA 1-3,2-9	pos	pos
6	TMaA 1-8,4-2	neg	pos
7	TMaA 1-4,4-3	pos	pos
8	TMaA 1-5,2-7,2-10,3-2	pos	neg
9	TMaA 1-6,2-7,4-3	neg	pos
10	TMaA 1-7,4-1	neg	pos
11	TMaA 1-7,3-1,4-1	pos	pos
12	TMaA 1-8,3-1,4-2	pos	neg
13	TMaA 1-8,2-8,4-4	neg	pos
14	TMaA 3-1	neg	neg
15	TMaA 3-2	neg	neg
16	TMaA 1-2,3-2,4-1	neg	pos
17	TMaA 3-2	neg	pos
18	TMaA 3-3	neg	neg
19	TMaA 3-3	neg	pos
20	TMaA 3-4	neg	neg
21	TMaA 3-4,4-1	neg	pos
22	TMaA 8-3A	pos	pos
23	TMaA 8-3A	pos	pos
24	TMaA 1-6,4-2,8-5A	neg	pos
25	TMaA 8-5A	neg	pos
26	TMaA 1-4	neg	pos
27	TMaA 1-3	neg	pos
28	TMaA 2-9	neg	neg
29	TMaA 3-1	pos	pos
30	TMaA 8-3a	neg	neg
31	TMaA 8-3a	pos	pos

Table 5.5: Descriptive Statistics for IHC Staining Results

Genomic Markers	Results	Total	Percentage (%)
p53	Positive	11	35.5
	Negative	20	64.5
p63	Positive	20	64.5
	Negative	11	35.5



(a) Positive Stained (b) Negative Stained

Figure 5.7: Example of IHC staining results

5.7 Summary

In this chapter, the selection of oral cancer cases for this research is described and 15 clinicopathologic variables which are used in this research have been identified. Descriptive statistics for the selected clinicopathological variables are analysed and explained in detail. Besides that, two genomic markers have been identified to be used in this research, these are, namely, *p53* and *p63*. 31 oral cancer cases have been selected with the help from oral pathologists and 15 TMaA slides have been prepared. Next, immunohistochemistry staining was performed on the TMaA slides by using the Dako REAL^{TM} EnVision^{TM} Detection Kit and the selected antibodies. Steps and procedures of the staining are clearly defined in this chapter. The results from the staining were analysed and scored by using an image analyzer system and the staining result is

considered positive if more than 10% of the nuclear is positively stained. These two types of markers are combined and served as the inputs for our developed oral cancer prognosis model, which are further explained in the following chapters.

CHAPTER 6

DEVELOPMENT OF ORAL CANCER PROGNOSTIC MODEL

6.1 Introduction

Data pre-processing is an important step in computer modelling, especially for medical modelling where the samples are usually noisy, incomplete and maybe small. In this chapter, data processing methods which are applied on the oral cancer prognosis dataset are discussed, which are namely, data cleansing, data discretization, data transformation and feature selection/input reduction.

There are numerous input variables that can be used for medical modelling, for example demographic variables, risk factors, clinical variables, pathological variables and genomic variables. Each type of variables may consist from ten up to hundreds of inputs. Furthermore, in medical research, it takes time to collect sufficient samples and thus, the sample size is usually small. Hence, there is a need to implement feature selection methods to identify significant variables that are important to the clinical outcomes and to avoid the over-fitting problem. In this research, the purpose of implementing feature selection method is to find an optimal number of features for the small sample of oral cancer prognosis data.

This chapter discusses the development of oral cancer prognostic model. First, the data pre-processing methods and feature selection methods which have been implemented in this research are discussed. The architecture for five feature selection methods, which

are genetic algorithm (GA), Correlation Coefficient (CC), Relief-F algorithm, CC-GA, and ReliefF-GA are discussed and compared. Next, the implementation of ANFIS classification model in the oral cancer prognostic dataset is discussed and the results are shown in Chapter 7.

6.2 Data Pre-processing

Data pre-processing describes the processing methods performed on raw data to transform the data into a format that will be more easily and effectively processed for the purpose of the modelling.

There are different methods used for data pre-processing, which include (Markov, 2011; Pyle, 1999):

(a) Data cleansing - filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

(b) Discretization & sampling - selecting a representative subset from a large population of data.

(c) Transformation – normalising and aggregating

(d) Feature selection/data reducing - extracting specified data that is significant in some particular context.

Data pre-processing is an important process in any data modelling prediction. The accuracy of predictive model depends largely on the quality of the data. Usually, medical dataset is incomplete, noisy and contains a lot of missing data. Therefore, we need to pre-process the selected dataset before we implement the feature selection methods. As described in Section 4.6, we obtained the oral cancer prognosis data from the MOCDTBS database. The original dataset consists of various information for

patients such as personal details, risk factors, oral health related quality of life assessment, clinical and pathological findings, details of surgery, post surgical pathology, details of radiotherapy and chemotherapy and dietary pattern. With the help from oral cancer experts (Zain & George, 2009), we have identified and selected 15 clinicopathological variables and 2 genomic variables, which are commonly used in oral cancer prognosis for the purpose of this research. Based on that, only 31 samples were selected based on the completeness of data and the availability of formalin fixed paraffin embedded tissues, the detail of the oral cancer prognosis dataset is described in Chapter 4.

6.2.1 Data Cleansing

The first step of data pre-processing is the data cleansing. From the selected dataset, we found out that there are 16 missing data and 4 incomplete data. The missing data are mostly on the risk habits. Incomplete data are usually survival information such as lost of follow-up cases in a 3-year study, in which the patient had died and the date had not been recorded.

There are some common methods to deal with missing values, such as case deletion, mean imputation, median imputation and mode imputation. Case deletion which discards the case with missing values in any one feature is not feasible in our research since our sample size is very small. Mean and median imputation is suitable for numeric or continuous data and mode imputation is for nominal (categorical) data. In this research, median imputation is taken for numeric data and mode imputation is taken for nominal data. Median imputation is selected because it is suitable to use when the distribution of the values is skewed and mean imputation is affected by the values of outliers (Acuna, & Rodriguez, 2004).

In this research, we have only three cases of lost to follow-up patients. We assumed that the patients "survived" for 3-year, which means no event occurred within 3-year of follow-up. The number of patients lost to follow-up is small, thus very little bias is likely to have resulted (Clark, 2003).

6.2.2 Data Discretization and Transformation

Next, we categorized and coded the data according to the groups and format that are needed for our predictive modelling. Besides that, three new variables are coded inferred by existing variables, they are status for 1-year survival, 2-year survival and 3-year survival. All the cases in this research are right censored cases; which means the event (death) happened after the follow-up period, and non-censored cases. A sample of oral cancer dataset is listed as in Table 6.1. The categorization results for the dataset are listed in the Appendix (b).

Table 6.1: A Sample of oral cancer dataset

No.	**A-ge**	**Ethni-city**	**Gen-der**	**Smo-ke**	**Dri-nk**	**Ch-ew**	**Site**	**Diagnosis**	**Invasion**	**Nodes**	**PT**	**PN**	**Stage**	**Size**	**Treatment**	**p53**	**p63**	**1-yr survival**	**2-yr survival**	**3-yr survival**
1	41	Malay	M	Yes	No	No	Right cheek	Moderate Differentiated	Non cohesive	Positive	T2	N2b	IV	0 – 2	Surgery + chemotherapy	pos	pos	Survive	Survive	survive
2	79	Indian	F	No	Yes	Yes	Gingiva	Well Differentiated	Non cohesive	Positive	T4	N0	IV	>6	Surgery + radiotherapy	neg	neg	Survive	Survive	survive
3	50	Indian	F	No	No	Yes	Buccal mucosa	Moderately Differentiated	Non cohesive	Positive	T4	N1	IV	>6	Surgery + radiotherapy	neg	pos	Survive	Survive	survive
4	48	Malay	F	No	No	No	Tongue	Moderately Differentiated	Non cohesive	Negative	T3	N0	III	>2 – 4	Surgery	pos	pos	Survive	Survive	survive
5	66	Chinese	M	No	Yes	No	Left side of tongue	Moderately-differentiated	Non cohesive	Negative	T3	N0	III	>2 – 4	Surgery + radiotherapy	pos	pos	Survive	Dead	Dead

6.2.3 Feature selection/Data Reduction

The number of variables in the dataset was considered too many (17 variables) if compared to the sample size (31 cases). Thus, feature selection method is needed to reduce the number of variables and to select only the variables that are significant to oral cancer prognosis. The details of feature selection methods implemented in this research are explained in Section 6.3.

6.3 Feature Selection Methods

Feature selection is used to select the inputs which are most significant in the modelling process, in order to produce more accurate outputs. The purpose of feature selection is to reduce the number of inputs in the modelling process, but retain or increase the accuracy of the outputs as compared to the full-input model. Thus, this can produce a more predictive and cost effective model. This is important especially in medical research where fewer inputs means lower test and diagnosis/prognosis costs.

In this research, the purpose of feature selection is to find an optimal number of features for the small samples of oral cancer prognosis data. Five feature selection methods had been selected and implemented in this research, which were genetic algorithm (GA), Pearson's correlation coefficient, Relief-F, CC-GA, and ReliefF-GA. The above methods were discussed in Section 2.6.

6.3.1 Genetic Algorithm (GA)

In this research, a GA algorithm for the oral cancer prognosis dataset was proposed. The solutions of the GA will form the clinicopathologic or genomic variables that will subsequently be used in the oral cancer prognosis and the output will indicate how well the solutions can predict oral cancer survival.

Before the implementation of GA feature selection method, a simple GA was run to find out the optimal number of inputs (*n-input model*) from the 17 inputs of clinicopathologic and genomic data. The numbers of inputs with the lower error rate are chosen. The error rate for each *n*-input model is shown in Table 6.2 which shows that for Group 1, there are four models with the lowest error rate of 0.3871, which are the 3-input, 4-input, 5-input, and 6-input model. Meanwhile, for Group 2, the model with the lowest error rate is the 3-input model with an error rate of 0.2581. In this case, for comparison purposes, the number of inputs between 3-input to 7-input are chosen. Hence *n* is set as *n* = 3, 4, 5, 6, 7 for the feature selection method.

Table 6.2: Error rate for *n*-input model

	Group 1	Group 2
2-input	0.4193	0.2903
3-input	0.3871	0.2581
4-input	0.3871	0.2903
5-input	0.3871	0.3226
6-input	0.3871	0.3548
7-input	0.4571	0.3548
8-input	0.4839	0.4194
9-input	0.5161	0.4516

The pseudo-code of the proposed GA is listed in Figure 6.1 and is repeated for the *n*-input model. The details of the GA components are discussed below, and the flowchart for GA is illustrated in Figure 6.2.

(i) Solution Encoding

As discussed in section 2.6.1, each feature is represented by binary digits of *0* or *1*. For example, in the oral cancer prognosis dataset, if the solution is a *011001000010000* string of 15 binary digits, it indicates that features 2, 3, 6, and 11 are selected as the feature subset.

```
While selecting initial population with n-input
    Generate initial population randomly without repetition variables
End while
Evaluate the fitness function of each individual using classification error
rate estimated using 10-fold cross-validation
While stopping criteria not exceeded
    Select parents from the population
    Perform crossover operation
    Perform mutation operation
    Evaluate the fitness function using classification error rate
    estimated using 10-fold cross-validation
    Replace the fittest individual
End while
Return the best solution for n-input model
```

Figure 6.1: Pseudo-code for the proposed GA

(ii) Initial population

In this proposed GA feature selection method, if the features are all different, the subset is included in the initial population. If not, it is regenerated until an initial population with the desired size is created.

(iii) Fitness function

The function is used to classify between two groups, which are alive and dead. The error rate of the classification is calculated using a 10-fold cross-validation. The fitness function is the final error rate obtained. The subset of variables with the lowest error rate is being selected.

(iv) Selection, Crossover, Mutation and Stopping Criteria

The selection, crossover, mutation and stopping criteria used in this method are listed in Table 6.3.

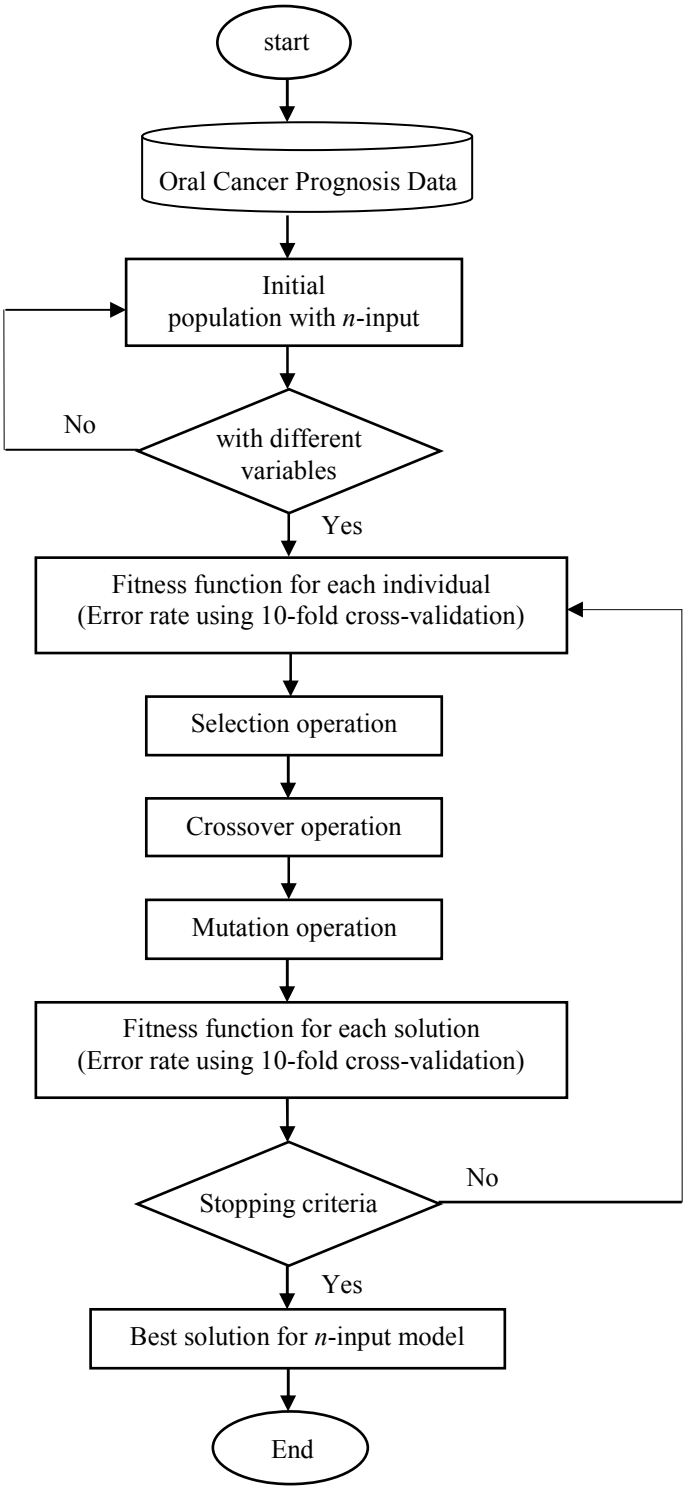


Figure 6.2: Genetic algorithm feature selection flowchart

Table 6.3: Selection, crossover, mutation and stopping criteria for the GA feature selection method

Selection	Roulette wheel
Crossover	Scattered. Crossover fraction = 0.5
Mutation	Uniform. Mutation rate = 0.30
Stopping Criteria	Number of generation = 100 or Time limit = 600s whichever occur first

6.3.2 Pearson's Correlation Coefficient (CC)

In this research, the correlation coefficient, *r*, is calculated and ranked for each of the feature input and the one with the highest *r* is selected. For example, for the 3-input model, the top three inputs with the highest *r* value is selected. This is repeated for the 4-input model to the 7-input model for both Group 1 and Group 2. The flowchart for this method is shown in Figure 6.3.

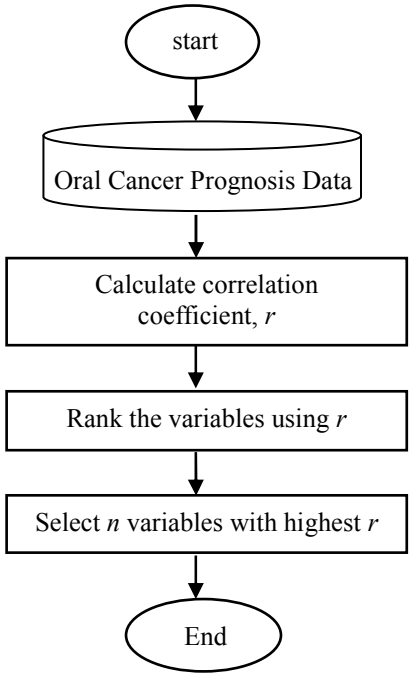


Figure 6.3: Correlation coefficient feature selection flowchart

6.3.3 Relief-F Algorithm

Relief-F belongs to the filter approach. In this method, each feature input is ranked and weighted using the *k*-nearest neighbours classification, in which $k = 1$. The top features with large positive weights are selected for both groups of dataset. Figure 6.4 shows the flowchart for the Relief-F method. The features selected by using this method are listed in the next chapter.

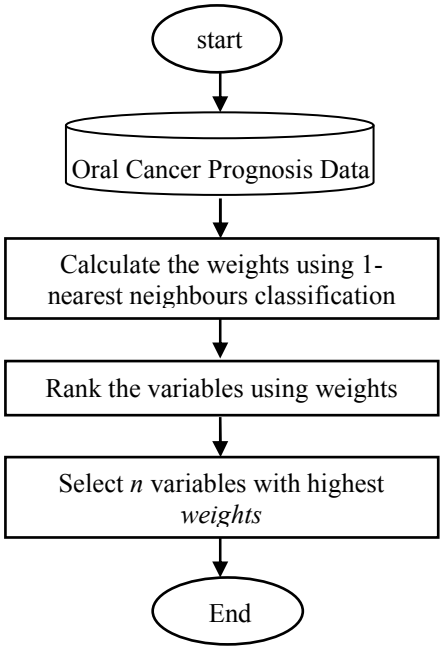


Figure 6.4: Relief-F feature selection flowchart

6.3.4 Correlation Coefficient and Genetic Algorithm (CC-GA)

This is the hybrid feature selection approach which consists of two stages: first, it is a filter approach which calculates the correlation coefficient, *r*, and second, it is a wrapper approach of GA. In the first stage, ten features with the highest *r* are selected and fed into the second stage of the GA approach. The procedures of GA are the same as described in Section 6.3.1. In the second stage, *n*-input is selected for both groups. Figure 6.5 shows the flowchart for the CC-GA method.

6.3.5 Relief-F and Genetic Algorithm (ReliefF-GA)

This hybrid feature selection approach consists of two stages: first, it is a filter approach of Relief-F, and second, it is a wrapper approach of GA. In the first stage, ten features with the highest weights are selected and fed into the second stage of the GA approach. The procedures of GA are the same as described in Section 6.3.1. In the second stage, *n*-input is selected for both groups and the features selected are listed in chapter 7. Figure 6.6 shows the flowchart for the ReliefF-GA method.

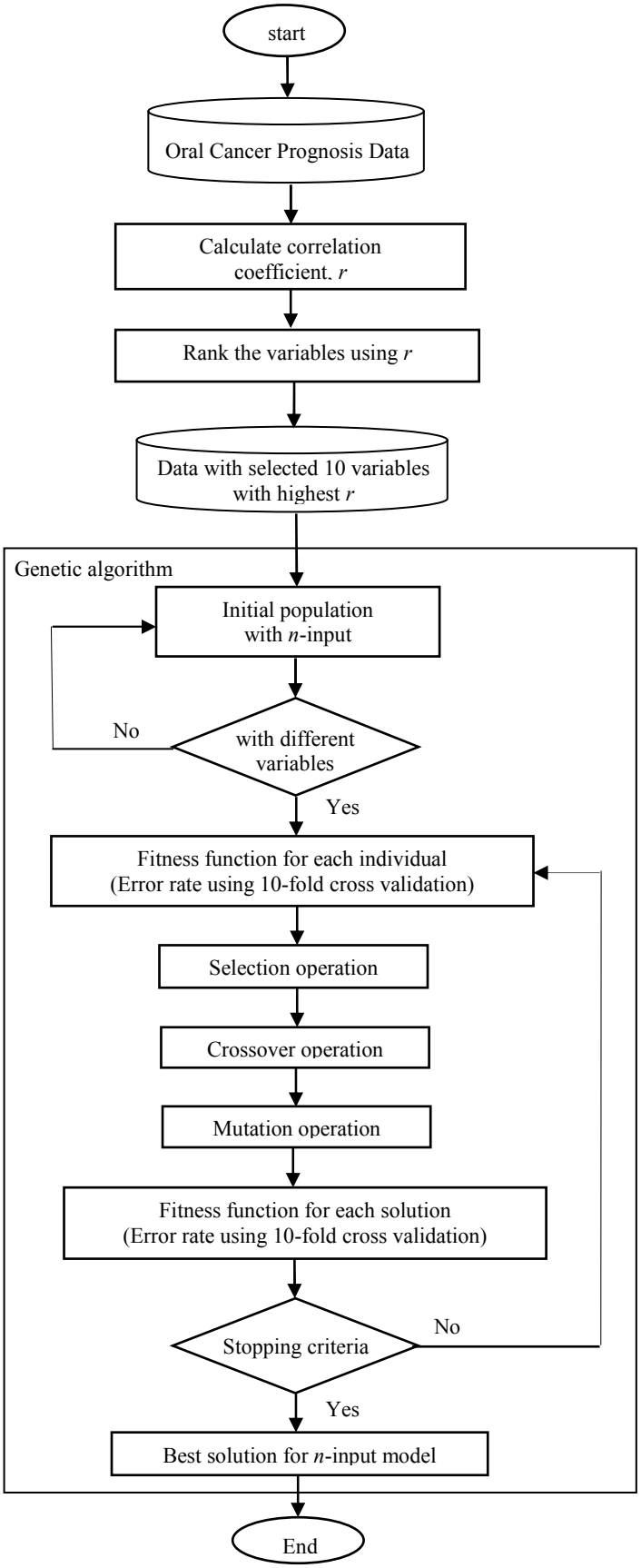


Figure 6.5: CC-GA feature selection flowchart

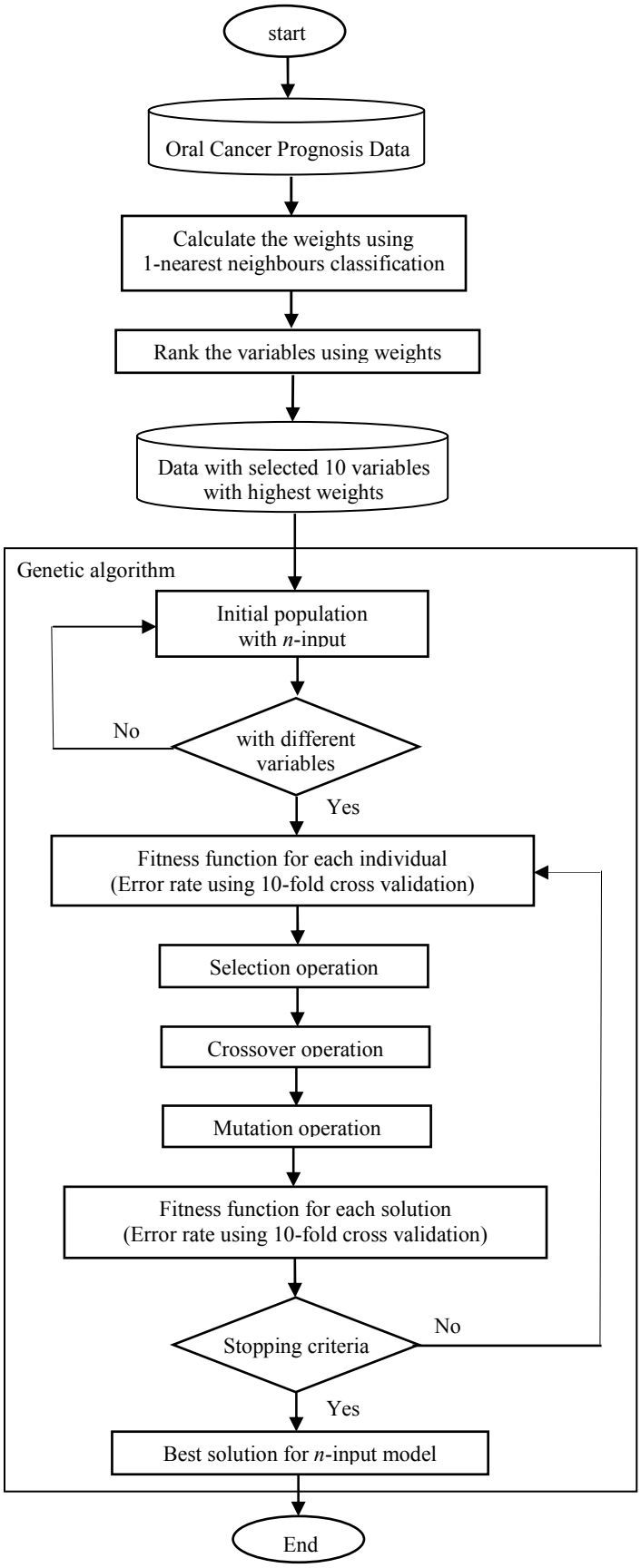


Figure 6.6: ReliefF-GA feature selection flowchart

6.4 ANFIS Classification Model

The proposed ANFIS model is implemented for both Group 1 and Group 2 for the *n*-input models generated from the five proposed feature selection methods. The details for the proposed ANFIS model are described in section 2.3.1. The proposed ANFIS has *n*-inputs, with two to four membership functions for each input, and (m_2^{n1} x m_3^{n2} x m_4^{n3}) rules (refer to section 4.3.3). The type of membership function used is the Gaussian and the number of membership functions for each input variable is shown in Table 6.4. The membership functions for each input variables are shown in Appendix (c). Figure 6.7 shows an example of the membership functions for input variable 'age'. Each ANFIS was run for 5 epochs. A 5-fold cross-validation is implemented on the dataset.

Table 6.4: Membership functions for each input variable

Name	No. of membership function	Name of Membership function
Age	4	40-50, >50-60, >60-70<,>70
Eth	3	Malay, Chinese, Indian
Gen	2	Male, Female
Smoke	2	Yes, No
Drink	2	Yes, No
Chew	2	Yes, No
Site	4	Buccal mucosa, tongue, floor, others
Subtype	3	Well differentiated, moderate differentiated, poorly differentiated
Inv	2	Non-cohesive, cohesive
Node	2	Negative, positive
PT	4	T1, T2, T3, T4
PN	4	N0, N1, N2A, N2B
Stage	4	I, II, III, IV
Size	4	0-2cm, >2-4cm, >4-6cm, >6cm
Treat	3	Surgery only, Surgery+Radiotherapy, Surgery+Chemotherapy
p53	2	Negative, positive
p63	2	Negative, positive

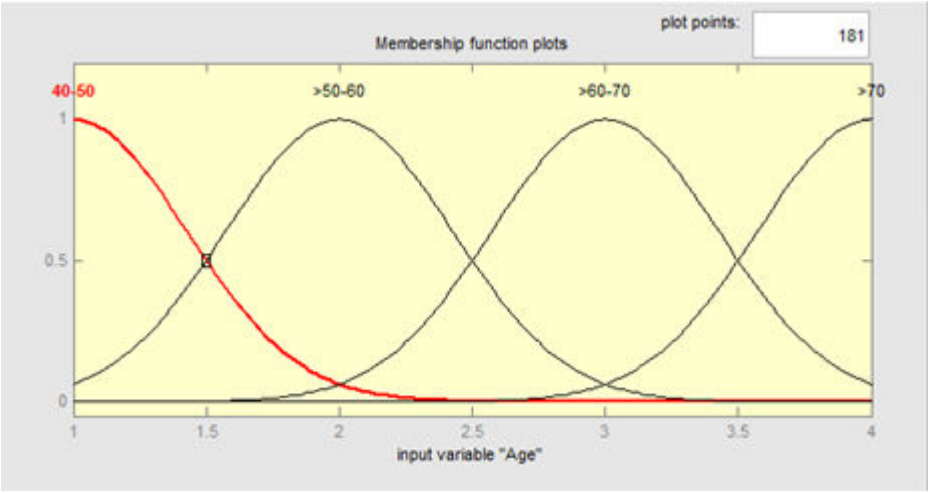


Figure 6.7: Membership functions for input variable "Age"

6.5 Summary

This chapter describes the data pre-processing methods namely data cleansing, data discretization, data transformation and feature extraction/feature selection used for the oral cancer prognosis dataset. Five types of feature selection methods were chosen and compared; these are genetic algorithm (GA), Pearson's correlation coefficient (CC), Relief-F algorithm, CC+GA and ReliefF+GA. The proposed ANFIS classification model is discussed. The membership functions for each input variable are listed in Table 6.4. The results for the feature selection methods and the ANFIS classification model are discussed in Chapter 7.

The sample size for oral cancer prognosis data is very small, thus, the feature selection method is a must to reduce the number of input variables to avoid the over-fitting problem. Feature selection methods are suitable for medical research which has the key features of limited time, cost and tissue samples.

CHAPTER 7

RESULTS AND DISCUSSIONS

7.1 Introduction

In this chapter, we discuss the proposed oral cancer prognostic model to classify whether the patients are alive or dead after 1-3-year of diagnosis.

The oral cancer dataset was divided into 2 groups, which were Group 1 with clinicopathologic variables (15 variables) only and Group 2 with both of the clinicopathologic and genomic variables (17 variables). The feature selection methods as described in Chapter 6 was implemented on both groups in order to select the *n*-input models, with the optimum feature selected.

Next, the proposed ANFIS classification model with 5-fold cross-validation was implemented on the *n*-input models generated from the feature selection methods. For validation purpose, the ANFIS classification results were compared with two AI methods which are artificial neural network (ANN) and support vector machine (SVM), the statistical method of logistic regression (LR), different permutations of *n*-input model and also the full model. Lastly, the ANFIS prognostic model was used to classify 1-year oral cancer prognosis and 2-year oral cancer prognosis.

The three main objectives of this chapter are first, to show the classification results are better in Group 2 (clinicopathologic and genomic variables) if compared to Group 1 (clinicopathologic variables only). Second, to obtain an optimum subset of features for the oral cancer prognosis, and third, to show that the ANFIS classification model is the

optimum tool for oral cancer prognosis if compared to other methods such as ANN, SVM and LR.

7.2 Feature Selection Methods

Before the implementation of the feature selection methods, first, the oral cancer prognosis dataset was divided into two groups; Group 1 consists of clinicopathologic variables only (15 variables) and Group 2 consists of clinicopathologic and genomic variables (17 variables). The feature selection methods were implemented to both groups and the selected features are listed in Table 7.1 and Table 7.2 respectively. In Table 7.2, almost all the feature selection methods included the genomic variable as one of the key features, except for the ReliefF-3-input and ReliefF-4-input.

Table 7.3 listed the number of times a particular feature was selected for each of the feature selection method for Group 1 and Group 2 and Table 7.4 summarised the most selected features for each of the feature selection method for both groups. From Table 7.4, it is noted that for both Group 1 and Group 2, the most selected features for CC, ReliefF and ReliefF-GA are the same as the 3-input model selected for each method respectively.

However, in order to obtain the optimum subset of features, the features selected needed to be tested and validated using classification methods. In the next section, each *n*-input model for both groups were tested with the proposed ANFIS classification system and compared with the artificial neural network (ANN), support vector machine (SVM) and logistic regression (LR) classification methods. Classification accuracy and the area under the Receiver-Operating-Characteristic (ROC) curve for each model were being calculated. In addition, the accuracy of the *n*-input models was compared with the most

selected features from Table 7.4, other combination of *n*-input models and the full model (the model with all 17 variables)

Table 7.1: Feature Subset Selected for Group 1*

Method	**Feature Subset Selected**
GA	
3-input	*Gen,Smo,PN*
4-input	*Dri,Inv,PN,Size*
5-input	*Dri,Node,PT,PN,Size*
6-input	*Age,Gen,Smo,Inv,PT,Size*
7-input	*Age,Eth,Chew,Inv,Node,PN,Size*
CC	
3-input	*Age,Inv,PN*
4-input	*Age,Gen,Inv,PN*
5-input	*Age,Gen,Inv,PN,Size*
6-input	*Age,Gen,Inv,PN,Sta,Size*
7-input	*Age,Gen,Dri,Inv,PN,Sta,Size*
ReliefF	
3-input	*Eth,Dri,Sta*
4-input	*Age,Eth,Dri,Sta*
5-input	*Age,Eth,Dri,Sta,Tre*
6-input	*Age,Eth,Gen,Dri,Sta,Tre*
7-input	*Age,Eth,Gen,Dri,PT,Sta,Tre*
CC-GA	
3-input	*PT,PN,Sta*
4-input	*Dri,Inv,PN,Size*
5-input	*Age,Gen,Inv,PN,Size*
6-input	*Gen,Dri,Node,PT,PN,Sta*
7-input	*Gen,Dri,Chew,Inv,Node,PN,Size*
ReliefF-GA	
3-input	*Gen,Inv,Node*
4-input	*Gen,Dri,Inv,Node*
5-input	*Gen,Dri,Inv,Node,PT*
6-input	*Eth,Gen,Dri,Inv,Node,PT*
7-input	*Age,Eth,Gen,Smo,Dri,Node,Tre*

*Group 1 - clinicopathologic variables only

Table 7.2: Feature Subset Selected for Group 2*

Method	Feature Subset Selected
GA	
3-input	*Inv,Node,p63*
4-input	*Gen,Inv,Size,p53*
5-input	*Age,PT,PN,Size,p53*
6-input	*Age,PT,PN,Size,Tre,p53*
7-input	*Age,Eth,Smo,PT,PN,Size,p53*
CC	
3-input	*Inv,PN,p63*
4-input	*Age,Inv,PN,p63*
5-input	*Age,Gen,Inv,PN,p63*
6-input	*Age,Gen,Inv,PN,Size,p63*
7-input	*Age,Gen,Inv,PN,Size,p53,p63*
ReliefF	
3-input	*Age,Eth,Dri*
4-input	*Age,Eth,Dri,Tre*
5-input	*Age,Eth,Dri,Tre,p53*
6-input	*Age,Eth,Dri,Tre,p53,p63*
7-input	*Age,Eth,Gen,Dri,Tre,p53,p63*
CC-GA	
3-input	*Inv,Node,p63*
4-input	*Gen,Inv,Size,p53*
5-input	*Age,Dri,PN,Size,p53*
6-input	*Gen,Inv,Node,PN,Size,p53*
7-input	*Gen,Dri,Inv,Node,PN,Size,p53*
ReliefF-GA	
3-input	*Dri,Inv,p63*
4-input	*Dri,Inv,Tre,p63*
5-input	*Age,Gen,Smo,Dri,p63*
6-input	*Age,Gen,Smo,Dri,Inv,p63*
7-input	*Age,Eth,Inv,Sta,Tre,p53,p63*

*Group 2 - clinicopathologic and genomic variables

Table 7.3: The Number of Times Feature is Selected

(a) Group 1

Features	GA	CC	ReliefF	CC-GA	ReliefF-GA
Age	2	5	4	1	1
Eth	1	0	5	0	2
Gen	2	4	2	3	5
Smo	2	0	0	0	1
Dri	2	1	5	3	4
Chew	1	0	0	1	0
Site	0	0	0	0	0
Subtype	0	0	0	0	0
Inv	3	5	0	3	4
Node	2	0	0	2	5
PT	2	0	1	2	2
PN	4	5	0	5	0
Sta	0	2	5	2	0
Size	4	3	0	3	0
Tre	0	0	3	0	1

(b) Group 2

Features	GA	CC	ReliefF	CC-GA	ReliefF-GA
Age	3	4	5	1	3
Eth	1	0	5	0	1
Gen	1	3	1	3	2
Smo	1	0	0	0	2
Dri	0	0	5	2	4
Chew	0	0	0	0	0
Site	0	0	0	0	0
Subtype	0	0	0	0	0
Inv	2	5	0	4	4
Node	1	0	0	3	0
PT	3	0	0	0	0
PN	3	5	0	3	0
Sta	0	0	0	0	1
Size	4	2	0	4	0
Tre	1	0	4	0	2
p53	4	1	3	4	1
p63	1	5	2	1	5

Table 7.4: Most selected features for feature selection methods

Feature selection methods	Group 1	Group 2
GA	*Inv, PN, Size*	*Age, PT, PN, Size, p53*
CC	*Age, Inv, PN*	*Inv, PN, p63*
ReliefF	*Eth, Dri, Sta*	*Age, Eth, Dri*
CC-GA	*Gen, Dri, Inv, PN, Size*	*Inv, Size, p53*
ReliefF-GA	*Gen, Dri, Inv, Node*	*Dri, Inv, p63*

7.3 ANFIS Classification Model

The ANFIS model was implemented for both Group 1 and Group 2 for the *n*-input models generated from the five proposed feature selection methods. The details for the proposed ANFIS model were provided in section 2.3.1 and section 6.4. The results obtained from the implementation of these models are given in Table 7.5 to Table 7.8. The results using data for Group 1 are shown in Table 7.5 and 7.6 while results of Group 2 are shown in Table 7.7 and 7.8.

Table 7.5: Classification accuracy for ANFIS in Group 1

Feature selection	**ANFIS**				
Method	3-input	4-input	5-input	6-input	7-input
GA	70.95	67.42	64.76	58.57	57.62
CC	58.10	74.76	51.43	57.62	64.29
ReliefF	61.43	50.59	58.10	64.29	64.29
CC-GA	44.76	67.62	63.81	64.29	57.62
ReliefF-GA	67.14	60.48	67.62	51.90	64.76

Table 7.6: AUC for ANFIS in Group 1

Feature selection	**ANFIS**				
method	3-input	4-input	5-input	6-input	7-input
GA	0.66	0.61	0.63	0.55	0.54
CC	0.53	0.70	0.43	0.50	0.58
ReliefF	0.53	0.50	0.50	0.54	0.54
CC-GA	0.44	0.57	0.55	0.54	0.52
ReliefF-GA	0.55	0.59	0.59	0.47	0.57

Table 7.7: Classification accuracy for ANFIS in Group 2

Feature selection method	**ANFIS**				
	3-input	4-input	5-input	6-input	7-input
GA	74.76	67.62	41.90	58.57	35.71
CC	58.10	58.10	51.90	48.57	61.90
ReliefF	54.29	44.29	48.10	67.14	67.14
CC-GA	74.76	70.48	54.76	61.43	64.29
ReliefF-GA	93.81	93.81	65.71	64.76	68.10

Table 7.8: AUC for ANFIS in Group 2

Feature selection method	**ANFIS**				
	3-input	4-input	5-input	6-input	7-input
GA	0.74	0.70	0.40	0.58	0.36
CC	0.48	0.52	0.48	0.46	0.59
ReliefF	0.47	0.38	0.53	0.62	0.62
CC-GA	0.70	0.71	0.57	0.61	0.65
ReliefF-GA	0.90	0.90	0.63	0.62	0.67

For Group 1, there are two models with the accuracy of more than 70%, these are namely, GA-3-input and CC-4-input model (as shown in Table 7.5). The model with the best accuracy is the CC-4-input with an accuracy of 74.76% and an AUC of 0.70 (shown in Table 7.5 and 7.6). The features selected by this model are *age, gender, invasion,* and *PN* (refer Table 7.1).

For Group 2 (Table 7.7 and 7.8), there are five models with a accuracy above 70%, these are namely, GA-3-input, CC-GA-3-input, CC-GA-4-input ReliefF-GA-3-input and ReliefF-GA-4-input. The best models are ReliefF-GA-3-input and ReliefF-GA-4-input with the accuracy of 93.81% and AUC of 0.90 and the features selected for ReliefF-GA-3-input are *drink*, *invasion,* and *p63* and features selected for ReliefF-GA-4-input are *drink*, *invasion, treatment* and *p63* (refer Table 7.2).

7.4 Other classification models

In this section, the oral cancer dataset are tested using other classification models. Two AI methods which are artificial neural network (ANN) and support vector machine (SVM), and a statistical method which is logistic regression (LR) are selected and tested and the results are discussed and verified in section 7.5.

7.4.1 Artificial Neural Network

The artificial neural network (ANN) that was employed in this research is the feed forward (FF) neural network, which is the most common type of ANN. The FF neural network was trained using the Levenberg-Marquardt algorithm. In this research, one hidden layer with five neurons (achieved the best results) was used in the FF neural network and a 5-fold cross-validation was implemented on the dataset. The average classification accuracy and the area under ROC curve (AUC) for ten runs were taken. The results generated from the neural network experiments are shown in Table 7.9, and 7.10 for Group 1 and Table 7.11 and 7.12 for Group 2 respectively.

Table 7.9: Classification accuracy for feed forward neural network in Group 1

Feature selection method	**Feed Forward Neural Network****				
	3-input	4-input	5-input	6-input	7-input
GA	45.52	52.43	45.05	48.38	45.33
CC	54.48	53.57	51.29	51.29	52.33
ReliefF	51.52	41.62	46.05	46.05	44.10
CC-GA	49.24	49.48	46.67	48.29	50.48
ReliefF-GA	50.24	52.86	56.76	47.00	50.05

** Average for 10 runs

Table 7.10: AUC for feed forward neural network in Group 1

Feature selection method	**Feed Forward Neural Network****				
	3-input	4-input	5-input	6-input	7-input
GA	0.53	0.53	0.47	0.52	0.50
CC	0.61	0.59	0.58	0.51	0.53
ReliefF	0.48	0.47	0.49	0.48	0.48
CC-GA	0.51	0.52	0.49	0.49	0.51
ReliefF-GA	0.55	0.59	0.58	0.51	0.54

** Average for 10 runs

Table 7.11: Classification accuracy for feed forward neural network in Group 2

Feature selection method	**Feed Forward Neural Network****				
	3-input	4-input	5-input	6-input	7-input
GA	45.14	51.48	45.81	46.14	47.71
CC	46.24	49.38	46.14	57.38	55.48
ReliefF	40.62	43.24	47.71	49.48	48.76
CC-GA	49.38	53.90	47.05	44.76	55.19
ReliefF-GA	84.62	73.38	48.00	51.57	45.86

** Average for 10 runs

Table 7.12: AUC for feed forward neural network in Group 2

Feature selection method	**Feed Forward Neural Network****				
	3-input	4-input	5-input	6-input	7-input
GA	0.50	0.55	0.49	0.50	0.51
CC	0.46	0.49	0.50	0.58	0.57
ReliefF	0.48	0.49	0.50	0.51	0.50
CC-GA	0.52	0.60	0.52	0.48	0.57
ReliefF-GA	0.83	0.75	0.52	0.53	0.47

** Average for 10 runs

None of the models from Group 1 achieved an accuracy above 70%. For Group 2, the FF neural network together with ReliefF-GA-3-input model achieved the best result at accuracy of 84.62% and an AUC of 0.83. Figure 7.1 shows the mean square error for the neural network training, testing and validation results. It shows that the best validation mean squared error (1.1822×10^{-15}) was obtained after 4 epochs of training. Figure 7.2 shows the training regression for training, test and validation data for ReliefF-GA-3-input model. For a perfect fit, the data should fall along a 45 degree line,

where the network outputs are equal to the targets, with R = 1. For this experiment, the fit is reasonably good for all data sets, with R for training set = 0.86723, R for validation set = 1 and R for testing set = 0.99558. The R for all data sets is 0.91952.

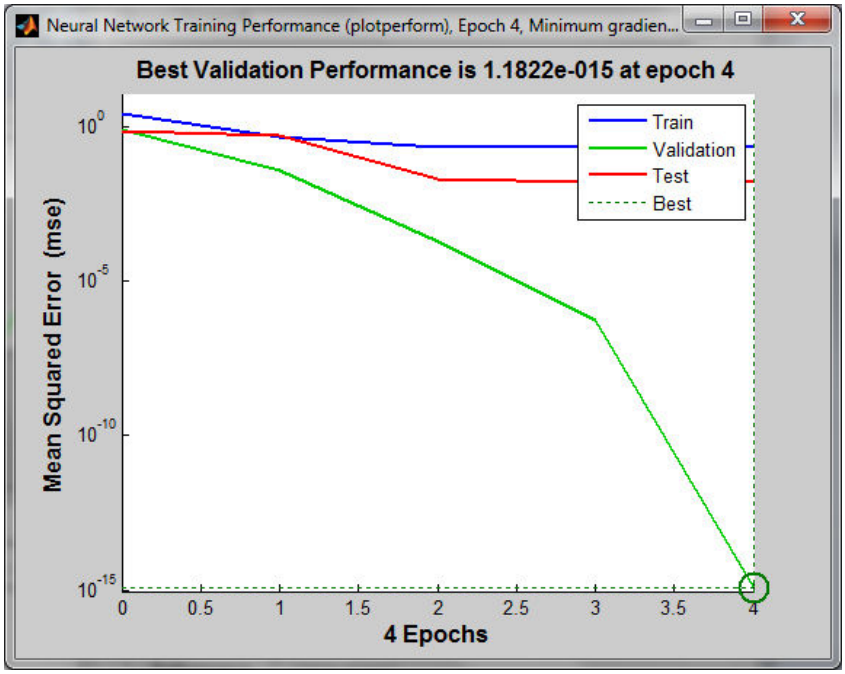


Figure 7.1: Mean Squared Error for ReliefF-GA-3-input model

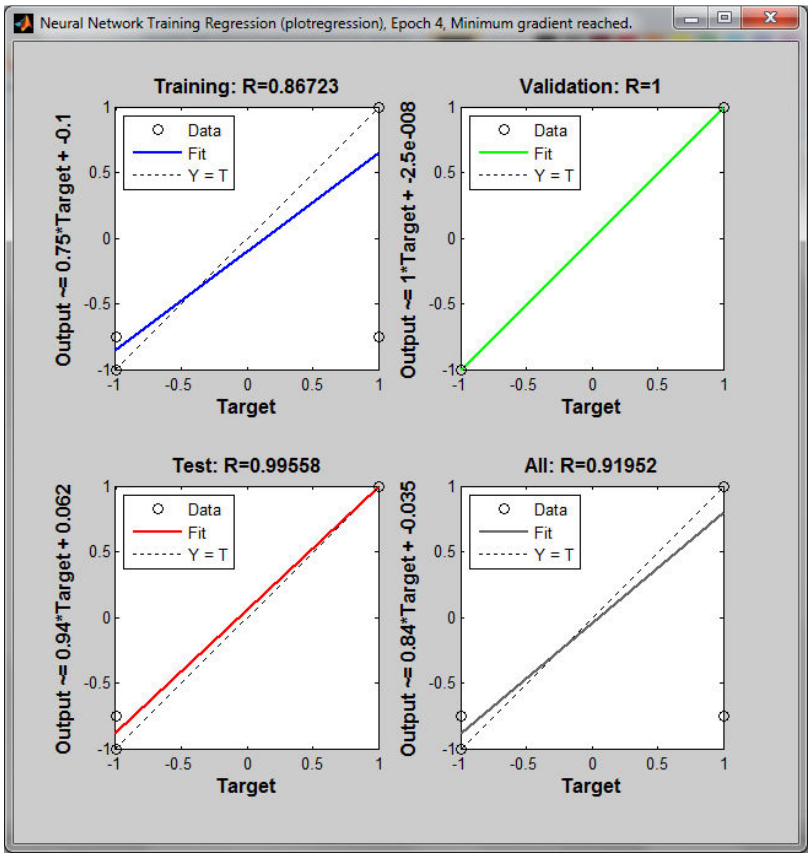


Figure 7.2: Training regression for ReliefF-GA-3-input model

7.4.2 Support Vector Machine

The support vector machine (SVM) tool that was used is the LIBSVM. LIBSVM is a library for support vector machines and it is one of the most widely used SVM software (Chih-Chung, 2011). A 5-fold cross-validation was implemented on the dataset as well. The results are tabulated in Table 7.13 to Table 7.16 respectively.

Table 7.13: Classification accuracy for SVM in Group 1

Feature selection method	**Support Vector Machines (SVM)**				
	3-input	4-input	5-input	6-input	7-input
GA	60.95	61.43	58.10	58.10	61.43
CC	60.95	60.95	58.10	51.43	51.43
ReliefF	54.29	50.95	51.43	48.10	50.95
CC-GA	63.81	61.43	58.10	58.10	58.10
ReliefF-GA	64.29	64.29	64.29	64.29	54.76

Table 7.14: AUC for SVM in Group 1

Feature selection method	**Support Vector Machines (SVM)**				
	3-input	4-input	5-input	6-input	7-input
GA	0.53	0.51	0.48	0.46	0.49
CC	0.53	0.53	0.46	0.41	0.41
ReliefF	0.44	0.42	0.42	0.40	0.45
CC-GA	0.55	0.51	0.46	0.48	0.49
ReliefF-GA	0.50	0.50	0.50	0.50	0.46

Table 7.15: Classification accuracy for SVM in Group 2

Feature selection method	**Support Vector Machines (SVM)**				
	3-input	4-input	5-input	6-input	7-input
GA	74.76	54.76	70.95	60.95	50.95
CC	64.76	64.76	64.76	67.62	67.62
ReliefF	54.29	54.29	44.29	48.10	34.76
CC-GA	74.76	54.76	61.43	58.10	61.43
ReliefF-GA	74.76	71.43	74.76	74.43	54.76

Table 7.16: AUC for SVM in Group 2

Feature selection method	**Support Vector Machines (SVM)**				
	3-input	4-input	5-input	6-input	7-input
GA	0.70	0.51	0.65	0.55	0.42
CC	0.55	0.55	0.55	0.56	0.62
ReliefF	0.44	0.44	0.36	0.46	0.28
CC-GA	0.70	0.51	0.50	0.54	0.57
ReliefF-GA	0.70	0.68	0.70	0.66	0.53

Table 7.13 to Table 7.16 show that the classification results generated from SVM are generally better in Group 2 when compared to Group 1 with some exceptions. None of the model from Group 1 could achieve an accuracy above 70%. Whereas, there are seven models from Group 2 with an accuracy of above 70%, which are the GA-3-input, GA-5-input, CC-GA-3-input, ReliefF-GA-3-input, ReliefF-GA-4-input, ReliefF-GA-5-input and ReliefF-GA-6-input. The best accuracy in Group 2 is obtained by the GA-3-input, CC-GA-3-input, ReliefF-GA-3-input, and ReliefF-GA-5-input with an accuracy of 74.76% and an AUC of 0.70.

7.4.3 Logistic Regression

Logistic regression (LR) is selected as the benchmark test for the statistical method and the results are compared with the AI methods as discussed earlier. A 5-fold cross-validation was implemented in the dataset and the results are tabulated in Table 7.17 to Table 7.20 respectively.

Table 7.17: Classification accuracy for logistic regression in Group 1

Feature selection method	**Logistic Regression**				
	3-input	4-input	5-input	6-input	7-input
GA	64.29	67.62	64.76	68.10	64.29
CC	64.29	60.48	67.62	67.62	64.29
ReliefF	50.59	50.59	48.10	41.43	44.29
CC-GA	67.62	67.62	61.43	70.95	64.76
ReliefF-GA	54.29	51.43	61.43	47.62	48.10

Table 7.18: AUC for logistic regression in Group 1

Feature selection method	**Logistic Regression**				
	3-input	4-input	5-input	6-input	7-input
GA	0.56	0.60	0.55	0.64	0.60
CC	0.56	0.57	0.61	0.61	0.58
ReliefF	0.44	0.44	0.39	0.34	0.39
CC-GA	0.57	0.60	0.51	0.73	0.67
ReliefF-GA	0.54	0.52	0.62	0.55	0.51

Table 7.19: Classification accuracy for logistic regression in Group 2

Feature selection method	**Logistic Regression**				
	3-input	4-input	5-input	6-input	7-input
GA	74.76	63.81	67.14	54.76	54.29
CC	71.43	71.43	61.43	68.10	61.43
ReliefF	50.59	48.10	48.10	44.76	41.43
CC-GA	74.76	63.81	60.48	64.29	60.48
ReliefF-GA	74.76	74.76	71.43	58.10	61.43

Table 7.20: AUC for logistic regression in Group 2

Feature selection method	**Logistic Regression**				
	3-input	4-input	5-input	6-input	7-input
GA	0.70	0.64	0.57	0.43	0.47
CC	0.67	0.67	0.59	0.65	0.59
ReliefF	0.45	0.39	0.41	0.43	0.41
CC-GA	0.70	0.64	0.61	0.63	0.54
ReliefF-GA	0.70	0.70	0.68	0.55	0.60

As in the other classification methods, the results from Group 2 are generally better than the results obtained in Group 1. Table 7.17 to Table 7.20 also show that the best accuracy is obtained by Group 1 at 70.95% and by CC-GA-6-input model. Whereas, for Group 2, GA-3-input, CC-GA-3-input, ReliefF-GA-3-input and ReliefF-GA-4-input achieved the best classification accuracy of 74.76% and the AUC of 0.70.

7.5 Discussion

This section summarises and compares the results generated from different classification methods as discussed in section 7.4. Table 7.21 and Table 7.22 summarize the best accuracy for the *n*-input model based on the feature selection method for Group 1 and Group 2. The summary is also depicted in the graph as shown in Figure 7.3 and Figure 7.4 respectively.

Table 7.21: Best accuracy for *n*-input model based on feature selection method for Group 1

Feature selection method	***n*-input model**				
	3-input	4-input	5-input	6-input	7-input
GA	70.95	67.62	64.76	68.10	64.29
CC	64.29	74.76	67.62	67.62	64.29
ReliefF	61.43	50.59	58.10	64.29	64.29
CC-GA	67.62	67.62	63.81	70.95	64.76
ReliefF-GA	67.14	64.29	67.62	64.29	64.76

Table 7.22: Best accuracy for *n*-input model based on feature selection method for Group 2

Feature selection method	***n*-input model**				
	3-input	4-input	5-input	6-input	7-input
GA	74.76	67.62	70.95	60.95	54.29
CC	71.43	71.43	64.76	68.10	67.62
ReliefF	54.29	54.29	48.10	67.14	67.14
CC-GA	74.76	70.48	61.43	64.29	64.29
ReliefF-GA	93.81	93.81	74.76	74.43	68.10

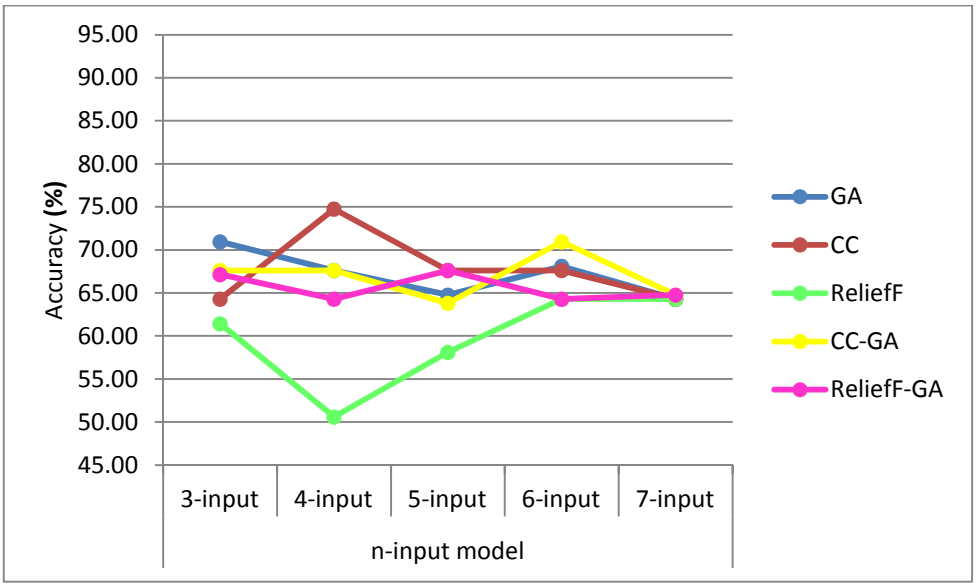


Figure 7.3: Graphs for best accuracy for n-input model based on feature selection method for Group 1

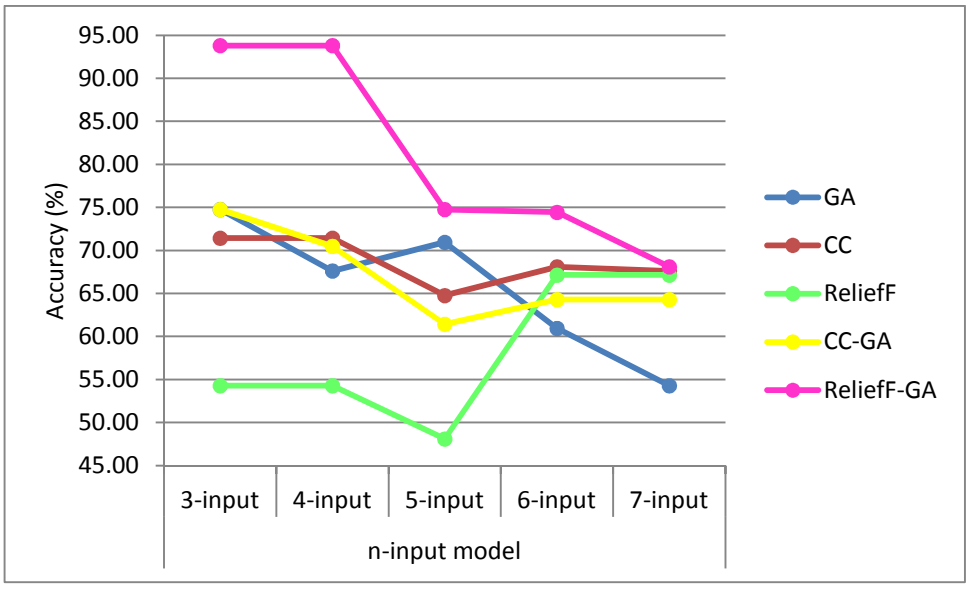


Figure 7.4: Graphs for best accuracy for n-input model based on feature selection method for Group 2

For Group 1 (Figure 7.3), the correlation coefficient (CC) feature selection method performed better than the other methods with the highest accuracy of 74.76% in 4-input model. There are three models that achieved accuracy of above 70%; the other two are GA-3-input and CC-GA-6-input (Table 7.21). ReliefF feature selection method obtained the worst results when compared to the other methods.

As regards to Group 2, the ReliefF-GA feature selection method outperformed the others in all *n*-input models, with the highest accuracy of 93.81%. There are ten models with an accuracy above 70% as shown in Table 7.22; this confirms that Group 2 which includes genomic variables achieved higher accuracy with feature selection methods. In addition, most of the models with higher accuracy are the lower input models with 3 or 4-input only.

Next, Table 7.23 and 7.24 lists the best accuracy by classification method and the graphs are depicted in Figure 7.5 and 7.6 for both Group 1 and Group 2.

Table 7.23: Best accuracy by classification method for Group 1

Feature selection method	**Classification method**			
	ANFIS	ANN	SVM	LR
GA	70.95	52.43	61.43	68.10
CC	74.76	54.48	60.95	67.62
ReliefF	64.29	51.52	54.29	50.59
CC-GA	67.62	50.48	63.81	70.95
ReliefF-GA	67.62	56.76	64.29	61.43

Table 7.24: Best accuracy by classification method for Group 2

Feature selection method	**Classification method**			
	ANFIS	ANN	SVM	LR
GA	74.76	51.48	74.76	74.76
CC	61.90	57.38	67.62	71.43
ReliefF	67.14	49.48	54.29	50.59
CC-GA	74.76	55.19	74.76	74.76
ReliefF-GA	93.81	84.62	74.76	74.76

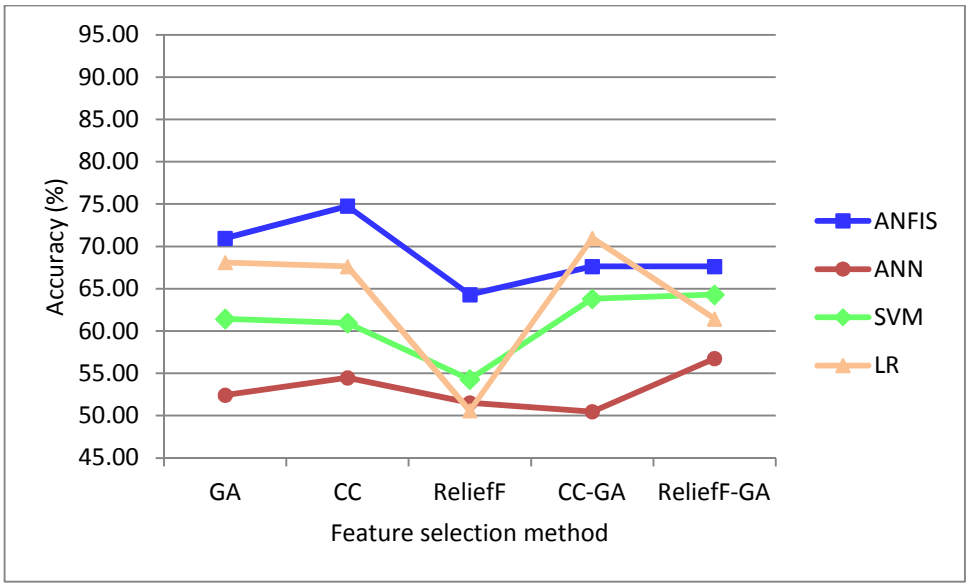


Figure 7.5: Graphs for best accuracy by classification method for Group 1

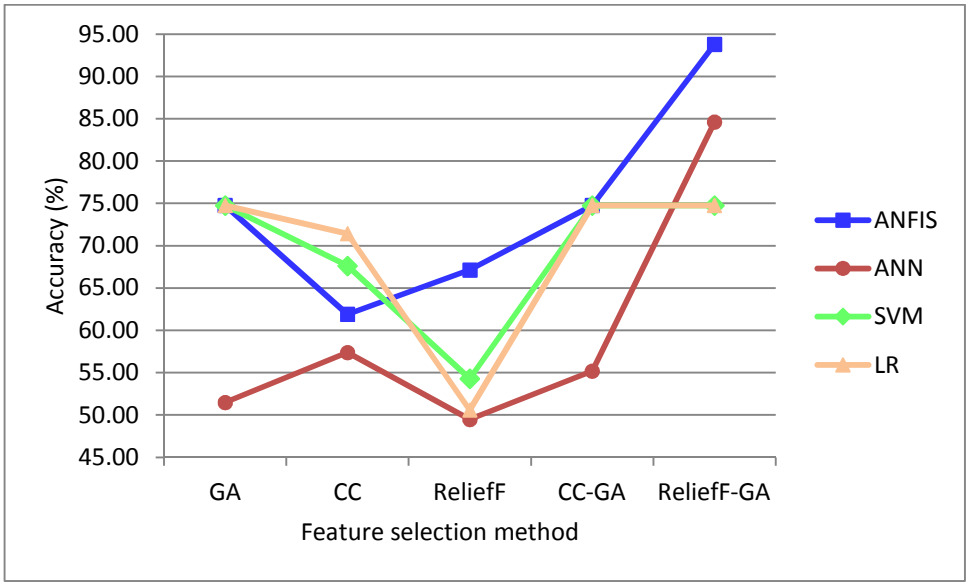


Figure 7.6: Graphs for best accuracy by classification method for Group 2

From Figure 7.5, ANFIS performed the best in Group 1 for all types of feature selection methods except CC-GA method. All the classification methods except for the ANN performed worst in ReliefF feature selection method. ANN had the lowest accuracy rate if compared to other methods.

Whereas, in Group 2, ANFIS outperformed the others except in CC feature selection method. The best accuracy is achieved by ANFIS in ReliefF-GA method with the accuracy of 93.81%. All classification methods performed better in CC-GA and ReliefF-GA feature selection methods. As with Group 1, ANN had the lowest classification rate except in ReliefF-GA method. In overall, the performance of the classification method is better in Group 2 as compared to Group 1. Table 7.25 summarises the best model with their selected features.

Table 7.25: Best models with accuracy, AUC, classification method and selected features

	Accuracy	AUC	Classification method	Selected features
Group 1				
CC-4-input	74.76	0.70	ANFIS	*Age,Gen,Inv,PN*
GA-3-input	70.95	0.66	ANFIS	*Gen,Smo,PN*
CC-GA-6-input	70.95	0.72	LR	*Gen,Dri,Node,PT,PN,Sta*
Group 2				
ReliefF-GA-3-input	**93.81**	**0.90**	**ANFIS**	***Dri,Inv,p63***
ReliefF-GA-4-input	**93.81**	**0.90**	**ANFIS**	***Dri,Inv,Tre,p63***
ReliefF-GA-3-input	84.62	0.83	ANN	*Dri,Inv,p63*
GA-3-input	74.76	0.74	ANFIS	*Inv,Node,p63*
CC-GA-3-input	74.76	0.70	ANFIS	*Inv,Node,p63*
CC-GA-3-input	74.76	0.70	SVM	*Inv,Node,p63*
CC-GA-3-input	74.76	0.70	LR	*Inv,Node,p63*
ReliefF-GA-3-input	74.76	0.70	SVM	*Dri,Inv,p63*
ReliefF-GA-3-input	74.76	0.70	LR	*Dri,Inv,p63*
Relief-GA-4-input	74.76	0.70	LR	*Dri,Inv,Tre,p63*
Relief-GA-5-input	74.76	0.70	SVM	*Age,Gen,Smo,Dri,p63*
Relief-GA-6-input	74.43	0.66	SVM	*Age,Gen,Smo,Dri,Inv,p63*
Relief-GA-4-input	73.38	0.75	ANN	*Dri,Inv,Tre,p63*
Relief-GA-4-input	71.43	0.68	SVM	*Dri,Inv,Tre,p63*
Relief-GA-5-input	71.43	0.68	LR	*Age,Gen,Smo,Dri,p63*
CC-3-input	71.43	0.67	LR	*Inv,PN,p63*
CC-4-input	71.43	0.67	LR	*Age,Inv,PN,p63*
CC-GA-4-input	70.48	0.71	ANFIS	*Gen,Inv,Size,p53*

From Table 7.25, the models with the highest accuracy are ReliefF-GA-3-input and ReliefF-GA-4-input from Group 2 with ANFIS classification, the accuracy is 93.81%

and AUC of 0.90. The features selected are *Drink, Invasion* and *p63* and *Drink, Invasion, Treatment,* and *p63* respectively. This is followed by the ReliefF-GA-3-input model from Group 2 with ANN classification, with the accuracy of 84.62% and AUC of 0.83. Most of the best models are generated from the ReliefF-GA feature selection method; this proves that the features selected by this method are the optimum features for the oral cancer prognosis dataset.

The results shown are in accordance with the objective of this research in which the classification performance is much better with the existence of genomic variables in Group 2. From the results in Table 7.25, the best feature selection method for oral cancer prognosis is ReliefF-GA with ANFIS classification. This proves that the ANFIS is the most optimum classification tool for oral cancer prognosis.

Since there are two top models with the same accuracy, hence, the simpler one will be chosen, which is the ReliefF-GA-3-input model with ANFIS classification, and the optimum subset of features are *Drink, Invasion* and *p63*. These findings are in accordance with some previous studies which have proved that these features are important prognosis factor for oral cancer survival. Alcohol consumption has always been considered as a risk factor and one of the reasons for poor prognosis of oral cancer (Cordon et al., 2001; Jefferies & Foulkes, 2001; Leite, 1997; Reichart, 2001; Zain, 2001).Walker et al., (2003) have shown that the depth of invasion is one of the most important predictors of lymph node metastasis in tongue cancer and in the different researches done by Asakage et al., (1998), Giacomarra et al., (1999), Morton et al., (1994), Williams et al., (1994), discovered a significant link between the depth of invasion and oral cancer survival. As regards to *p63*, Muzio, et al. (2005b) showed that p63 over expression associates with poor prognosis in oral cancer. In the next section,

the top best model, ReliefF-GA-3-input will be compared and validated with other permutation of 3-input features and also the full model.

7.6 Significance Testing

The significance test used in this research is the Kruskal-Wallis test. Kruskal-Wallis is a non-parametric test to compare samples from two or more groups and returns the *p*-value. Non-parametric tests means the tests are not severely affected by changes in a small portion of the data. For this research, we want to test is there any statistical significant difference between the accuracy results generated for the 3-input model of Group 2 for five feature selection methods. Thus, the null hypothesis is set as: H_0 = *There is no difference between the results of the different feature selection models.* If the *p*-value computed from the test is 0.05 or less, the H_0 is rejected, which means there is a difference between *the results of the different* feature selection methods; if the *p-value* > 0.05, the null hypothesis is accepted, which means there is no difference between the *results of the different* feature selection methods. The results and box-plots for the Kruskal-Wallis test are shown in Figure 7.7 and 7.8 respectively. Figure 7.7 is a standard ANOVA table, calculated using the ranks of the data rather than their numeric values. Ranks are found by ordering the data from smallest to largest across all groups, and taking the numeric index of this ordering. The entries in the ANOVA table are the sums of squares (SS), degrees of freedom (df), and mean square (MS) on the ranks. The chi-square statistic is used and the *p-value* measures the significance of the chi-square statistic. Figure 7.8 shows the box plots for each column of *x* (the value of *x*), in this case, the accuracy results by each type of feature selection methods.

The *p-value* (Prob>Chi-sq in Figure 7.7) that generated was 0.0312, which is less than 0.05, this means the H_0 is rejected and there is a significant difference between the feature selection methods.

Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	347.875	4	86.9688	10.62	0.0312
Error	274.625	15	18.3083		
Total	622.5	19			

Figure 7.7: Kruskal-Wallis ANOVA table

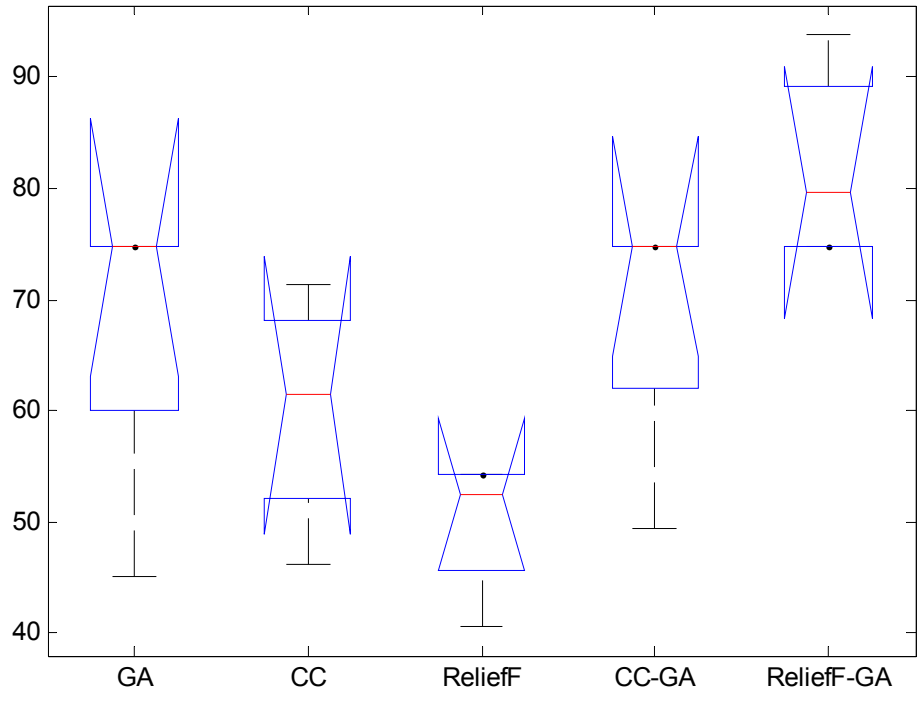


Figure 7.8: Box plots for Kruskal-Wallis test

7.7 Validation Testing

In this section, the best model of ReliefF-GA-3-input model is compared with other models with a random permutation of three inputs and also compared with the model with most selected features as listed in Table 7.4. The purpose is to validate that the features selected by the ReliefF-GA method are the optimum subset for oral cancer prognosis. In addition, the full-input model (the model with all the 17 variables) will be tested as well in order to verify that the reduced model can achieve the same or better results than the full model. In this testing, the classification method used is ANFIS due to its best performance in the section 7.4 and the results are tabulated in Table 7.26.

In Table 7.26, different permutation of the 3-input models are tested and classified using ANFIS. The three inputs are generated randomly and the best accuracy obtained is 80.48% with an AUC of 0.70. The features selected are *Drink, p53* and *p63*. With regards to the most selected features (refer to Table 7.4), the best result is achieved by the ReliefF-GA method with the accuracy of 93.81% and the AUC of 0.90 as shown in Table 7.26. The features selected are *Drink, Invasion* and *p63*, which are the same as the features selected by the ReliefF-GA-3-input model as shown in Table 7.2. This proved that the features selected using ReliefF-GA are the optimum features to oral cancer prognosis in this research.

Table 7.26: Validation test with random permutation of 3-input model, most selected features model and full input model for Group 2

Models	**ANFIS**	
	Accuracy (%)	**AUC**
Random permutation model		
Age, Inv, p63	64.76	0.63
Eth, Dri, p53	57.14	0.49
PT, PN, Sta	58.10	0.51
Gen, Node, Tre	70.95	0.59
Eth, Gen, Sub	39.05	0.32
Dri, p53, p63	**80.48**	**0.70**
Age, p53, p63	67.14	0.67
Gen, Dri, Inv	54.76	0.55
Site, Inv, Size	32.86	0.28
Age, Chew, Size	48.10	0.41
Smo, Site, PN	41.90	0.35
PM, Size, Tre	61.43	0.52
Sta, Size, p63	39.05	0.30
Dri, PN, p53	67.62	0.60
Model with most selected features		
Age, PT, PN, Size, p53 (GA)	44.76	0.42
Inv, PN, p63 (CC)	58.10	0.48
Age, Eth, Dri (ReliefF)	54.29	0.47
Inv, Size, p53 (CC-GA)	74.29	0.75
Dri, Inv, p63 (ReliefF-GA)	**93.81**	**0.90**
Full model		
Full model with ANFIS	N.A.*	N.A.*
Full model with NN	42.90	0.47
Full model with SVM	54.76	0.46
Full model with LR	54.76	0.59

*N.A. - Results not available due to over-fitting problem as the rule-base generated was too large

On the other hand, the full model with all the 17 variables is tested using different classification methods and the results are compared with the reduced model. The results of the full model are not promising and the results of full model using ANFIS cannot be generated due to the over-fitting problems as the rule base generated is too large.

Finally, the selected features are tested on the oral cancer dataset for 1-year and 2-year with ANFIS classification and the results are very promising with an accuracy for 1-

year prognosis of 93.33% and 2-year prognosis observed at 84.29%, the results are shown in Table 7.27.

Table 7.27: Classification results for 1-year, 2-year and 3-year oral cancer prognosis

Oral cancer prognosis	Accuracy (%)	AUC
1-year	93.33	0.90
2-year	84.29	0.77
3-year	93.81	0.90

7.8 Model Validation Study for Oral Cancer Clinicians

Human experts' prediction ability in medical prognosis can never be replaced by any computerised tools/models. The purpose of developing such computerised tools/models is to provide aids in the prediction by combining the prognosis generated by the model with the clinician's own estimate of prognosis in order to improve the accuracy of prognosis. Thus, the validation study for the clinicians is necessary to further validate and assess the performance of such developed computerised models (Goddard et al. 2011; Scott et al., 2011; Wyatt & Altman, 1995).

A model validation study was carried out involving five (5) oral cancer clinicians from the Faculty of Dentistry, University of Malaya, Malaysia. The group of clinicians ranges from senior lecturers to professors.

The objectives of this model validation study were:

(i) To measure the prediction accuracy of human expert predictions.

(ii) To measure the prediction consistency of human expert predictions.

The clinicians were asked to choose the most significant variables for oral cancer, to state whether the inclusion of the genomic markers would improve the accuracy of oral cancer prognosis and to make a 3-year prognosis for oral cancer based on the selected clinicopathologic variables. The form for the model validation study is attached in Appendix D.

This validation study is divided into two sections, which are section A and section B. There are two questions in section A. In Question 1, the clinicians were required to choose four (4) most significant variables for oral cancer prognosis from the list of clinicopathologic variables given and rank the variables accordingly. As in question 2, the clinicians were asked to give their opinion whether the inclusion of the genomic markers would improve the accuracy of oral cancer prognosis or not. For section B, the clinicians were asked to indicate their prognosis based on the combination of clinicopathologic variables listed in the three models, which are Model 1, Model 2 and Model 3.

7.8.1 Results and Analysis on the Model Validation Study for Oral Cancer Clinicians

(a) Section A - Question 1

Four most significant clinicopathologic variables listed by the clinicians were summarised and weighted. More weightage was given to the variable with higher rank. For example, the variable ranked number one was given a weightage of "4", the variable ranked number two was given a weightage of "3" and so on. The results are shown as in Table 7.28 and Figure 7.9 respectively. Table 7.28 shows the number of oral cancer clinicians for each variable and weightage while Figure 7.9 shows the bar chart of total weightage for each variable.

Table 7.28: Number of oral cancer clinicians for each variable and weightage

Variables	Weightage				Total
	4	3	2	1	
PN	2	2	0	0	14
Stage	2	1	1	0	13
Site	0	1	1	1	6
Inv	0	0	2	2	6
PT	1	0	0	1	5
Size	0	1	0	0	3
Treat	0	0	1	0	2
Subtype	0	0	0	1	1

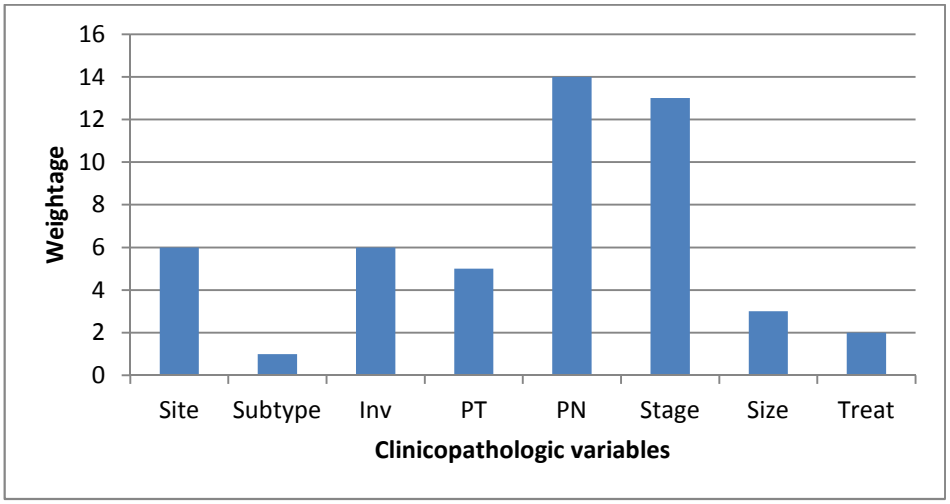


Figure 7.9: Bar chart for Section A - Question 1

From Figure 7.9, the top four most significant clinicopathologic variables chosen by the clinicians are *PN*, *Stage*, *Inv* and *Site*. Three out of these four variables, which are *PN*, *Stage* and *Inv* were selected as the most selected features for the feature selection methods as listed in Table 7.4. This finding shows that the developed feature selection methods were almost similar to the clinicians' selections on the clinicopathologic variables.

(b) Section A - Question 2

All the clinicians agreed that the inclusion of genomic markers will help to improve the accuracy of oral cancer prognosis. However, all of the clinicians agreed

that currently there is no specific genomic marker yet for oral cancer prognosis and the results from different studies on genomic markers are varied, thus it makes genomic markers difficult to put in real cancer practice.

(c) Section B

In this section, three models were selected and given to the clinicians for the prognosis. The models selected are the top three best models for Group 1 as listed in the Table 7.25, which are CC-4 input-ANFIS, GA-3 input-ANFIS and CC-GA-6 input-LR. Table 7.29 shows the information for these three models.

Table 7.29: Information for the selected models

	AI model in Group 1	**Clinicopathologic Variables**
Model 1	GA-3-input-ANFIS	*Gen, Smo, PN*
Model 2	CC-4-input-ANFIS	*Age, Gen, Inv, PN*
Model 3	CC-GA-6-input-LR	*Gen, Dri, Nodes, PT, PN, Sta*

In this validation study, only models from Group 1 were selected, as clinicians never make prognosis based on the genomic markers. Currently, there is no genomic marker accepted as prognostic value in oral cancer as mentioned in the discussions in Section A- Question 2.

Oral cancer clinicians' prognoses for the three models were measured. The accuracy, sensitivity, specificity and areas under receiver operating characteristic curve (AUC) of the oral cancer clinician prognosis are shown in Table 7.30 and the receiver operating characteristic (ROC) curve for the models are shown in Figure 7.10. Table 7.31 compares the accuracy and AUC for the oral cancer clinician prognosis and the AI prognosis.

Table 7.30: Accuracy, sensitivity, specificity and AUC of oral cancer clinician prognosis

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Model 1	67.70	80.00	45.45	0.64
Model 2	71.00	100.00	18.20	0.76
Model 3	67.70	80.00	45.45	0.66

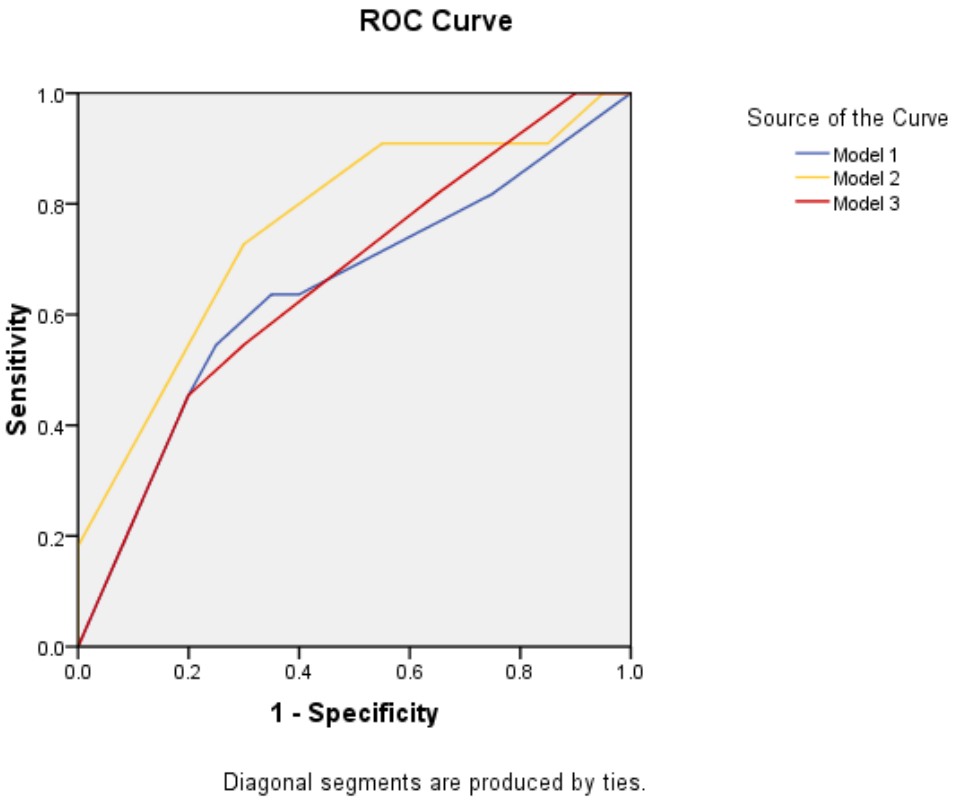


Figure 7.10: Receiver operating characteristic (ROC) curves for the oral cancer clinician prognosis

Table 7.31: Accuracy and AUC for oral cancer clinician prognosis and AI prognosis

	Oral cancer clinician prognosis		AI prognosis	
Model	Accuracy (%)	AUC	Accuracy (%)	AUC
Model 1	67.70	0.64	70.95	0.66
Model 2	71.00	0.76	74.76	0.70
Model 3	67.70	0.66	70.95	0.72

From Table 7.30, the accuracy and AUC for both oral cancer clinician prognosis and AI prognosis are about the same. Model 2 has the best performance among the three, with

an accuracy of 71.00% and an AUC of 0.76 for oral cancer clinician prognosis, and an accuracy of 74.76%, AUC of 0.70 for AI prognosis. These findings proved that our AI prognoses are correct and that similar to the oral cancer clinicians' prognoses. It is hope that with the inclusion of genomic markers, the accuracy of oral cancer prognosis could be improved as shown in our best AI model (ReliefF-GA-3-input-ANFIS model) in Group 2, with the accuracy of 93.81% and AUC of 0.90 (as shown in Table 7.25).

7.9 Summary

This chapter discusses and compares the results generated using the proposed 5 feature selection methods and the ANFIS classification model. First, the feature selection methods are applied on the two groups of oral cancer dataset, which are Group 1 with clinicopathologic variables only and Group 2 with clinicopathological and genomic variables. For both groups, *n*-input models are selected, with n = 3, 4, 5, 6, 7 and the selected features are listed in Table 7.1 and 7.2 respectively. The most selected features for each feature selection methods are summarised in Table 7.4.

Next, ANFIS classification model using the oral cancer prognosis dataset with features selected from the proposed feature selection methods is implemented. Two other AI classification methods, which are artificial neural network (ANN) and support vector machine (SVM) and one statistical method which is logistic regression (LR), are used to test and compare with the ANFIS model. All the classification experiments are performed on both Group 1 and Group 2.

In summary, first, ANFIS outperformed the rest of the classification methods for both Group 1 and Group 2. Second, the performances of Group 2 are generally much better than those from Group 1. Third, the best model is the ReliefF-GA-3-input with ANFIS,

and the optimum feature subset selected is *Drink, Invasion* and *p63*. The optimum feature subset is validated with other random permutation of the 3-input model, the model with most selected features for each feature selection method, and the full model and it was shown that the model with the optimum feature subset achieved the highest accuracy. A validation study for oral cancer clinicians was conducted in order to validate the results obtained with the developed AI models and the oral cancer clinicians. The results showed that the prognoses by the oral cancer clinicians were similar to the prognoses from the developed AI models.

In accordance with the aims of this chapter, we have shown that Group 2 with clinicopathologic and genomic variables performs better when compared to Group 1. Next, the optimum subset of features have been identified and verified, the key features for oral cancer prognosis are *Drink, Invasion* and *p63*. Lastly, the ANFIS classification model has been proved to be the optimum classification tool for oral cancer prognosis.

The sample size of this research is very small, which is 31 samples only, this may not be sufficient for some researchers in the classification research. However, this is inevitable in the medical research, especially for oral cancer research in Malaysia where we do not have many samples and there are many medical confidentiality problems as discussed in Chapter 1. In order to overcome this problem, a re-sampling technique, which is the cross-validation, is implemented in the classification experiments.

CHAPTER 8

Conclusion and Future Work

8.1 Research Summary

The overall aim of this research is to apply artificial intelligent (AI) techniques in oral cancer prognosis based on the clinicopathologic and genomic markers. To this end, a hybrid of AI oral cancer prognostic model with optimum feature subset has been developed and the end results are very promising. This section summarises the findings in the development of the oral cancer prognostic model in line with the research objectives.

Chapter 4 explains the overview path of this research and all the methodologies used. There are five objectives for this research. First, is to identify the most common clinicopathologic markers and second, to analyse the genomic markers from the results of immunohistochemistry (IHC) staining. These are described in Chapter 5 where the methods, preparations and procedures for acquiring oral cancer prognosis data were discussed. Third, is to determine the optimum clinicopathologic and genomic markers for oral cancer prognosis using feature selection methods. The details of feature selection methods were discussed in Chapter 6 and the results and discussions were further discussed in Chapter 7. Fourth, is to develop a prognostic model for oral cancer prognosis using ANFIS techniques and to prove that the proposed model is the optimum tool for oral cancer prognosis. For this objective, ANFIS is proposed to cater for the needs of the small sample size and the results and comparisons with other techniques (artificial neural network, support vector machine and logistic regression) were discussed in Chapter 7. Lastly, is to prove that the prognostic results are more accurate

with the presence of both clinicopathologic and genomic markers, as discussed in Chapter 7.

The clinicopathologic data used in this research are available from MOCDTBS at Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, University of Malaya. With the help from the oral cancer experts from OCRCC, 15 clinicopathologic variables were identified. As regards to the genomic data, due to time, cost and medical tissues limitation, only two genomic variables were identified and used in this research, which are *p53* and *p63*. 31 oral cancer cases were selected with the help from the staff of OCRCC and the tissues of the selected cases were prepared in the form of formalin-fixed paraffin embedded macroarray block (TMaA block). Immunohistochemistry (IHC) staining was performed on the selected tissues and the results of staining were analysed using the image analyser system. The results were categorised into two which are positive and negative staining, the tissue is considered positive if more than 10% of the nuclear is stained.

Next, pre-processing methods were implemented on the oral cancer prognosis dataset with clinicopathologic and genomic variables. Data cleansing, discretization and transformation were implemented on the selected dataset. The dataset was divided into two groups, with Group 1 (clinicopathologic variables) and Group 2 (clinicopathologic and genomic variables). 3-year prognostic data was used. Subsequently, feature selection methods were implemented with the objectives to reduce the number of input variables to avoid over-fitting and to find out an optimum feature subset for oral cancer prognosis. Five feature selection methods were implemented, which were genetic algorithm (GA), Pearson's correlation coefficient, Relief-F, hybrid CC-GA, and hybrid ReliefF-GA. The number of features selected was ranged from three to seven inputs (*n*-

input model) and the selected features from each method were tested using the proposed ANFIS classification model.

The proposed ANFIS classification model was proposed in order to classify whether the patients were alive or dead after subsequent years of diagnosis, in this case, 1-year to 3-year. ANFIS is implemented on the *n*-input models generated from the five feature selection methods. Due to the small sample size, a re-sampling technique which was the *k*-fold cross-validation was used. The results generated from ANFIS were compared with other AI methods (Artificial neural network and support vector machine) and the statistical method of logistic regression. Furthermore, the ANFIS classification model was validated using different permutations of the *n*-input model, most selected features model, full input model and also used to classify for 1-year and 2-year oral cancer prognosis. The analyses and findings from the proposed oral cancer prognostic model are:

(i) The performance of Group 2 (clinicopathologic and genomic variables) is better than Group 1 (clinicopathologic variables). This is in accordance with the objective of this research, which shows that the prognostic result is more accurate with the combination of clinicopathologic and genomic markers.

(ii) The model with the best accuracy is the ReliefF-GA-3-input model with the ANFIS classification model and the Kruskal-Wallis test carried shows that the results from this model shows a significant difference as compared to the 3-input model of GA, CC, ReliefF and CC-GA.

(iii) The optimum subset of features for oral cancer prognosis is *drink, invasion* and *p63* and this finding is in accordance with similar studies in the literature.

(iv) The ANFIS classification model achieved the best accuracy in oral cancer prognosis when compared to artificial neural network, support vector machine and statistical method of logistic regression.

(v) The prognostic result is more accurate with fewer inputs (reduced model) in comparison with the full model.

(vi) The hybrid ReliefF-GA-ANFIS prognostic model performed well in 1-year and 2-year oral cancer prognosis data.

As a conclusion, the hybrid AI model of ReliefF-GA-ANFIS with 3-input features of *drink, invasion* and *p63* achieved the best accuracy and is feasible to be used as an aid to clinicians for the prognosis of oral cancer.

8.2 Research Constraints

Medical informatics is a comparatively new research area in Malaysia, hence the medical databases that are available are very limited. This has limited the research activities in this area, as there are not enough medical samples or tissues available for the research experiments. Furthermore, most of the medical records available are kept in the hardcopy format (paper format), thus, it takes time to transform these data into the softcopy or computerised format.

Limited medical samples/tissues is a constraint for the genomic data, this is the main reason for the small sample size of the medical data. Moreover, it is time consuming to prepare the medical samples/tissues for the purpose of genomic data, from cutting of the tissues during the surgery, preparation of the tissues in the lab, staining using specific reagents/antibody, and analysis of the staining results by the oral cancer experts. The whole process takes several weeks to several months depending on the number of

samples. Besides that, a high cost is needed in obtaining the genomic data, costs incurred in the purchasing of the testing equipments (microscope, software, camera, computer, etc.) and materials (reagents, antibody, etc.), payment and honorarium for the laboratory technicians and other costs.

In this research, data for a maximum of 3-year prognosis was available. This is due to the incomplete records for cases of more than three years. For records of more than 3-year, another one or two years are needed to obtain sufficient cases for the proposed prognostic model. Due to the time and cost limitation, only 1-year to 3-year survival are included in this research, and only two genomic data are selected, which are *p53* and *p63*.

8.3 Research Contributions

The contribution of this research can be divided into four parts. First, it has been proven that the prognostic results are better with both clinicopathologic and genomic markers. This is supported by the research done by Catto et al., (2006), Exarchos et al. (2011), Futschik et al., (2003), Gevaert et al., (2006), Oliveira et al., (2008), Passaro et al., (2005), Seker et al., (2003), and Sun et al. (2007).

Second, a hybrid feature selection model of ReliefF-GA was proposed as the feature selection method for oral cancer prognosis. This hybird model had shown to predict prognosis better than the full input model. Thus, the proposed ReliefF-GA is feasible to use as a feature selection method for other medical dataset as well, especially for those researches which utilize both clinicopathologic and genomic markers.

Third, a 3-input model with the features of *drink, invasion* and *p63* had been identified as the optimum subset for oral cancer prognosis.

Fourth, the proposed ANFIS classification model was found to have high classification ability when compared to artificial neural network, support vector machine and logistic regression. The results showed that the ANFIS prognostic model is suitable for small sample size data with the proposed optimum feature subset. This finding will help the clinicians and oral cancer experts in determining the prognosis/survival of oral cancer patients with very few markers. However, more tests and experiments needed to be done in order to further verify the results obtained in this research as discussed in the future work section of 8.4. Nevertheless, this finding provides a good foundation for future computer-based intelligent prognostic modelling especially at the local scenario.

Fifth, through the identification of fewer markers for oral cancer prognosis, it is hoped that this will aid clinicians in carrying out prognostic procedures, and thus help them in making a more accurate prognosis in a shorter time at lower costs. Furthermore, the results of this research help patients and their family plan their future and lifestyle through a more reliable prognosis.

8.4 Future Work

This is the first research in Malaysia which implements AI techniques in oral cancer prognosis using both clinicopathologic and genomic data, the contributions of this research is discussed in section 8.3. However, there is still room for improvement for this research. Some suggestions for future work are listed below:

(i) Increase the sample size of the dataset by providing more medical samples thus making it closer to the real population and improving the prediction accuracy.

(ii) Include more genomic markers such as *EGFR*, *p16, CYP1A1*, *cyclin D1* and others as discussed in section 3.4.2.

(iii) The validation exercise of the proposed model could be extended to other classifiers such as Bayesian approach, *k*-nearest neighbour algorithm, genetic algorithm, decision tree and hybrid AI models.

(iv) Use other feature selection methods such as simulated annealing, information gain, gain ratio, decision tree, and others.

(v) Use DNA microarray as the alternate source for the genomic data. A microarray is a tool containing samples of many genes that are arranged in a regular pattern in a small membrane or glass slide for the purpose of gene expression analysis. By using microarrays, researchers can get the expression levels of hundreds or thousands of genes in a single experiment, thus, reducing time and cost.

8.5 Concluding Remarks

As a summary, the proposed prognostic model with ReliefF-GA and ANFIS provides a computer-based intelligent approach to oral cancer prognosis using only three combinations of clinicopathologic and genomic markers. This prognostic model is feasible to aid the clinicians in the decision support stage and to identify the high risk markers to better predict the survival rate for each oral cancer patient. However, it is not recommended for clinical use yet as more tests and validations are needed to be done in order to further verify the results obtained in this research. Although the sample size is small, it is hoped that this research will set a stepping stone to embark more Malaysians in a similar research.

REFERENCE

Abcam. (2010). IHC-Paraffin Protocol (IHC-P). Available from: www.abcam.com/technical. [Accessed 20 September 2010].

Abdul-Kareem, S., Baba, S., Zubairi, Y. Z., Prasad, U., & Wahid, M. I. A. (2002). *Prognostic Systems for NPC: A Comparison of the Multilayer Perceptron Model and the Reccurent Model.* Paper presented at the 9th International Conference on Neural Information Processing.

Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect in the classifier accuracy. *Classification, Clustering and Data Mining Applications*, 639-648.

Akay, M. F. (2008). Support Vector Machines Combined with Feature Selection for Breast Cancer Diagnosis. *Expert Systems with Applications, 36*(2, Part 2), 3240-3247.

Anantharaman, D., M.Chaubal, P., Kannan, S., A.Bhisey, R., & B.Mahimkar, M. (2007). Susceptibility to Oral Cancer by Genetic Polymorphisms at CYP1A1, GSTM1, and GSTT1 loci among Indians:Tobacco Exposure as a risk modulator. *Carcinogenesis, 28*(7), 1455-1462.

Arulchinnappan, S., Karunakaran, K., & Rajendran, G. (2011). The Use of Fuzzy Correlation to Identify People at Risk of Oral Cancer. *European Journal of Scientific Research, 52*(3), 332-338.

Asakage, T., Yokose, T., Mukai, K., Tsugane, S., Tsubono, Y., Asai, M., et al. (1998). Tumor thickness predicts cervical metastasis in patients with stage I/II carcinoma of the tongue. *Cancer, 82*, 1443-1448.

Baker, O. F., & Abdul-Kareem, S. (2008). *ANFIS Models for Prognostic and Survival Rate Analysis*. Paper presented at the IEEE International Conference on Management of Innovation & Technology.

Baker, O. F., & Kareem, S. A. (2007). *Using Genetic Algorithm to Evolves Algebraic Rule-Based Classifiers for NPC Prognosis*. Paper presented at the International Conference on Intelligent and Advanced Systems.

Barnard, L., & Lan, W. Y. (2008). Treatment of Missing Data: Beyond Ends and Means. *Journal of Academic Ethics, 6*, 173–176.

Baronti, F., & Starita, A. (2007). Hypothesis Testing with Classifier Systems for Rule-Based Risk Prediction. *EvoBIO 2007, LNCS 4447*, 24-34.

Bellaachia, A., & Guven, E. (2006). *Predicting Breast Cancer Survivability Using Data Mining Techniques.* Paper presented at the 2006 SIAM International Conference in Data Mining Maryland.

Berenji, H. R., & Khedkar, P. (1992). Learning and Tuning Fuzzy Logic Controllers Through Reinforcements. *IEEE Transactions on Neural Networks 3*(5), 724-740.

Blood, D. C. & Studdert, V. P. (2007). Saunders Comprehensive Veterinary Dictionary. (3rd ed.). Elsevier.

Bouda, M., Gorgoulis, V. G., Kastrinakis, N. G., & Giannoudis, A., et al. (2000). High Risk HPV types are frequently detected in potentially malignant and malignant oral lesions, but not in normal oral mucosa. *Mod Pathol, 113*, 644-653.

Braga-Neto, U. M., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics, 20*(3), 374-380.

Brinkman, B. M. N., & Wong, D. T. W. (2006). Disease mechanism and biomarkers of oral squamous cell carcinoma. *Current Opinion in Oncology, 18*(3), 228-233.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2*, 121-167.

Cancer. (2009) Available from: http://www.who.int/mediacentre/factsheets/fs297/en/. [Accessed 20th May 2010].

Castanho, M. J. d. P., Barros, L. C. d., Yamakami, A., & Vendite, L. L. (2008). Fuzzy Expert System: An Example in Prostate Cancer. *Applied Mathematics and Computation, 202*, 78-85.

Catto, J. W. F., Abbod, M. F., Linkens, D. A., & Hamdy, F. C. (2006). Neuro-Fuzzy Modeling: An Accurate and Interpretable Method for Predicting Bladder Cancer Progression. *The Journal of Urology, 175*, 474-479.

Chap, T. L. (1997). *Applied Survival Analysis*. Wiley-Interscience, John Wiley and Sons, Inc., New York.

Chen, P.-H., Shieh, T.-Y., Ho, P. S., Tsai, C.-C., Yang, Y.-H., Lin, Y.-C., et al. (2007). Prognostic Factors Associated with the Survival of Oral and Pharyngeal Carcinoma in Taiwan. *BMC Cancer, 7*(101).

Chih-Chung, C., & Chih-Jen, L. (2011). LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 27:21--27:27.

Chih-Wei, H., Chang, C.-C., & Lin, C.-J. (2010). A Practical Guide to Support Vector Machine *Technical Report*: National Taiwan University.

Clark, T., Bradburn, M., & Love, S. A., DG. (2003). Survival Analysis Part 1: Basic Concepts and First Analyses. *British Journal of Cancer, 89*, 232-238.

Correlation Coefficient. (2010) Available from: http://mathbits.com/mathbits/tisection/statistics2/correlation.htm. [Accessed 10th November 2010].

Colozza, M., Cardoso, F., & Sotiriou, C. (2005). Bringing molecular prognosis and prediction to the clinic. *Clin Breast Cancer, 6*, 61-76.

Cordon, O., Herrera, F., Hoffmann, F., & Magdalena, L. (2001). *Genetic Fuzzy Systems-Evolutionary Tuning and Leanrning of Fuzzy Knowledge Bases*. World Scientific Publishing.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning 20*, 273-297.

Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics, 2*, 59-78.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods. *Artificial Intelligence in Medicine, 34*, 113-127.

Dom, R. M. (2009). *A Fuzzy Regression Model for the Prediction of Oral Cancer Susceptibility.* PhD Thesis, University of Malaya, Kuala Lumpur, Malaysia.

Dom, R. M., Abdul-Kareem, S., Abidin, B., Jallaludin, R. L. R., Cheong, S. C., & Zain, R. B. (2008). Oral Cancer Prediction Model for Malaysian Sample. *Austral-Asian Journal of Cancer, 7*(4), 209-214.

Dom, R. M., Kareem, S. A., Zain, R., & Abidin, B. (2007). *An Adaptive Fuzzy Regression Model for the Prediction of Dichotomous Response Variables*. Paper presented at the Fifth International IEEE Conference on Computational Science and Applications.

Editors, A. H. D. (2007). The American Herritage Medical Dictionary. Houghton Mifflin Harcourt.

Efron, B., & Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Exarchos, K., Goletsis, Y., & Fotiadis, D. (2011). Multiparametric Decision Support System for the Prediction of Oral Cancer Reoccurrence. *IEEE Transactions on Information Technology in Biomedicine*, In press

F.Baker, O. (2010). *Application of Soft Computing Techniques in the Prediction of Survival in Cancer.* PhD Thesis, University of Malaya, Kuala Lumpur, Malaysia.

Fausett, L. (1993). *Fundamentals of Neural Networks: Architectures, Algorithms And Applications* (1st ed.). Prentice Hall.

Fielding, L. P., Fenoglio-Preiser, C. M., & Freedman, L. S. (1992). The Future of Prognostic Factors in Outcome Prediction for Patients with Cancer. *Cancer, 70*, 2367-2377.

Foundation, O. C. (2010). Oral Cancer Facts. Available from: http://www.oralcancerfoundation.org/facts/index.htm. [Accessed 23rd February 2010].

Fu, W. J., Caroll, R. J., & Wang, S. (2005). Estimating Misclassification Error with Small Samples via Bootstrap Cross-validation. *Bioinformatics, 21*(9), 1979-1986.

Fuentes-Uriarte, J., Garcia, M., & Castillo, O. (2008). *Comparative Study of Fuzzy Methods in Breast Cancer Diagnosis*. Paper presented at the Fuzzy Information Processing Society, 2008, New York.

Futschik, M. E., Sullivan, M., Reeve, A., & Kasabov, N. (2003). Prediction of Clinical Behaviour and Treatment for Cancers. *Applied Bioinformatics, 2*(3 Suppl), S53 - S58.

Gerard, L. C. C., Rampal, S., & Yahaya, H. (2005). *Third Report of the National Cancer Registry Cancer Incidence in Malaysia (2005)*. National Cancer Registry, Ministry of Health Malaysia Retrieved from http://www.makna.org.my/NCR/.

Gershenson, C. (2003). Neural Network For Beginners: Cornell University Library.

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, D., & Moor, B. D. (2006). Predicting the Prognosis of Breast Cancer by Integrating Clinical and Microarray Data with Bayesian Networks. *Bioinformatics, 22*(14), e184-e190.

Giacomarra, V., Tirelli, G., Papanikolla, L., & Bussani, R. (1999). Predictive factors of nodal metastases in oral cavity and oropharynx carcinomas. *Laryngoscope, 109*, 795–799.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*, 121-127.

Good, P. I. (2004). *Permutation, Parametric, and Bootstrap Tests of Hypotheses* (3rd ed.). New York. Springer.

Hamid, S., Yang, Y. H., Peng, K. N. L., Ismail, S. M., Zain, R. B., Lim, K. P., et al. (2008). MDM2 SNP309 does not confer an increased risk to oral squamous cell carcinoma but may modulate the age of disease onset. *Oral Oncology, 10*(1016).

Hammond, P., & Speight, P. M. (1998). Screening for Risk of Oral Cancer and Pre-cancer. *Intelligent Data Analysis In Medicine and Pharmacology*.

Hassan, M. R., Hossain, M. M., Begg, R. K., Ramamohanarao, K., & Morsi, Y. (2010). Breast-Cancer identification using HMM-fuzzy approach. *Computers in Biology and Medicine, 40*, 240-251.

Hayward, J., Alvarez, S. A., Ruiz, C., Sullivan, M., Tseng, J., & Whalen, G. (2010). Machine learning of clinical performance in a pancreatic cancer database. *Artificial Intelligence in Medicine, 49*, 187-195.

Immunohistochemistry. (2010) Available from: http://www.protocol-online.org/prot/Immunology/Immunohistochemistry/index.html. [Accessed 25 August 2010].

Institute, N. C. (2009) What You Need To Know About Oral Cancer. Available from: http://www.cancer.gov/cancertopics/wyntk/oral/page2. [Accessed 20th October 2010]

Institute, N. C. (2010) NCI Dictionary of Cancer Terms. Available from: http://www.cancer.gov/dictionary?CdrID=45618. [Accessed 12th October 2010].

Jang, J. S. R. (1993). ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Transactions on Systems, Man, and Cybernetics, 23*(3), 665-685.

Jang, J. S. R. (1996). *Input Selection for ANFIS Learning*. Paper presented at the Fifth IEEE International Conference on Fuzzy Systems.

Jefferies, S., & Foulkes, W. D. (2001). Genetic mechanisms in squamous cell carcinoma of the head and neck. *Oral Oncology, 37*, 115-126.

Jerez, J. M., Molina, I., Garcia-Laencina, P. J., Alba, E., Ribelles, N., Martin, M., et al. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine, 50*, 105-115.

Jones, A. S., Taktak, A.G.F., Helliwell, T.R., Fenton, J.E., Birchall, M.A., Husband, D.J., & Fisher, A.C. (2006). An Artificial Neural Network Improves Prediction of Observed Survival in Patients with Laryngeal Squamous Carcinoma. *Eur Arch Otorhinolaryngol, 263*, 541-547.

Kaehler, S. D. (1993) Fuzzy Logic Tutorial Available from: http://www.seattlerobotics.org/encoder/mar98/fuz/fl_part1.html. [Accessed 31st May 2010].

Kareem, S. A. (2001). *Application of Artificial Neural Network fot the Prognosis of Nasopharyngeal Carcinoma.* PhD Thesis, University of Malaya, Kuala Lumpur, Malaysia.

Kawazu, T., Kazuyuki, A., Yoshiura, K., Nakayama, E., & Kanda, S. (2003). Application of Neural Networks to the Prediction of Lymph Node Metastasis in Oral Cancer. *Oral Radiol. 2003, 19*, 137-142.

Kent, S. (1996). Diagnosis of Oral Cancer using Genetic Programming. *Technical Report*.

Kleinbaum, D. G. (1995). *Survival Analysis - A Self-Learning Text*. Springer-Verlag, New York.

Kononenko, I. (1994). *Estimating Attributes: Analysis and Extension of RELIEF*. Paper presented at the ECML-94 Proceedings of the European conference on machine learning on Machine Learning.

Leite, I. C. G., & Koifman, S. (1998). Survival analysis in a sample of oral cancer patients at a reference hospital in Rio de Janeiro, Brazil. *Oral Oncology, 34*(1998), 347-352.

Li, D. C., Hsu, H. C., Tsai, T. I., Lu, T. J., & Hu, S. (2007a). A New Method to Help Diagnose Cancers for Small Sample Size. *Expert Systems with Applications, 33*, 420-424.

Li, H., Li, D., Zhang, C., & Nie, S. (2007b). An Application of Machine Learning in the Criterion Updating of Diagnosis Cancer. *International Conference in Neural Networks and Brain 2005, 1*, 187-190.

Lin, C.-T., & Lee, C. S. G. (1991). Neural-Network based Fuzzy Logic Control and Decision System. *IEEE Transactions on Computers, 40*, 1320-1336.

Lin, R.-H., & Chuang, C.-L. (2010). A hybrid diagnosis model for determining the types of the liver disease. *Computers in Biology and Medicine, 40*, 665-670.

Li, X. M., Di, B., Shang, Y., Li, J., & Cheng, J. (2005). Clinicopathologic Features and Prognostic Factors of Cervical Lymph Node Metastasis in Oral Squamous Cell Carcinoma. *Chinese Journal of Cancer, 24*(2), 208-201.

Liu, J., Wyatt, J. C., & Altman, D. G. (2006). Decision tools in health care: focus on the problem, not the solution. *BMC Medical Informatics and Decision Making, 6*(4).

Marinakisa, Y., Douniasb, G., & Jantzenc, J. (2009). Pap smear diagnosis usinga hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. *Computers in Biology and Medicine, 39*, 69--78.

Markov, Z. (2011). Data Preprocessing. Available from: http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-3.html. [Accessed 16th May 2011].

Marques, C., Koifmanb, S., Koifmanb, R., Boffettac, P., Brennanc, P., & Hatagimaa, A. (2006). Influence of CYP1A1, CYP2E1, GSTM3 and NAT2 genetic polymorphisms in oral cancer susceptibility: Results from a case-control study in Rio de Janeiro. *Oral Oncology, 42*(6), 632-637.

MathWorks. (2010). Matlab User Guide for Global Optimization Toolbox 3. Retrieved from http://www.mathworks.com/help/pdf_doc/gads/gads_tb.pdf

Mehrotra, R., & Yadav, S. (2006). Oral Squamous cell carcinoma: Etiology, pathogenesis and prognostic value of genomic alterations. *Indian Journal of Cancer, 43*(2), 60-66.

Mitra, S., & Hayashi, Y. (2000). Neuro–Fuzzy Rule Generation: Survey in Soft Computing Framework. *IEEE Transactions on Neural Networks, 11*(3), 748-768.

Mitchell, T. M. (1997). *Machine Learning*. New York. McGraw-Hill.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics, 21*(15), 3301–3307.

Morrow, A. (2007). Oral Cancer: Types. Available from: http://www.omnimedicalsearch.com/conditions-diseases/oral-cancer-types.html. [Accessed 24th April 2009].

Morton, R., Ferguson, C., Lambie, N., & Whitlock, R. (1994). Tumor thickness in early tongue cancer. *Archives of Otolaryngology-Head & Neck Surgery, 120*, 717–720.

Mustafa, W. M. W., Ghani, W. M. N., Karen, N. L. P., Rahman, Z. A. A., Araham, M. T., Zain, R. B., et al. (2007). Survival of Oral Cancer Patients in Malaysia- A Multi-

Centre Audit. *J Dent Res 86 (Special Issue B), 101(SEA) (www.dentalresearch.org)*.

Muzio, L. L., D'Angelo, M., Procaccini, M., Bambini, F., Calvino, F., Florena, A. M., et al. (2005a). Expression of cell cycle markers and human papillomavirus infection in oral squamous cell carcinoma: Use of fuzzy neural networks. *International Journal of Cancer, 115*, 717–723.

Muzio, L. L., Santarelli, A., Caltabiano, R., Rubini, C., Pieramici, T., & Trevisiol, L. (2005b). p63 overexpression associates with poor prognosis in head and neck squamous cell carcinoma. *Human Pathology, 36*, 187-194

Nauck, D., & Kruse, R. (1993). *A Fuzzy Neural Network Learning Fuzzy Control Rules and Membership Functions by Fuzzy Error Backpropagation*. Paper presented at the IEEE International Conference in Neural Networks, San Francisco

Nauck, D., & Kruse, R. (1995). *NEFCLASS - A Neuro-Fuzzy Approach For The Classification Of Data*. Paper presented at the ACM Symposium on Applied Computing, Nashville.

NCBI. (2010). NCBI Reference Sequence. Available from: http://www.ncbi.nlm.nih.gov/RefSeq/. [Accessed 7th June 2010].

NeuroDimension. (2011) Genetic Server and Genetic Library. Available from: http://www.nd.com/products/genetic.htm. [Accessed 6th May 2011].

Obitko, M. (1998) Introduction to Genetic Algorithm. Available from: http://www.obitko.com/tutorials/genetic-algorithms/ga-basic-description.php. [Accessed 31st May 2010].

Oliveira, L. R., Ribeiro-Silve, A., Costa, J. P. O., Simoes, A. L., Di Matteo, M. A. S., & Zucoloto, S. (2008). Prognostic factors and survival analysis in a sample of oral squamous cell carcinoma patients. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology, 106*(5), 685-695.

Omar, Z. A., Ali, Z. M., & Tamin, N. S. I. (2006). Malaysian Cancer Statistics - Data and Figure, Peninsular Malaysia 2006.

OmniMedicalSearch.com. (2007) Oral Cancer: Types. Available from: http://www.omnimedicalsearch.com/conditions-diseases/oral-cancer-types.html. [Accessed 24th April 2009].

Passaro, A., Baronti, F., & Maggini, V. (2005). *Exploring Relationships Between Genotype and Oral Cancer Development Through XCS*. Paper presented at the GECCO'05.

Pyle, D. (1999). *Data Preparation for Data Mining*. San Francisco. Morgan Kaufmann.

Rao, V. S. H., & Kumar, M. N. (2011). A New Intelligence Based Approach for Computer-Aided Diagnosis of Dengue Fever. *IEEE Transactions on Information Technology in Biomedicine*, (In press).

Razak, A. A., Saddki, N., Naing, N. N., & Abdullah, N. (2010). Oral cancer survival among Malay patients in Hospital Universiti Sains Malaysia, Kelantan. *Asian Pacific Journal of Cancer Prevention, 11*(2), 187-191.

Regnier-Coudert, O., McCall, J., Lothian, R., Lamb, T., McClinton, S., & N'Dow, J. (2011). Machine learning for improved pathological staging of prostate cancer: A performance comparison on a range of classifiers. *Artificial Intelligence in Medicine*, In press.

Reichart, P. A. (2001). Identification of Risk Groups for Oral Precancer and Cancer and Preventive Measures. *Clin. Oral Invest, 5*, 207-213.

Rios, D. (2010) Neural networks: A requirement for intelligent systems. Available from: http://www.learnartificialneuralnetworks.com/. [Accessed 18th January 2012].

Ross, S. M. (2010). *Introductory Statistics* (3rd ed.). Academic Press, Elsevier.
Rosner, B. (2006). *Fundamentals of Biostatistics* (6th ed.). California. Thomson Higher Education.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517.

Saritas, I., Ozkan, I. A., & Sert, I. U. (2010). Prognosis of prostate cancer by artificial neural networks. *Expert Systems with Applications, 37*, 6646–6650.

Sarle, W. (2010) What are cross-validation and bootstrapping? Available from: http://www.faqs.org/faqs/ai-faq/neural-nets/part3/section-12.html. [Accessed 8th June 2010].

Scott, G. P. T., Shah, P., Wyatt, J. C., Makubate, B., & Cross, F. W. (2011). Making electronic prescribing alerts more effective: scenario-based exerimental study in junior doctors. *Journal of the American Medical Informatics Association, 18*(6), 789-798.

Sebastiani, P., Yu, Y. H., & Ramoni, M. F. (2003). Bayesian Machine Learning and Its Potential Applications to the Genomic Study of Oral Oncology. *Adv. Dent Res, 17*, 104-108.

Seker, H., Odetayo, M., Petrovic, D., Naguib, R. N. G., Bartoli, C., Alasio, L., et al. (2000). *A Fuzzy Measurement-Based Assessment of Breast Cancer Prognostic Markers.* Paper presented at the Proceedings of IEEE EMBS International Conference on Information Technology Applications in Biomedicine, 2000.

Seker, H., Odetayo, M. O., Petrovic, D., & Naguib, R. N. G. (2003). A Fuzzy Logic Based-Method for Prognostic Decision Making in Breast and Prostate Cancers. *IEEE Transactions on Information Technology in Biomedicine, 7*(2), 114-122.

Siow-Wee, C., Abdul-Kareem, S., Zain, R. B., & Merican, A. F. M. A. (2010). *A Bootstrap-ANFIS Framework for Oral Cancer Prognosis Based on Clinical and Genomic Markers.* Paper presented at the 3rd IEEE International Conference on Computer Science and Information Technology, Chengdu, China.

Siow-Wee, C., Kareem, S. A., Kallarakkal, T. G., Merican, A. F., Abraham, M. T., & Zain, R. B. (2011). Feature Selection Methods for Optimizing Clinicopathologic Input Variables in Oral Cancer Prognosis. *Asia Pacific Journal of Cancer Prevention, 12(10)*, 2659-2664.

Sivaraksa, M., Herzallah, R., & Lowe, D. (2008). Unclassifiability in Medical Prognosis: An Example Using Breast Cancer Gene Marker.

Society, A. C. (2010). What are oral cavity and oropharyngeal cancers? Available from: http://www.cancer.org/Cancer/OralCavityandOropharyngealCancer/DetailedGuide/oral-cavity-and-oropharyngeal-cancer-what-is-oral-cavity-cancer. [Accessed 23rd August 2010].

Song, H. J., Lee, S. G., & Park, G. T. (2005). *A Methodology of Computer Aided Diagnostic System on Breast Cancer.* Paper presented at the IEEE Conference on Control Applications, Torronto, Canada.

Speight, P. M., & Hammond, P. (2001). The Use of Machine Learning in Screening for Oral Cancer *Artificial Neural Networks in Cancer Diagnosis, Prognosis, and Patient Management* (pp. 55-72): CRC Press.

Spiegelhalter, D.J. (1983). Evaluation of clinical decision-aids, with an application to a system for dyspepsia. Statistics in Medicine, 2, 207-216.

Sun, Y., Goodison, S., Li, J., Liu, L., & Farmerie, W. (2007). Improved Breast Cancer Prognosis Through the Combination of Clinical and Genetic Markers. *Bioinformatics, 23*(1), 30-37.

Sunitha, C., & Gabriel, R. (2004). Oral Cancer At A Glance. *The Internet Journal of Dental Science, 1*(2).

Talavera, L. (2005). *An evaluation of filter and wrapper methods for feature selection in categorical clustering.* Paper presented at the 6th International Symposium on Intelligent Data Analysis, Madrid, Spain.

Tan, M. N., Mohd Ghazali, S. H., Tan, C. E., Ghani, W. M. N., & Zain, R. B. (2005). Habitual and Nutritional Factors for Oral Cancer among Malaysians. . *J Dent Res 84 (Special Issue B), 028(SEA) (www.dentalresearch.org)*.

Teni, T., Pawar, S., Sanghvi, V., & Saranath, D. (2002). Expression of Bcl-2 and Bax In Chewing Tobacco-Induced Oral Cancers and Oral Lesions from India. *Pathology Oncology Research, 8*(2), 109-114.

Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2008). Breast Cancer Survivability via AdaBoost Algorithms. *Proc. 2nd Australasian Workshop on Health Data and Knowledge Management*, 55-64.

Thurfjell, N., Coates, P. J., Boldrup, L., Lindgren, B., Bäcklund, B., Uusitalo, T., et al. (2005). Function and Importance of p63 in Normal Oral Mucosa and Squamous Cell Carcinoma of the Head and Neck. *Current Research in Head and Neck Cancer, 62*, 49-57.

Walker, D., Boey, G., & McDonald, L. (2003). The pathology of oral cancer. *Pathology, 35*(5), 376–383.

Wasson, J.H., Harold, C.S., Raymond, K.N., and Goldman, L. (1985). Clinical Prediction Rules — Applications and Methodological Standards. *The New England Journal of Medicine*, 313, 793-799.

What you need to know about Oral Cancer. (2004) Available from: http://www.cancer.gov/cancertopics/wyntk/oral/page2. [Accessed 3rd June 2010].

Williams, J., Carlson, G., Cohen, C., Derose, P., Hunter, S., & Jurkiewicz, M. (1994). Tumor angiogenesis as a prognostic factor in oral cavity tumors. *The American Journal of Surgery, 168*, 373–380.

W.H.O. (2008) Are the number of cancer cases increasing or decreasing in the world? Available from: http://www.who.int/features/qa/15/en/index.html. [Accessed 20th May 2010].

Wyatt, J. C., & Altman, D. G. (1995). Prognostic Models: Clinically useful of quickly forgotten? *British Medical Journal, 311*, 1539-1541.

Xu, R., Cai, X., & Wunsch II, D. C. (2005). *Gene Expression Data for DLBCL Cancer Survival Prediction with A Combination of Machine Learning Technologies.* Paper presented at the Proceedings of the 2005 IEEE Engineering in Medicine and Biology.

Yu, C. H. (2003). Resampling methods: Concepts, Applications, and Justification. *Practical Assessment, Research & Evaluation, 8*(19).

Zain, R. B., & Ghazali, N. (2001). A Review of Epidemiological Studies of Oral Cancer and Precancer in Malaysia. *Annals of Dentistry University of Malaya 8*, 50-56.

Zain, R. B., & George, T. (2009). [Personal Communication. Department of Oral Pathology and Oral Medicine and Periodontology, Faculty of Dentistry, University of Malaya].

Zain, R. B., Ghani, W. M. N., Razak, I. A., Latifah, R. J. R., Samsuddin, A. R., Cheong, S. C., et al. (2009). Building Partnership in Oral Cancer Research in a DevelopingCountry - Processes and Barriers. *Asian Pacific Journal of Cancer Prevention, 10*, 513-518.

Zhang, Y. X. (2007). Artificial neural networks based on principal component analysis input selection for clinical pattern recognition analysis. *The International Journal of Pure and Applied Analytical Chemistry (Talanta), 73*(2007), 68-75.

Zhong, L., Ma, C.-Y., Zhang, H., Yang, L.-J., Wan, H.-L., Xie, Q.-Q., et al. (2011). A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA–CG–SVM method. *Computers in Biology and Medicine, 41*, 1006-1013.

Zigeuner, R., Tsybrovskyy, O., Ratschek, M., Rehak, P., Lipsky, K., & Langner, C. (2004). Prognostic Impact of p63 and p53 in Upper Urinary Tract Transitional Cell Carcinoma. *Adult Urology, 63*(6), 1079-1083.

LIST OF PUBLICATIONS RELATED TO THIS RESEARCH

[1] Siow-Wee, C., Abdul-Kareem, S., Zain, R. B., & Merican, A. F. (2010). A Bootstrap-ANFIS Framework for Oral Cancer Prognosis Based on Clinical and Genomic Markers. Paper presented at the 3rd IEEE International Conference on Computer Science and Information Technology, Chengdu, China. (ISI Conference Proceeding)

[2] Siow-Wee, C., Abdul-Kareem, S., Kallarakkal, T.G., Merican, A.F., Abraham, M.T., Zain, R.B. (2011). Feature Selection Methods for Optimizing Input Variables in Oral Cancer Prognosis. Asia Pacific Journal of Cancer Prevention, 12, 2659-2664. (ISI Publication)

[3] Siow-Wee, C., Abdul-Kareem, S., Merican, A.F., Zain, R.B. (2013). A Hybrid Prognostic Model for Oral Cancer based on Clinicopathologic and Genomic Markers, Sains Malaysiana. (ISI Publication) (Submitted)

[4] Siow-Wee, C., Abdul-Kareem, S., Merican, A.F., Zain, R. B. (2013). Oral Cancer Prognosis Based on Clinicopathologic and Genomic Markers Using a Hybrid of Feature Selection and Machine Learning Methods, BMC Bioinformatics (ISI Publication) (Submitted)

APPENDIX

(a) Oral Cancer Prognosis Dataset

No.	**Location**	**Age**	**Date of 1st diagnosis**	**Ethnicity**	**Gender**	**Smoke**
1	TMaA 1-1,4-3	67	190706	INDIAN	F	No
2	TMaA 1-1,2-7	73	181105	INDIAN	F	No
3	TMaA 1-2,2-9	59	210704	CHINESE	F	Yes
4	TMaA 1-2,2-8	60	70206	MALAY	F	No
5	TMaA 1-3,2-9	59	160106	INDIAN	F	No
6	TMaA 1-8,4-2	59	70405	INDIAN	F	No
7	TMaA 1-4,4-3	51	230106	INDIAN	F	No
8	TMaA 1-5,2-7,2-10,3-2	48	310506	INDIAN	F	No
9	TMaA 1-6,2-7,4-3	70	240804	INDIAN	F	No
10	TMaA 1-7,4-1	55	61003	INDIAN	M	Yes
11	TMaA 1-7,3-1,4-1	63	151003	INDIAN	F	No
12	TMaA 1-8,3-1,4-2	64	291204	INDIAN	F	No
13	TMaA 1-8,2-8,4-4	64	140105	INDIAN	F	No
14	TMaA 3-1	70	10306	INDIAN	F	No
15	TMaA 3-2	71	310306	INDIAN	F	No
16	TMaA 1-2,3-2,4-1	66	30505	INDIAN	F	No
17	TMaA 3-2	54	221204	MALAY	M	Yes
18	TMaA 3-3	57	111006	INDIAN	F	No
19	TMaA 3-3	41	90804	MALAY	M	Yes
20	TMaA 3-4	79	90404	INDIAN	F	No
21	TMaA 3-4,4-1	50	20904	INDIAN	F	No
22	TMaA 8-3A	48	280307	MALAY	F	No
23	TMaA 8-3A	66	Apr-04	CHINESE	M	No
24	TMaA 1-6,4-2,8-5A	66	280104	INDIAN	F	No
25	TMaA 8-5A	69	141107	INDIAN	F	No
26	TMaA 1-4	48	140703	INDIAN	M	No
27	TMaA 1-3	62	NA	MALAY	F	No
28	TMaA 2-9	48	201204	INDIAN	F	No
29	TMaA 3-1	73	NA	CHINESE	M	No
30	TMaA 8-3a	58	250406	INDIAN	F	No
31	TMaA 8-3a	64	51005	INDIAN	M	Yes

No.	**Drink**	**Chew**	**Site**	**Diagnosis & Grading**	**Invasion**
1	Yes	Yes	Buccal	OSCC-Well Differentiated	non cohesive
2	No	Yes	Buccal	OSCC- Moderately Differentiated	non cohesive
3	No	No	Buccal	OSCC- Moderately Differentiated	non cohesive
4	No	Yes	Buccal	OSCC- Moderately Differentiated	non cohesive
5	No	Yes	Buccal	OSCC-Moderately Differentiated	cohesive
6	No	Yes	Others	OSCC- Well Differentiated	non cohesive
7	No	Yes	Buccal	OSCC-Well Differentiated	cohesive
8	No	Yes	Buccal	OSCC- Moderately Differentiated	cohesive
9	No	Yes	Buccal	OSCC-moderately-differentiated	non cohesive
10	No	Yes	Buccal	OSCC-moderately differentiated	non cohesive
11	No	Yes	Others	OSCC-moderately differentiated	non cohesive
12	No	Yes	Others	OSCC-well-differentiated	non cohesive
13	No	Yes	Others	OSCC- Moderately Differentiated	non cohesive
14	No	Yes	Buccal	OSCC-Well Differentiated	non cohesive
15	No	Yes	Buccal	OSCC-Well Differentiated	non cohesive
16	No	Yes	Others	OSCC-well-differentiated	cohesive
17	No	No	Others	OSCC-well-differentiated	non cohesive
18	No	Yes	Floor	OSCC-Well Differentiated	Non cohesive
19	No	No	Buccal	OSCC-Moderate Differentiated	Non cohesive
20	Yes	Yes	Others	OSCC-Well Differentiated	Non cohesive
21	No	Yes	Buccal	OSCC- Moderately Differentiated	Non cohesive
22	No	No	Tongue	OSCC-Moderately Differentiated	Non cohesive
23	Yes	No	Tongue	OSCC-moderately-differentiated	non cohesive
24	No	Yes	Others	OSCC-basaloid (mod-diff.)	non cohesive
25	Yes	Yes	Buccal	OSCC-Moderately Differentiated	non cohesive
26	No	Yes	Buccal	OSCC-Well Differentiated	cohesive
27	No	Yes	Buccal	OSCC-Moderately Differentiated	noncohesive
28	Yes	Yes	Others	OSCC-moderately-differentiated	cohesive
29	No	Yes	Floor	OSCC-Moderately Differentiated	Non-cohesive
30	No	Yes	Tongue	OSCC-Moderately Differentiated	Non cohesive
31	Yes	No	Floor	OSCC-basaloid (mod-diff.)	cohesive

No.	**Nodes**	**PT**	**PN**	**Stage**	**Size**	**Treatment**
1	POSITIVE	T1	N2b	IV	>6	Surgery, post-op radiotherapy
2	NEGATIVE	T2	N0	II	>4 – 6	Surgery, post-op radiotherapy
3	NEGATIVE	T2	N0	II	>4 – 6	Surgery, post-op radiotherapy
4	NEGATIVE	T2	N0	II	>2 – 4	Surgery, post-op chemotherapy
5	POSITIVE	T2	N2b	IV	>2 – 4	Surgery, post-op radiotherapy
6	NEGATIVE	T1	N0	I	0 – 2	Surgery alone
7	NEGATIVE	T1	N0	I	>6	Surgery alone
8	NEGATIVE	T4	N0	IV	>4 – 6	Surgery, post-op radiotherapy
9	NEGATIVE	T3	N0	III	>4 – 6	Surgery, post-op radiotherapy
10	NEGATIVE	T1	N0	I	>2 – 4	Surgery alone
11	POSITIVE	T2	N2b	IV	>4 – 6	Surgery, post-op radiotherapy
12	NEGATIVE	T4	N0	IV	>4 – 6	Surgery, post-op radiotherapy
13	POSITIVE	T4	N0	IV	>2 – 4	Surgery, post-op radiotherapy
14	POSITIVE	T4	N2b	IV	>6	Pre-op radiotherapy, surgery
15	POSITIVE	T3	N2b	IV	>4 – 6	Surgery alone
16	NEGATIVE	T4	N0	IV	>4 – 6	Surgery, post-op radiotherapy
17	NEGATIVE	T4	N0	IV	>4 – 6	Surgery, post-op radiotherapy
18	POSITIVE	T3	N2a	IV	0 – 2	Surgery alone
19	POSITIVE	T2	N2b	IV	0 – 2	Surgery, post-op chemotherapy
20	POSITIVE	T4	N0	IV	>6	Surgery, post-op radiotherapy
21	POSITIVE	T4	N1	IV	>6	Surgery, post-op radiotherapy
22	NEGATIVE	T3	N0	III	>2 – 4	Surgery alone
23	NEGATIVE	T3	N0	III	>2 – 4	Surgery, post-op radiotherapy
24	POSITIVE	T4	N2b	IV	>6	Surgery, post-op radiotherapy
25	NEGATIVE	T4	N0	IV	0 – 2	Surgery alone
26	NEGATIVE	2	0	II	0 – 2	Surgery, post-op radiotherapy
27	NEGATIVE	4	0	IV	>4-6	Surgery + radiotheraphy
28	POSITIVE	4	1	IV	>4 – 6	Surgery, post-op chemotherapy
29	NEGATIVE	4	0	IV	>4-6	Surgery + radiotherapy
30	POSITIVE	T3	2b	IV	0 – 2	Surgery alone
31	POSITIVE	4	1	IV	>4 – 6	Surgery, post-op chemotherapy

No.	**p53**	**p63**	**1-year survival**	**2-year survival**	**3-year survival**
1	neg	neg	Survive	Lost	Lost
2	neg	pos	Survive	Survive	Lost
3	pos	neg	Survive	Dead	Dead
4	neg	pos	Survive	Dead	Dead
5	pos	neg	Survive	Survive	survive
6	neg	pos	Survive	Survive	survive
7	pos	pos	Survive	Survive	survive
8	pos	neg	Survive	Survive	survive
9	neg	pos	Survive	Survive	survive
10	neg	pos	Survive	Survive	survive
11	pos	pos	Survive	Dead	Dead
12	pos	neg	Dead	Dead	Dead
13	neg	pos	Survive	Survive	survive
14	neg	neg	Dead	Dead	Dead
15	neg	neg	Dead	Dead	Dead
16	neg	pos	Survive	Lost	Lost
17	neg	pos	Survive	Survive	survive
18	neg	neg	Dead	Dead	Dead
19	pos	pos	Survive	Survive	survive
20	neg	neg	Survive	Survive	survive
21	neg	pos	Survive	Survive	survive
22	pos	pos	Survive	Survive	survive
23	pos	pos	Survive	Dead	Dead
24	neg	pos	Survive	Survive	survive
25	neg	pos	Survive	Dead	Dead
26	neg	pos	Survive	Survive	survive
27	neg	pos	Survive	Survive	survive
28	neg	neg	Survive	Survive	survive
29	pos	pos	Survive	Survive	survive
30	neg	neg	Survive	Dead	Dead
31	pos	pos	Survive	Survive	Dead

(b) Categorization for Oral Cancer Prognosis Dataset

Categorization for Oral Cancer Prognosis Dataset

Feature	Codes/Value
Gender	1 - Male 2 - Female
Ethnicity	1 - Malay 2 - Chinese 3 - Indian
Age	1 - 40-50 2 - >50-60 3 - >60-70 4 - >70
Smoke	1 - No 2 - Yes
Drink	1 - No 2 - Yes
Chew	1 - No 2 - Yes
Site	1 - Buccal mucosa 2 - Tongue 3 - Floor 4 - Others
Size	1 - 0-2cm 2 - >2-4cm 3 - >4-6cm 4 - >6cm
Invasion	1 - Non-cohesive 2 - Cohesive

Feature	Codes/Value
Nodes	1 - Negative 2 - Positive
Subtype	1 - Well differentiated 2 - Moderate differentiated 3 - Poorly differentiated
PT	1 - T1 2 - T2 3 - T3 4 - T4
PN	1 - N0 2 - N1 3 - N2A 4 - N2B
Stage	1 - Stage I 2 - Stage II 3 - Stage III 4 - Stage IV
Treatment	1 - Surgery only 2 - Surgery + Radiotherapy 3 - Surgery + Chemotherapy
p53	1-Nagative 2-Positive
p63	1-Nagative 2-Positive
Survive	-1 - Survive 1 - Dead

(c) Membership functions for input variables

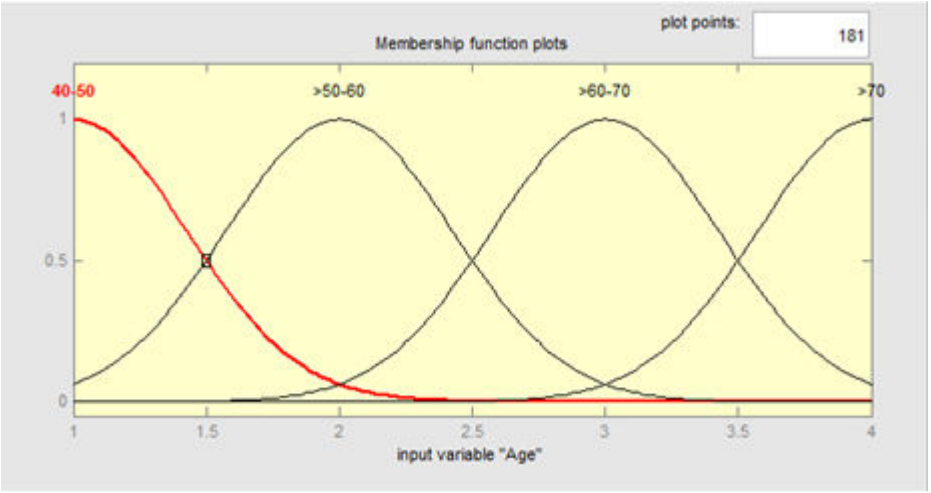


Figure B1: Membership functions for input variable "Age"

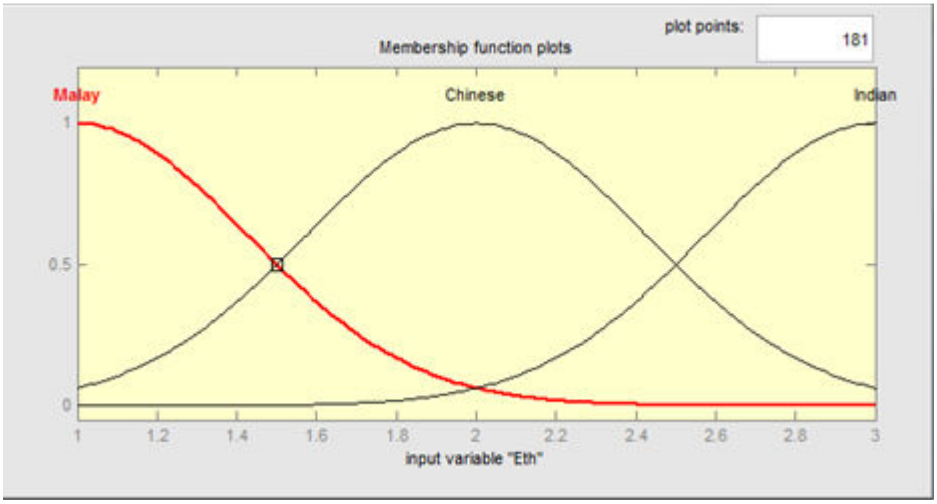


Figure B2: Membership functions for input variable "Eth"

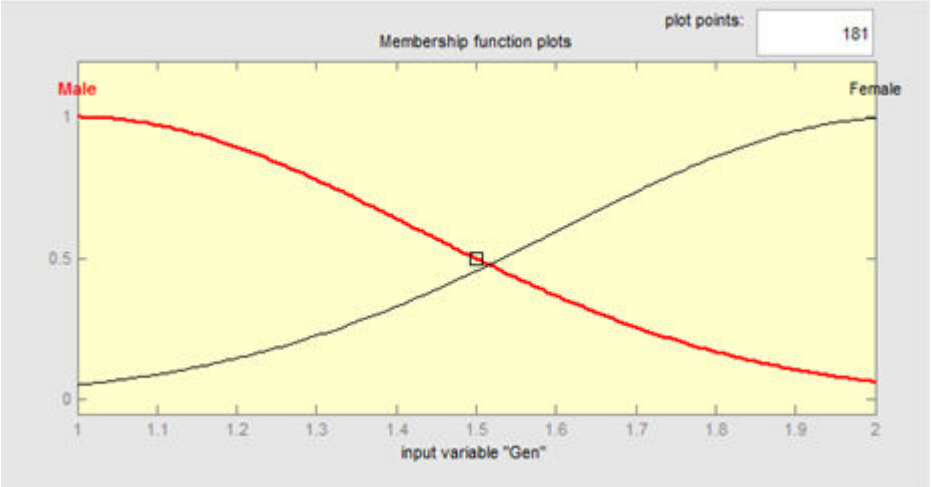


Figure B3: Membership functions for input variable "Gen"

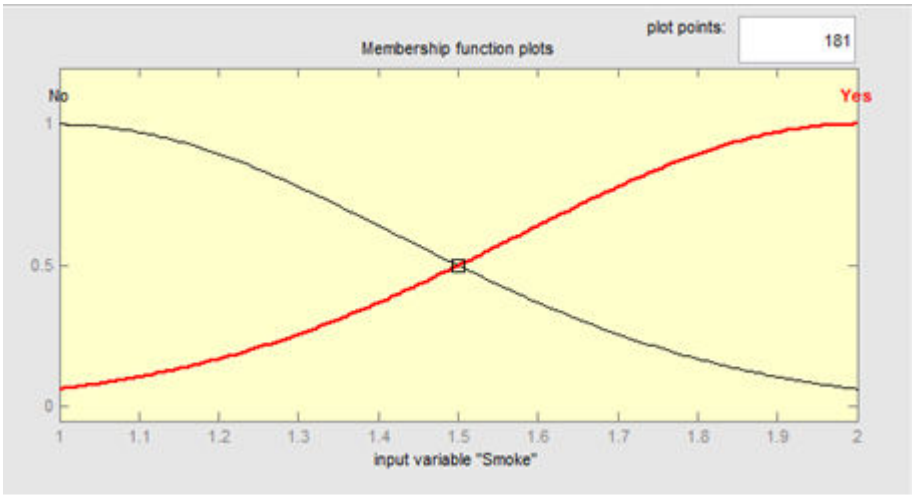


Figure B4: Membership functions for input variable "Smoke"

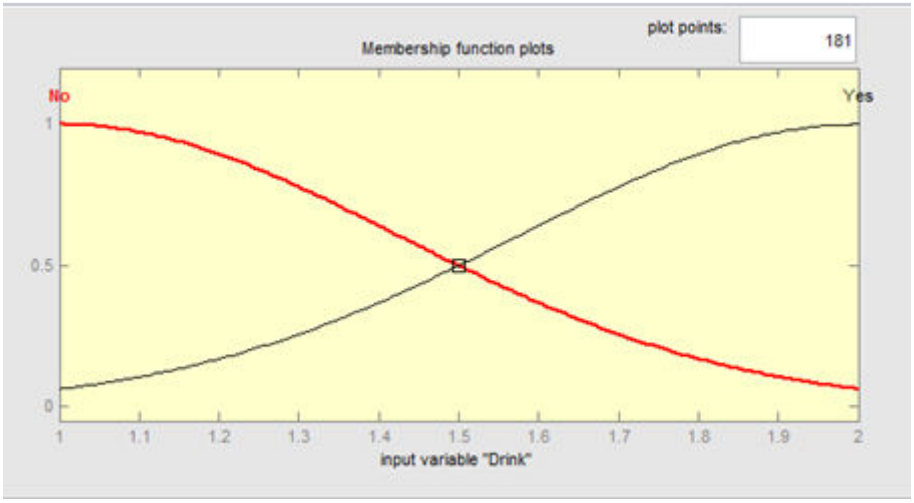


Figure B5: Membership functions for input variable "Drink"

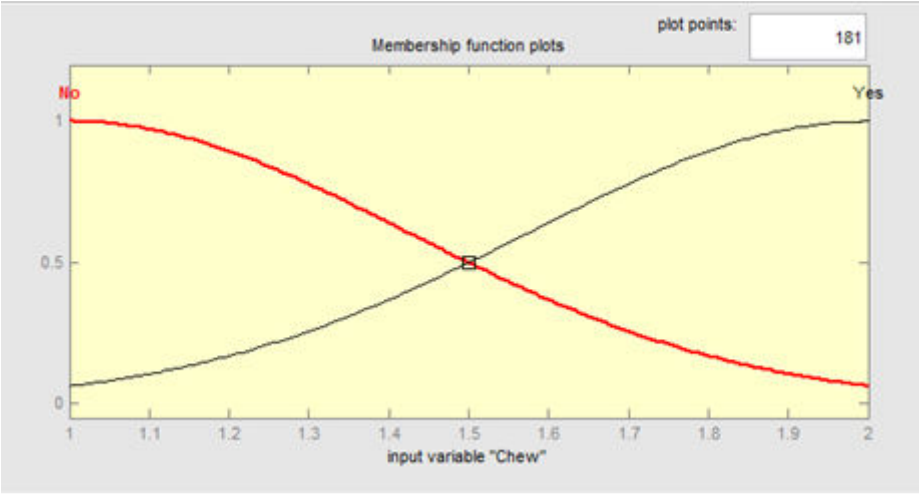


Figure B6: Membership functions for input variable "Chew"

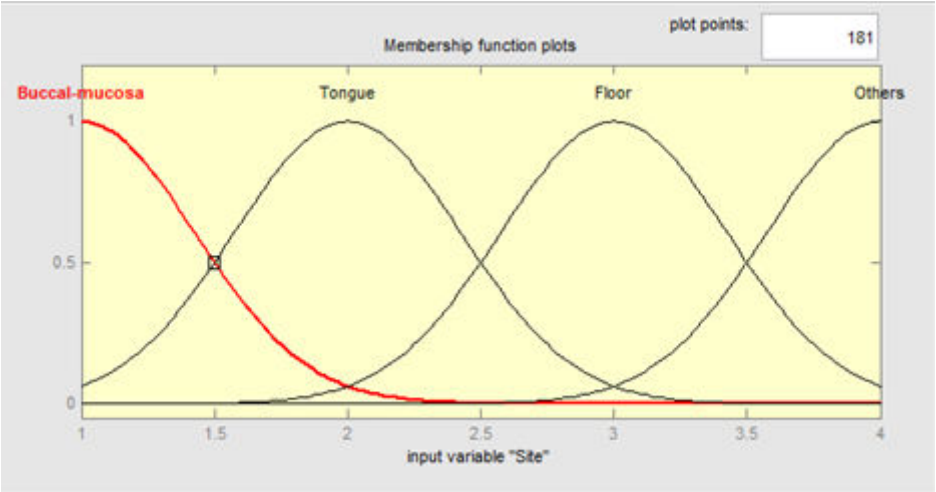


Figure B7: Membership functions for input variable "Site"

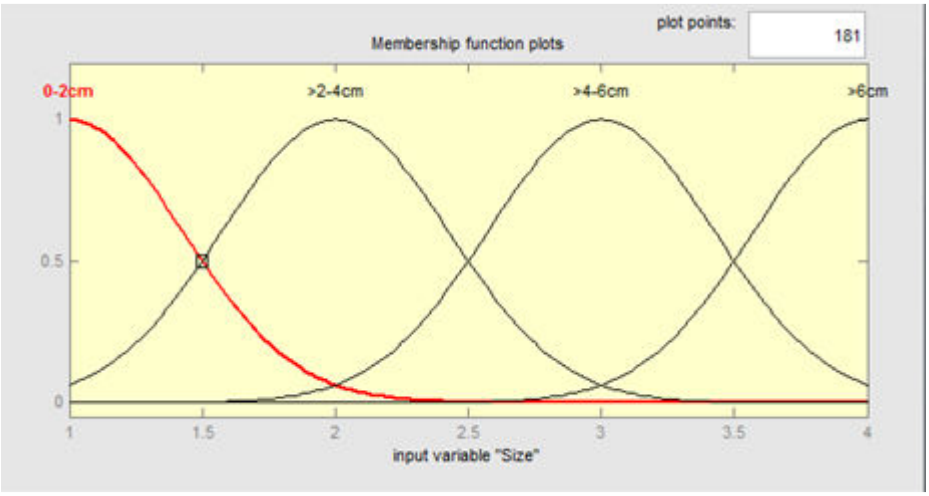


Figure B8: Membership functions for input variable "Size"

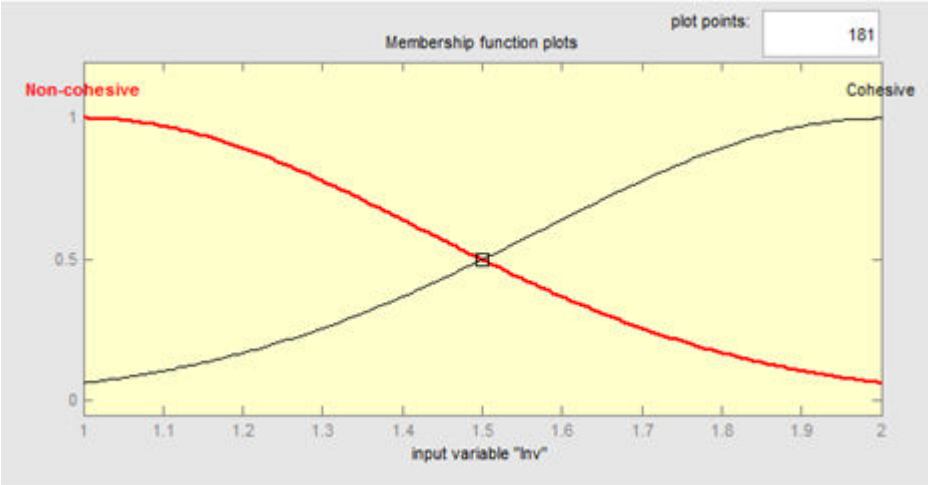


Figure B9: Membership functions for input variable "Inv"

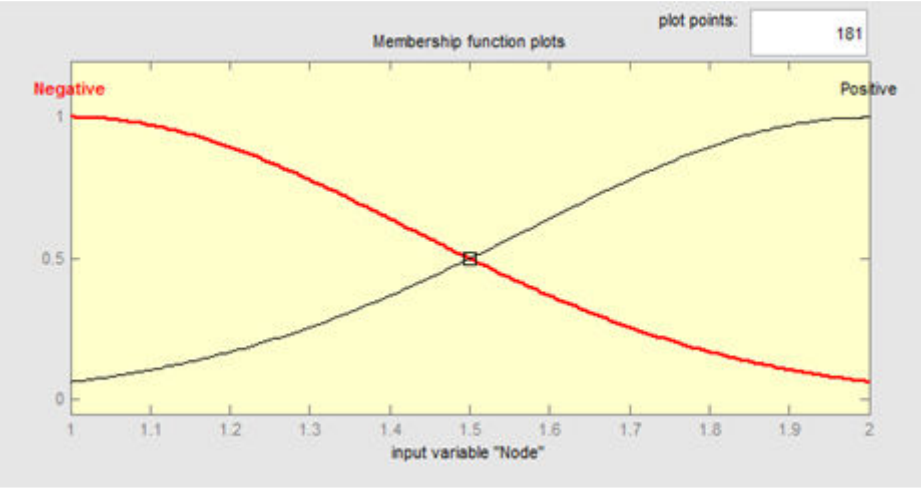


Figure B10: Membership functions for input variable "Node"

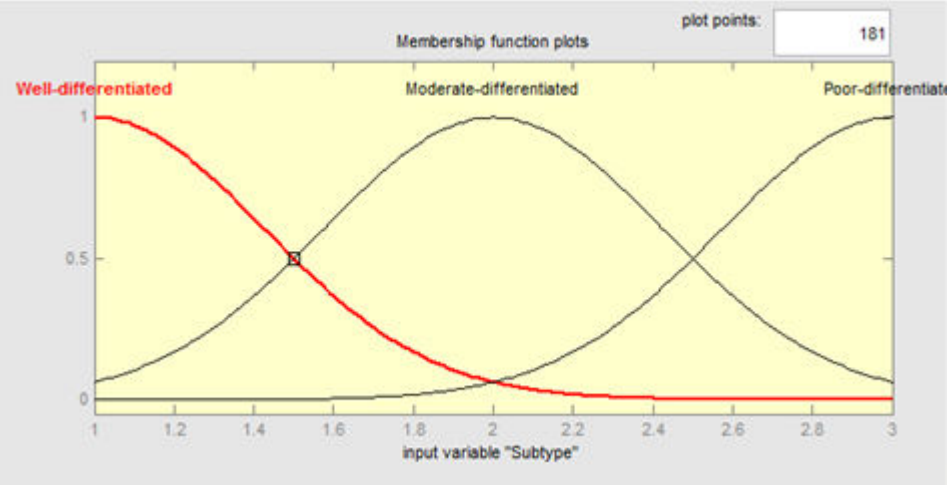


Figure B11: Membership functions for input variable "Subtype"

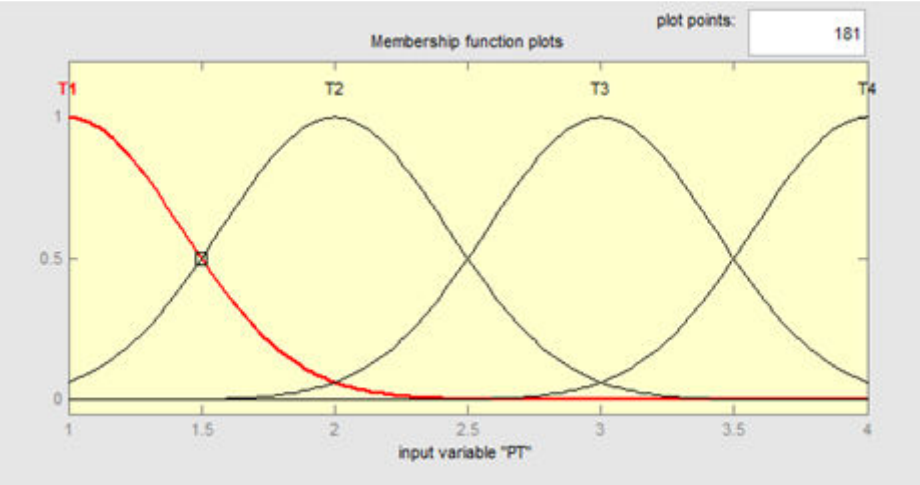


Figure B12: Membership functions for input variable "PT"

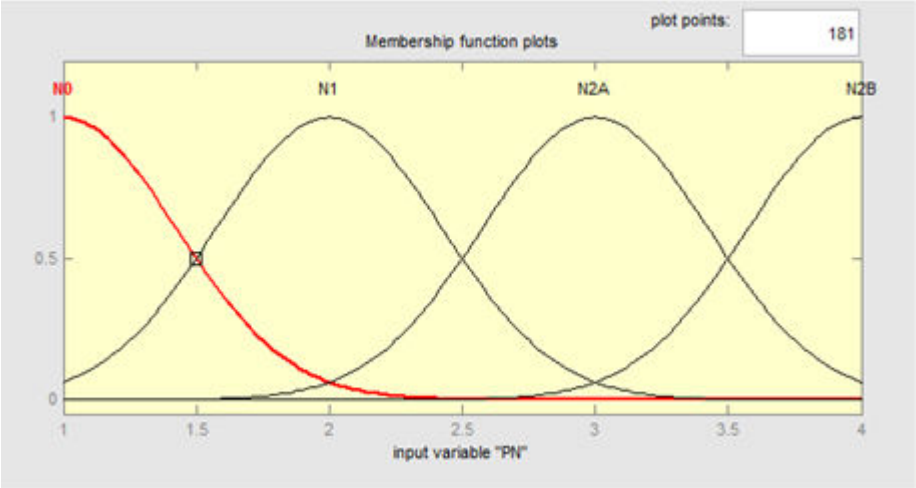


Figure B13: Membership functions for input variable "PN"

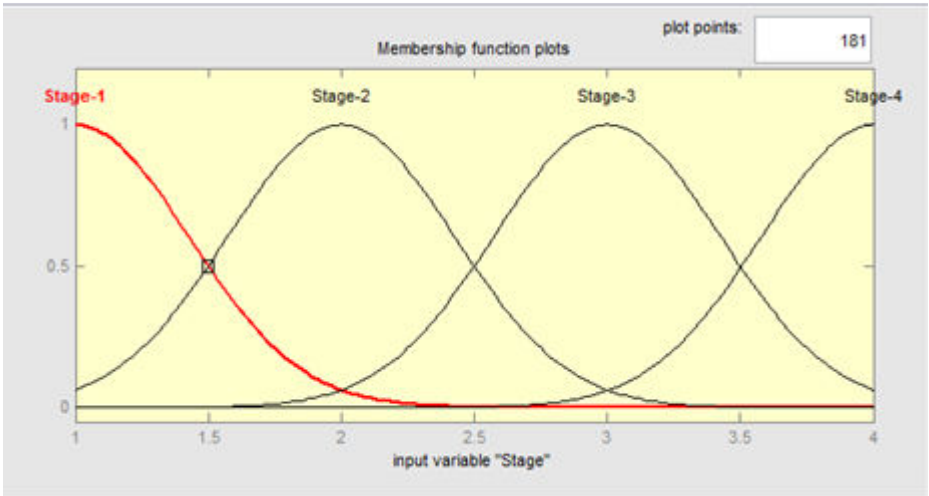


Figure B14: Membership functions for input variable "Stage"

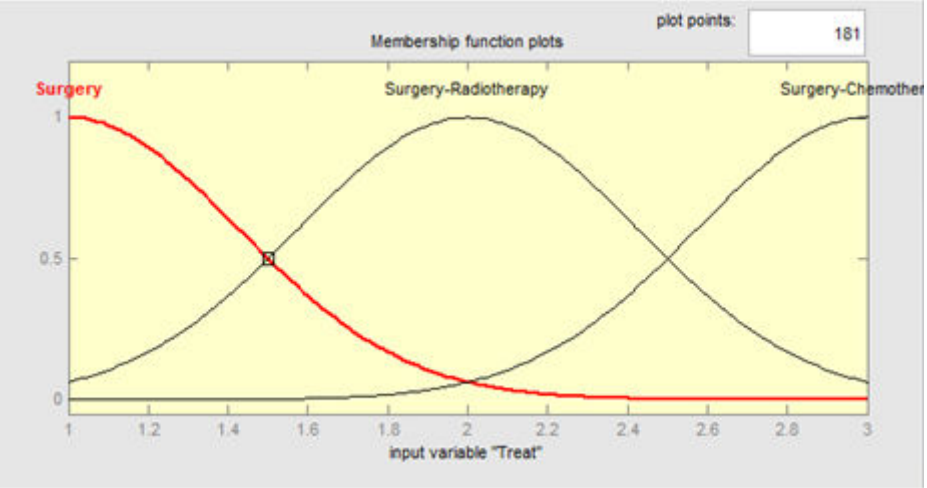


Figure B15: Membership functions for input variable "Treat"

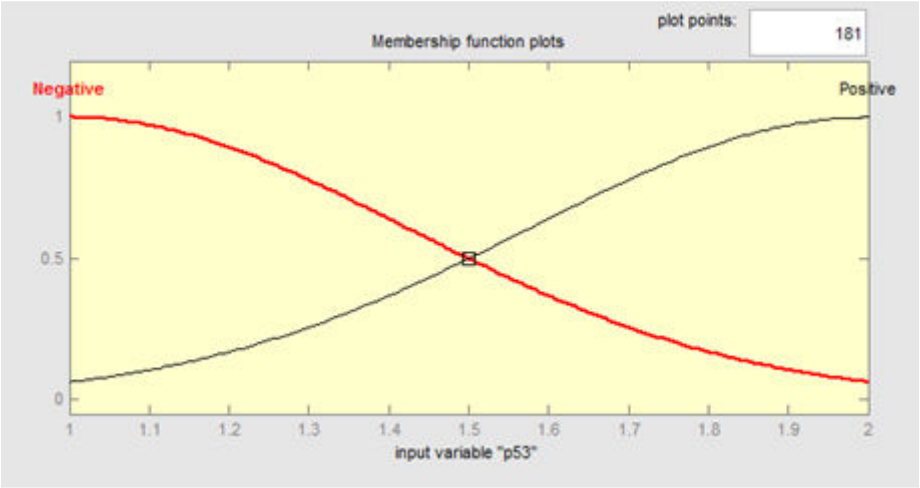


Figure B16: Membership functions for input variable "p53"

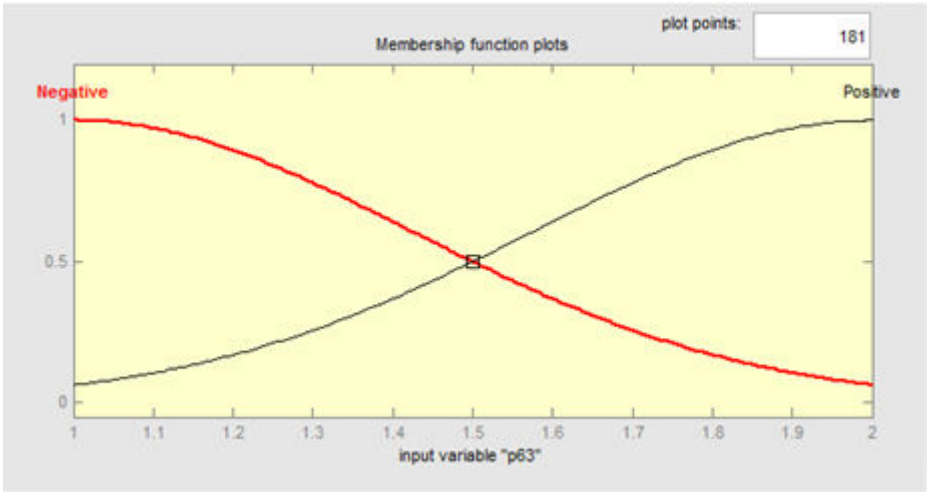


Figure B17: Membership functions for input variable "p63"

(d) Model Validation Study for Oral Cancer Clinicians

Introduction:

Artificial intelligent (AI) techniques are becoming useful as an alternative approach to conventional medical diagnosis or prognosis. AI techniques are good for handling noisy and incomplete data, and significant results can be attained despite the small sample size. Various AI techniques have been applied in medical research such as artificial neural networks, fuzzy logic, genetic algorithm and other hybrid methods. The main aim of this research is to apply AI techniques in the prognosis of oral cancer based on the parameters of the correlation of clinicopathologic and genomic markers. To this end, a hybrid AI model for oral cancer prognosis is developed. In order to validate the developed AI prognostic model, a validation study which involved oral cancer clinicians is required.

The objectives of this validation exercise were:

(i) To measure the prediction accuracy of human expert predictions

(ii) To measure the prediction consistency of human expert predictions

Researcher:
Chang Siow Wee
Faculty of Computer Science and Information Technology
University of Malaya
Email: changsiowwee@yahoo.com , siowwee@um.edu.my
H/P: 016-3852738

Co-researcher:

1. Assoc. Prof. Datin Dr. Sameem Abdul Kareem
   Faculty of Computer Science and Information Technology, UM.
2. Assoc. Prof. Amir Feisal Merican Aljunid Merican
   Institute of Biological Science, Faculty of Science, UM.
3. Prof. Rosnah Binti Mohd Zain
   Department of Oral Pathology and Oral Medicine and Periodontology,
   Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, UM.

Instruction: Respondents are required to make a 3-year prognosis for oral cancer based on the selected variables. That is, clinicians are supposed to predict whether an individual is having poor prognosis or better prognosis.

This exercise contains Sections A, and B. Please indicate your responses in the indicated space provided.

Variable descriptions:

Clinicopathologic variables:

	Name	Description
1.	Age	Age at diagnosis
2.	Eth	Ethnicity
3.	Gen	Gender
4.	Smoke	Smoking habit
5.	Drink	Alcohol drinking habit
6.	Chew	Quid chewing habit
7.	Site	Primary site of tumor
8.	Subtype	Subtype and differentiation for SCC
9.	Inv	Pattern of Invasion front
10.	Node	Neck nodes
11.	PT	Pathological tumor staging
12.	PN	Pathological lymph nodes
13.	Stage	Overall stage
14.	Size	Size of tumor
15.	Treat	Type of treatment

Section A

1. From the list of clinicopathologic variables listed above, please choose:
   Four (4) most significant variables for oral cancer prognosis (Please rank accordingly)

   i. _____

   ii. _____

   iii. _____

   iv. _____

2. In your opinion, would the inclusion of the genomic markers improve the accuracy of oral cancer prognosis?

   Yes ☐ No ☐

   Comments/Reasons:

   _____

   _____

   _____

Section B

Model 1: Based on the combination of three (3) clinicopathologic variables below, please indicate your prognosis.

Gender (Male/ Female)	**Smoking (Yes/No)**	**PN (N0/N1/ N2A/N2B)**	**3-year Prognosis (P=Poor prognosis/ B=Better prognosis)**
Female	No	N0	
Female	No	N1	
Female	No	N2A	
Female	No	N2B	
Female	Yes	N0	
Male	No	N0	
Male	Yes	N0	
Male	Yes	N1	
Male	Yes	N2B	

Model 2: Based on the combination of four (4) clinicopathologic variables below, please indicate your prognosis.

Age (40-50/ >50-60/ >60-70/ >70)	**Gender (Male/ Female)**	**Pattern of Invasion (Non-cohesive/cohesive)**	**PN (N0/N1/ N2A/N2B)**	**3-year Prognosis (P=Poor prognosis/ B=Better prognosis)**
40-50	Female	Cohesive	N0	
40-50	Female	Non-cohesive	N1	
40-50	Male	Cohesive	N0	
40-50	Male	Non-cohesive	N2B	
>50-60	Female	Cohesive	N0	
>50-60	Female	Cohesive	N2B	
>50-60	Female	Non-cohesive	N0	
>50-60	Female	Non-cohesive	N2A	
>50-60	Female	Non-cohesive	N2B	
>50-60	Male	Non-cohesive	N0	
>60-70	Female	Cohesive	N0	
>60-70	Female	Non-cohesive	N0	
>60-70	Female	Non-cohesive	N2B	
>60-70	Male	Non-cohesive	N0	
>60-70	Male	Non-cohesive	N1	
>70	Female	Non-cohesive	N0	
>70	Female	Non-cohesive	N2B	
>70	Male	Non-cohesive	N0	

Model 3: Based on the combination of six (6) clinicopathologic variables below, please indicate your prognosis.

Gender (Male/ Female)	**Drink (Yes/No)**	**Nodes (Negative/ Positive)**	**PT (T1/T2/ T3/T4)**	**PN (N0/N1/ N2A/N2B)**	**Stage (I/II/ III/IV)**	**3-year Prognosis (P=Poor prognosis/ B=Better prognosis)**
Female	No	Negative	T1	N0	I	
Female	No	Negative	T2	N0	II	
Female	No	Negative	T3	N0	III	
Female	No	Negative	T4	N0	IV	
Female	No	Positive	T2	N2b	IV	
Female	No	Positive	T3	N2a	IV	
Female	No	Positive	T3	N2b	IV	
Female	No	Positive	T4	N0	IV	
Female	No	Positive	T4	N1	IV	
Female	No	Positive	T4	N2b	IV	
Female	Yes	Negative	T4	N0	IV	
Female	Yes	Positive	T1	N2b	IV	
Female	Yes	Positive	T4	N0	IV	
Female	Yes	Positive	T4	N1	IV	
Male	No	Negative	T1	N0	I	
Male	No	Negative	T2	N0	II	
Male	No	Negative	T4	N0	IV	
Male	No	Positive	T2	N2b	IV	
Male	Yes	Negative	T3	N0	III	
Male	Yes	Positive	T4	N1	IV	

~ The End ~

THANK YOU.