

# Chapter 1

## INTRODUCTION

Information technology (IT) plays an important role in many fields such as astronomy, engineering, construction, business, aeronautics and medicine. Computer applications in the medical field resulted in the birth of medical informatics. Medical informatics is the application of computer technology, communications and information technology and system in medical fields such as medical care, medical education and medical research. The medical information gathered can be shared, analysed, and communicated to other parties for the support of specific medical research, education and quality patient care to enable the understanding and improve the precision, timeliness and reliability of decision-making by experts in their own right (Shortliffe, 1995; Collen et al., 1990; Warner et al., 1997).

In Malaysia, medical informatics is a relatively young field. To ensure that the nation does not lag behind in the medical field, the Malaysian government has planned and gradually implemented a massive healthcare initiative aimed at building up the necessary infrastructure to empower its own citizens to manage their own health via interconnected online platforms addressing specific areas of daily living such as proper dieting and self-examinations. Telemedicine is one of the areas targeted for intensive development through the Multimedia Super Corridor (MSC) project, which combines medical and healthcare with several application technologies such as artificial intelligence, multimedia communication and information system. The objective of MSC-based telemedicine is to establish a healthcare system leveraging advanced information and multimedia so as to deliver inaccessible services at the individual, family and community level. The MSC telemedicine model was designed to encompass four key pilot projects namely:

- Customised or personalised Health information and education
- Continuing medical education
- Teleconsultation
- Lifetime Health Plan (LHP)

The implementation of these projects will result in the robust application of computer technology in medicine and healthcare. One such example is the complete computerization of the hospital information system, which includes patient registration system, pharmacy system, laboratory system, medical records, and remote consultation. It is envisaged that the existence of such a complicated, flexible, cost-efficient and consumer friendly healthcare system would lay the groundwork of a healthy nation with healthy individuals, families and communities banding together to enhance the current way of life (Syed et al., 1998).

### **1.1 Cancer and Prognosis**

Cancer is defined as a group of healthy cells in any part of the body such as liver or prostate, which suddenly and inexplicably begins to divide uncontrollably. Essentially, the inbuilt program of these healthy cells is somewhat triggered by yet to be determined agents, causing them to divide in a manner that defies a specific, normal pattern. Different types of cancer behave and result differently; i.e. different growth patterns, response to treatment regimes, and survival rates (Charles & Buckner, 2002).

Breast cancer is one of the more common cancers to afflict the female population. Breast cancer is a malignant tumour that develops from uncontrolled growth of cells in the breast. A malignant tumour is composed of cells that invade or spread to other parts of the body. The exact cause of the breast cancer is not really known; it is most likely to

be a combination of genetic and environmental factors. Breast cancer is described in greater detail in chapter 2.

Women throughout the world are often disturbed by the thought of being diagnosed with breast cancer. In the year 2000, there were 1,050,346 cases of breast cancer reported with 372,969 deaths from breast cancer worldwide (Globocan, 2000). The number of reported cancer cases in Peninsular Malaysia in the year 2002 was 26,089, comprising 11,815 males and 14,274 females. Breast cancer was the most commonly diagnosed cancer in Malaysian women with 4337 cases reported to the NCR (NCR, 2002). The incidence was estimated at 45.1 per 100,000 with a chance of 1 out of 19 women getting breast cancer in their lifetime.

Breast cancer awareness campaigns amongst women have been carried out in almost every part of the world including Malaysia. Women are encouraged to see their doctors at an early age. They are also encouraged to carry out monthly breast self-examinations to detect any suspicious lumps. Many women feel embarrassed to see a doctor if they discover a lump in their breast and resort to traditional treatment in the hopes of curing their affliction. However, their actions decrease their survival rates as the disease should be diagnosed early to improve their chance of cure.

The diagnosis of cancer is usually carried out through a series of tests, and is confirmed with a biopsy. After the confirmation of cancer, the stage of the cancer is ascertained in order for an effective treatment solution to be drawn up. A combination of treatment methods, namely, surgery, chemotherapy and radiotherapy are usually recommended by clinicians in order to place the cancer in remission and at the same time avoid recurrence.

An individual with cancer who has been diagnosed usually asks the doctor whether the diseases can be cured with treatment. Medical definition of prognosis is the prediction of the future course and the outcome of disease process, which may either concern their natural course of their outcome after treatment (Abu Hanna & Lucas, 2001). Prognosis is the principal factor in determining the treatment that will immediately follow the diagnosis of the disease. Prognosis is important because the type and intensity of the medications are based on it. However, prognosis is only a prediction and like all predictions it is not a hundred percent accurate.

Prognosis of an individual with cancer can be affected by many factors, such as the type of cancer, the cancer stage, how aggressive the cancer is, the individual's age, the effectiveness of treatment methods and others. The aim of this project is to analyse cancer patient's data based on these factors in order to estimate prognosis.

Survival analysis describes the analysis of data that corresponds to the time from when an individual enter a study until the occurrence of some particular event or end-point. It is concerned with the comparison of survival curves for different combinations of risk factors and commonly uses statistical models to facilitate the comparison (Don McNeil, 1996). Statistical analyses are applied to a large group of people and may be meaningless for an individual who may want to plan his treatment options.

## **1.2 Research Objectives and Aim**

The objective of this research is to analyse and thence to predict the survival of breast cancer patients based on recognised prognostic factors and compare the accuracy of survival prognosis using different pre-processing techniques, training algorithms and neural network models.

Entirety, the aim of the research is to show that artificial neural network (ANN) is an effective and efficient tool to predict the survival of individual cases in the domain of breast cancer using based on recognised prognostic factors. This would assist patients and their families to plan their treatment options and budget their finances accordingly.

An ANN is a viable alternative tool for the analysis of survival data, and recently have become very popular in medical survival predictions. The application of ANN in cancer prognosis have been carried out by many researchers, namely, Street (1999), Burke et al. (1997), De Laurentiis and Ravdin (1994), Ravdin and Clark (1992), Ruth M.Ripley (1998) and Abdul-Kareem et al. (2001). A brief description on the previous works in survival analysis using neural network application is given in chapter 3.

### **1.3 Data and Analysis**

This research is based on breast cancer cases collected from the University of Malaya Medical Centre (UMMC), Kuala Lumpur from 1993 to 2001. There are a total of 1001 cases in the dataset, and these have been collected since patients come for their first diagnosis. Variables in the data set include age, race, stage, TNM (Tumour, Nodes and Metastases) classification, lymph node involvement, estrogen receptor, grade, and treatment options. The dataset is further described in chapter 4.

The dataset have to be cleaned of errors and free from redundancy in order to be useful for analysis so as to generate an accurate result. The collection of patients' records may be incomplete and may include erroneous submissions. Thus, incomplete records and records with empty fields need to be removed from the dataset.

Traditionally, the analysis of survival data is done by using common statistical methods, and these have been well reported in medical statistics textbooks and journals (Collet, 1994; David & Stanley, 1998; Lee, 1992). Part of the work done in this project was to carry out the statistical analysis of the breast cancer data set. The experiments and results carried out using Kaplan-Meier model in the domain of breast cancer is discussed in Appendix B. Kaplan Meier Analysis carried out in this research was to confirm the significance of variables such as stage, TNM classification, lymph node involvement, and grade as prognostic indicators.

ANN is a group of intelligence technology utilized exclusively for complex data analysis. The field of ANN was born in an attempt to overcome the limitations of the computer's ability to perform certain tasks. A neural network gathers its knowledge by detecting common patterns and relationships in raw data, then learning from such relationships and adapting the results as per requirement. The patterns reflect will enable it to assign the correct guidelines to new combinations of data. The neural network architecture and its processing is described in greater detail in chapter 3.

Generally, neural networks are applied in many areas such as modelling, pattern recognition, bioelectric signal processing, diagnostics and prognosis (Konel et al., 1997). Research on applications of ANN in medicine have been done since the late 80s as an aid to diagnosis and treatment. The active application of such endeavours can be observed through a number of research papers and journals. The ability of neural networks to generalise to new cases based on existing patterns is used as a basis to compute and predict the survival of individual cases.

In medical research, multi layer perceptrons that use backpropagation training are most commonly used (Burke et al., 1997). Backpropagation networks are networks where signals travel in one direction from input neuron to an output neuron without returning to its source. Backpropagation network consists of at least three layers of units: an input layer, at least one hidden layer and an output layer. Since backpropagation networks have attained good results, we have chosen this network to begin our training of the neural network.

The experiments in this research were conducted using ANN technology to predict the prognosis of breast cancer. Several experiments using different data pre-processing techniques, various training algorithms and survival intervals are done to lend credibility to the entire exercise. The experiments and results of various training algorithms are discussed in chapter 5. A summary of the thesis, the contributions of this project to the field of medical informatics and suggestions for future work are discussed in chapter 6.