CHAPTER 1
INTRODUCTION

This chapter tells about overview of cancer and microarray data, characteristic of microarray data and its classification which become background of the problem, problem statement, objective of this study related to the problem, scope and delimitation used in this study, contribution and significant of this study, and thesis organization.

1.1 Overview of Cancer and Microarray Data

Cancer is a term used for diseases in which abnormal cells divide without control and are able to invade other tissues. Based on WHO cancer fact sheets [1], cancer is a leading cause of death worldwide, accounting for 8.2 million deaths in 2012. Cancer cells can spread to other parts of the body through the blood and lymph systems [1].

Conventional methods for monitoring and diagnosing cancers there are now very dependent on human observation to detect certain features. A cancer diagnosis is usually performed using imaging systems and analysis of morphological and clinical data. In recent decades Microarray take an important role in the diagnosis of cancer and improve the accuracy of cancer diagnosis compared to traditional techniques. By using Microarray can be seen the level of gene expression in specific cell samples to analyze thousands of genes simultaneously [2].

Humans have tens of thousands of genes, and the development of DNA microarrays by Patrick O. Brown, Joseph DeRisi, David Botstein, and colleagues in the mid-1990s made it possible to examine the expression of thousands of genes at once. DNA Microarrays data are glass microscope slides onto which genes are attached at fixed and ordered locations. Each gene sequence is identified by a location of a spot in the array [2]. Using a Microarray printer, the DNA is spotted directly onto the slide. With microarrays, it is possible to examine a gene expression within a single sample or to compare gene expressions within two tissue samples, such as in tumor and non-tumor tissues [2]. In other side, microarray data typically consists thousands of genes with only tens of samples which is very prone to curse of dimensionality.

1.2 Background of Problem

The characteristic of microarray data is small sample but huge dimension. Dimensions of microarray data which obtained from repository are ranged from 2,000 – 24,481. Meanwhile, the number of samples are below 100 except for ovarian cancer data which contains 253 samples and Prostate cancer data which contains 102 samples. Since that, there was a challenge for researcher to provide solutions for Microarray Data classification with high performance both in accuracy and running time.

1.3 Problem Statement

As explained in Chapter 1.1 and 1.2, the characteristic of microarray data is small sample (about hundreds sample) and huge dimension (about thousands dimension). Since that, there is a challenge to provide solutions for Microarray Data classification with better performance especially in term of accuracy.

1.4 Objective

The objectives of this study is to provide a scheme for microarray data classification which has better performance compared to previous research. The scheme addressed the huge dimension problem using a dimension reduction method namely Principal Component Analysis. Classifier used is SVM with kernel function as improvement. The more detailed objectives are:

1. To provide a scheme for microarray data classification with better accuracy compared to previous research specifically for same data set.
2. To analyze the impact of PCA usage on microarray data classification by comparing accuracy and running time with classification without PCA.
3. To analyze the usage of kernel functions on SVM for microarray data classification.

1.5 Scope and Delimitation

This study will study about microarray data classification for cancer detection. Some delimitation which is in this study such as:

a. The data which are used are data from Kent-ridge bio-medical data set repository [3]. The data specification is listed in Table 1-1.

Table 1-1 Data Specification

Cancer Data	Number of		
	Classes	**Samples**	**Features**
Breast	2	78	24,481
Nervous System	2	60	7,129
Colon	2	62	2,000
Lung	2	32	12,533
Leukemia	3	57	12,582
Ovarian	2	253	15,154
Prostate	2	102	12,600

b. The performance of the proposed scheme is measured based on the accuracy and running time.

1.6 Contribution and Significant

Based on the previous works already described, this research is focused on reduction dimension of microarray data classification to support cancer detection. The research will be useful for:

- Medical doctor and laboratory analyst, this proposed scheme help detecting cancer.
- Patient, this scheme may make the process of cancer diagnosis become faster.
- Researcher, proposed scheme will give an insight how PCA and kernel tricks can affect microarray data classification.

1.7 Thesis Organization

The thesis book consists of some chapters and each chapter is related to another chapter. The organization of thesis is arranged as follows.

a) Chapters 1 is Introduction. It tells about overview of cancer and microarray data, characteristic of microarray data and its classification which become background of the problem, and the problem.

b) Chapter 2 is Literature Review. It tells about existing works related to microarray data classifications.

c) Chapter 3 is Methodology and System Design. It explains the methodology used in this study and the system design of the scheme.
d) Chapter 4 is Implementation and Analysis. It shows the experimental result and provide the analysis of the experimental result. The analysis includes influence of dimension reduction technique and classification method in term of performance.
e) Chapter 5 is Conclusion and Future Works. It tells the conclusion from this study and future work for next development related to this study.