The number of cancer patients in the world is increasing rapidly. Cancer is one of the leading causes of deaths in the world. It is a prominent disease of genes gone bad. In cancer, genes that control the orderly replication of cells get damaged, allowing the cell to reproduce without restraint. This eventually spreads into neighbouring tissues and sets up growth throughout the body. All types of cancers are genetic i.e. they are triggered by altered genes. Most cancers come from random mutations that develop in body cells during one's lifetime, either when cells are going through cell division or in response to injuries from environmental agents such as radiation or chemicals (http://cancer.health-cares.net).

Over the past few decades, classification and diagnosis of cancer patients were based on the examination of the organ where cancer developed. Diagnosis of cancer involves biopsy of a tumour or a suspected tissue in the body. Cancerous cells can vary from being quite similar to normal cells in appearance to being totally different without any similarity. Physicians use this histopathological examination to determine the prognosis, or the prospect for the patient's recovery from the disease. However, appearances can be deceiving thereby affecting the accuracy of results subject to experience of the pathologists. The diagnostic process is done through surgical operations which might expose the patients to different kinds of risks and is expensive.

Patrick Brown and associates of the Stanford University College of Medicine developed a technology known as DNA microarrays or DNA chips in the ninety's. With the advent of this technology, it is possible now to measure the gene expression profile of thousands of genes simultaneously in a single experiment. This technology has revolutionized classification and diagnosis of cancers. This also provides a sounder basis for deciding on a course of treatment. This technology classifies tumours by molecular characteristics, rather than by cell features. The rationale behind this technology is that the overall behaviour of a cancer can be determined by expression of genes within it (Cooper, 2001). It should, therefore, be possible to identify sets of genes whose expression or lack of expression defines each property of tumour, including its
precise diagnosis and clinical behaviour (Cooper, 2001). In recent years, research has shown that accurate cancer diagnosis is achievable by analysing the gene expression of a person suffering from cancer. In a landmark research study, Golub et al. (1999) applied microarray technology to classify leukemia, using microarray analysis based on neighbourhood analysis and the use of tumour class predictor. This procedure distinguished between acute myeloid leukemia and acute lymphocytic leukemia. Alizadeh et al. (2000) extended this approach for analysis of diffuse large B-cell lymphoma (DLBCL), the most common subtype of Hodgkin’s lymphoma. Analysis of microarray expression profiles led to the definition of two molecularly distinct categories of DLBCL (Cooper, 2001). Ross et al. (2000) used cDNA microarrays to determine the expression of 8000 genes within the 60 cell lines (NCI60 cell lines) used at National Cancer Institute for anticancer drugs.

Several classification and clustering techniques from data mining and machine learning have been applied to distinguish cancerous cells from non-cancerous cells and for the classification of different types of cancers using gene expression profile of different patients (Kohavi and John, 1997; Khan et al., 2001; Guyon et al., 2003; Liu et al., 2004; Shah and Kusiak, 2007). However, these techniques are unable to provide accurate results on microarray datasets. Classification of microarray datasets faces many challenges. One of the major challenges is that these datasets are characterized by large number of genes and small number of samples. This small number of samples compared to the large number of genes wakes up the curse of dimensionality (Bellman, 1961). In addition, a large number of genes are housekeeping genes and thus do not contribute significantly to classification. Also, many genes are correlated to each other. These redundant and irrelevant genes have negative effect on classification accuracy of the classifier. They also increase data acquisition cost and learning time. To improve classification accuracy, there is a need to reduce dimension of such datasets and determine set of genes relevant for classification.

Dimension reduction is possible in two ways: Feature Selection and Feature Extraction (Guyon and Elisseeff, 2003). Feature Selection refers to reducing the dimensionality of the measurement space by discarding redundant, noisy and irrelevant features (Duda et al., 2000). Feature Extraction refers to utilizing all the information
contained in the measurement space to obtain a new transformed space and then important features are selected from the new transformed space (Jain et al., 2000). The choice between feature selection and feature extraction depend upon the application domain and specific training data available. Feature selection leads to saving in measurement cost and the selected features retain their original physical interpretation. In addition, the retained features may be important for understanding the physical process that generates the patterns. On the other hand, transformed features generated by feature extraction may provide a better discriminative ability than the best subset of given features but these new features may not provide any physical meaning (Jain et al., 2000).

In microarray datasets, one is not only interested in classifying the given sample based on gene expression but also in identifying important genes which helps in the classification of cancerous and non-cancerous patients. Finding such genes can help druggists to design appropriate medicines targeting only those genes. Hence, dimension reduction is carried out with feature selection rather than feature extraction. Feature selection not only allows building faster and efficient learning algorithm by removing irrelevant, redundant and noisy features but also provides better understanding of genes which cause a particular disease.

Extracting a subset of informative genes and removing irrelevant or redundant genes for accurate classification has become an important area of research in last few years. The process of selecting a set of important genes without losing any information is known as gene selection (Guyon and Elisseeff, 2003). There are two major approaches to feature/gene selection: filter and wrapper approach (Kohavi and John, 1997; Guyon and Elisseeff, 2003; Ruiz et al., 2006). Most filter methods use statistical characteristics of data for feature selection which requires less computation. It independently measures the importance of features without involving any classifier (Guyon and Elisseeff, 2003). Since the filter approach does not take into account the learning bias introduced by the final learning algorithm, it may not be able to select the most relevant set of features for the learning algorithm (Liu and Yu, 2005). Wrapper method evaluates feature subsets using a learning algorithm. They find features better suited to the predetermined learning algorithm resulting in better performance.
However, they are computationally more expensive as the classifier must be trained for each candidate subset (Kohavi and John, 1997; Liu and Yu, 2005).

Recently some hybrid approaches have been suggested in literature (Xing et al., 2001; Ruiz et al., 2006; Tang et al., 2007; Chen and Zhao, 2008). These approaches combine more than one method to obtain a subset of genes which can distinguish a cancerous sample from a normal sample. It has been observed that these approaches give better results than using a single approach.

The objective of this thesis is to build a classifier for microarray datasets that can effectively predict whether the person is suffering from cancer or not by looking at the gene expression profile of the patient. As these datasets suffer from curse of dimensionality, gene selection is required to reduce the dimension of these datasets. So we aim to determine a set of discriminatory genes that can classify data efficiently. This can be achieved by removing noisy, redundant and irrelevant genes from the original set of genes. In this thesis, we propose four different hybrid approaches for selecting a set of relevant and non-redundant genes to classify microarray datasets. We have used four different classifiers and six microarray datasets to evaluate the proposed approaches for gene selection.

The thesis is organised as follows: Chapter 2 discusses microarrays and the datasets used in experiments. Classification methods are discussed in chapter 3. Chapter 4 reviews the basics of dimension reduction and different techniques used for gene selection. Chapters 5-8 describe proposed gene selection techniques.

In chapter 5, a novel two-stage ensemble approach to determine a subset of relevant genes for reliable cancer classification is proposed. Instead of relying on one gene ranking method, the proposed method considers union of informative genes selected by different gene ranking methods. This reduces chance of missing informative genes. This set of informative genes may contain redundant genes as ranking methods do not take into account the relationship between different genes. In second stage, sequential forward search is used with a measure that selects relevant and non-redundant genes.
In chapter 6, we discuss a two stage algorithm for finding a small subset of relevant genes responsible for classification of high dimensional microarray datasets. This algorithm takes advantage of both filter and wrapper method. It is based on the principle of Mutual Information and Cross Entropy (MICE). In first stage of algorithm, mutual information is employed to select a set of relevant genes and cross entropy is used to determine independent genes. In second stage, a wrapper based sequential forward feature selection method is used to obtain a set of optimal genes for a given classifier.

In chapter 7, a novel two stage algorithm based on clustering is proposed. In first stage of algorithm, the entire feature space is divided into k clusters using different clustering techniques. Three different clustering techniques have been investigated. Similarity measure used for clustering is maximal information compression index. Once genes are clustered, an informative gene is selected from each cluster using t-statistics and a pool of non-redundant genes is created. Second stage uses a wrapper based sequential forward feature selection method to obtain a smaller set of relevant and non-redundant genes for a given classifier.

In chapter 8, an algorithm based on genetic algorithm is proposed. Genetic algorithm cannot be directly applied to high dimensional datasets due to high computation cost. Hence, a two stage algorithm is proposed. In first stage, size of the original gene set is reduced. For reducing original gene set, two different schemes are investigated. Second stage uses genetic algorithm in conjunction with SVM to select a smaller set of discriminatory genes for classification. A new fitness function for genetic algorithm is also proposed.

In Chapter 9, proposed gene selection techniques have been compared in terms of classification accuracy and number of genes selected. Proposed algorithms have also been compared with the state-of-art methods. The comparison comprehensively concludes that proposed algorithms achieve better classification accuracy even with lesser number of genes.