Gene Filtering and Classification of Microarray data

By

Aneeqa Ali

NUST201464102MSEECS61314F

Supervisor:

Dr. Mian Muhammad Hamayun

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST), Islamabad,
Pakistan.
(July, 2018)

Approval

It is certified that the contents and form of the thesis entitled "**Gene Filtering and Classification of Microarray data**" submitted by **Aneeqa Ali** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Mian Muhammad Hamayun**

Signature: ____________

Date: ____________

Committee Member 1: **Dr. Anis ur Rehman**

Signature: ____________

Date: ____________

Committee Member 2: **Dr. Asad Waqar Malik**

Signature: ____________

Date: ____________

Committee Member 3: **Dr. Asad Ali Shah**

Signature: ____________

Date: ____________

Dedication

I dedicate this thesis to my parents, siblings and friends who have always motivated me in my life.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at National University of Sciences & Technology (NUST) School of Electrical Engineering & Computer Science (SEECS) or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Aneeqa Ali**

Signature: ________________

Acknowledgments

First of all, I am grateful to Allah Almighty for giving me the knowledge and understanding to complete this research work. Secondly, I would like to thank my Supervisor Dr. Mian Muhammad Hamayun for his consistent support and guidance throughout my thesis.

A special thanks to all the GEC members for their suggestions and advice, which were very helpful in every step of this research.

I would like to thank my parents and siblings for their full support and endless love all along. Finally, special thanks to my friends for their constant encouragement.

Aneeqa Ali

Abstract

Microarrays have been widely used by scientific community to study and analyze the expression of large number of genes simultaneously. The advance of microarray technology provides a huge amount of genomic data which leads to the necessity of efficient methods for its analysis. This technology gains special attention in the field of cancer research because with better classification of tumors, it would be easier to efficiently diagnose and treat cancerous cells. Efficient classification of microarray data is still a problem because of increased dimensions of the feature space and a very small sample size. Effective methods are required to reduce the dimensionality and improve the classification accuracy to by extracting meaningful information from the datasets. In this study, we aim to find a combination of different feature selection and classification methods that work best in terms of accuracy and number of features selected. Our proposed approach uses Correlation Based Feature Selection CFS (using Forward Search) as feature selection method combined with an ensemble based on SVM , Random Forest, Bagging and Bayesian Generalized Linear Models (BayesGLM). A number of experiments are conducted on the benchmark datasets: colon cancer, prostate cancer, leukemia and breast cancer. We demonstrated that our proposed approach outperforms or give comparable results with already published approaches in literature.

Contents

Page

List of Abbreviations

AUC	Area Under Curve
BayesGLM	Bayesian Generalized Linear Model
cDNA	Complementary DeoxyriboNucleic acid
CFS	Correlation based Feature Selection
CV	Cross Validation
DNA	DeoxyriboNucleic acid
FCBF	Fast Correlation Based Feature selection
IG	Information Gain
KNN	K Nearest Neighbour
KSVM	Kernel Support Vector Machine
mRNA	Messenger RiboNucleic acid
ROC	Receiver Operating Characteristic
SU	Symmetrical Uncertainty
SVM	Support Vector Machine

List of Figures

List of Tables

Chapter 1

Introduction

Microarray technology is able to provide a treasure of information on the expression level of thousands of genes. These genes are used for prognostic and diagnostic purposes of different diseases. Formerly researchers were able to study only a few genes at one time but an opportunity for studying the expression level of the whole genome is provided using this progressive technology. The results obtained from microarray experiments helps us to understand the genes that are regulated at the molecular level under the disease condition in clinical medicine as well as in biology [2].

A DNA microarray consists of thousands of DNA spots arranged on a chip. Each spot contains a gene. When a gene is expressed in a cell, it generates messenger RNA (mRNA). Detection of mRNA can be performed on the microarray. Firstly in microarray experiment, healthy and diseased tissue samples are collected . Then, messenger RNA (mRNA) is isolated from the samples. A copy of DNA also known as complementary DNA (cDNA) is made by tagging the mRNA with fluorescent materials. This cDNA when applied to microarray binds itself to the base pairs in each of the spots on the array. This process is called hybridization [3]. Fig 1.1 shows a typical DNA microarray experiment.

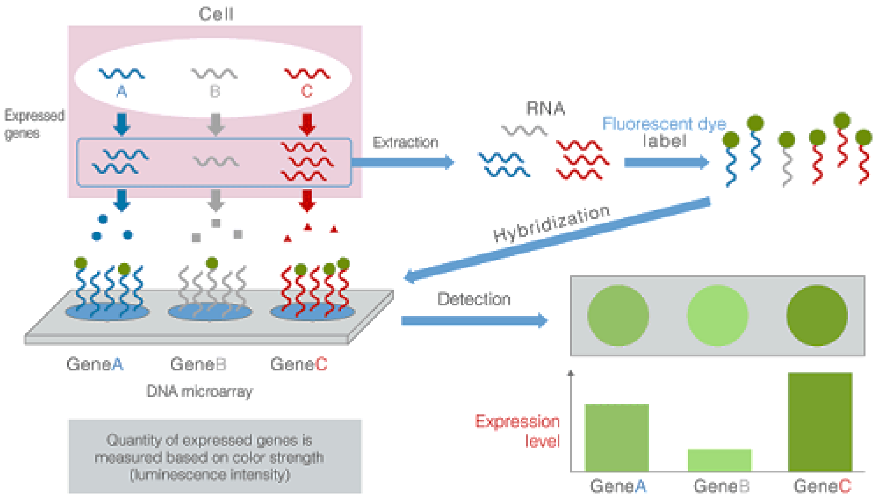


Figure 1.1: A Typical DNA Microarray Experiment [1]

Multiple data processing steps are applied to the microarray slide which includes collection of data from the image, quality control and normalization. A 2D array D containing thousand of thousands of columns (genes) and several rows (samples) is created as resultant dataset.

$$\mathbf{D} = \begin{bmatrix} x_1^1 & \cdots & x_1^n & c_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_m^1 & \cdots & x_m^n & c_m \end{bmatrix}.$$

Figure 1.2: Microarray Data Format

To date, this technology has gained a lot of attention in the cancer studies. Cancer is among the few fatal genetic diseases that is either caused by epigenetic changes or by the acquired mutations that ultimately leads to the altered gene expression of the cancer cells. Hence, microarray technology is used for the clinical diagnosis by discovery of biomarkers and also by identification of genes that are either up regulated or down regulated and plays a role in specific cancer [4]. However, in terms of clinical application, such approach is costly, tedious and is not practical for every patient. Nowadays, microarray technology cannot benefit the researchers because of multiple limitations of algorithms that are used to analyze data. Cancer progression is empowered by creating a set of marker genes with data classification. The number of genes is very crucial in microarray data analysis because usage of only a few genes cannot produce reliable outcomes and using a large number of genes introduces additional noise with decreases information [5]. So, for each cancer type, there is a need to locate an ideal set of genes that helps to classify various samples with high precision.

There are two main steps involved in microarray gene expression classification task: Feature Selection and Classification. Feature selection reduces the dimensionality and makes classification easier by only selecting relevant genes. Classification is a critical step for the actual prediction of classes. In order to effectively classify the microarray data, different machine learning algorithms and feature selectors has been discussed in chapter 2.

This research focuses on ensemble based classification system built with four base classifiers, Kernel Support Vector Machines (KSVM), Random Forest [6], Bagging [7] and Bayesian Generalized Linear Model (bayesGLM) [8]. Correlation Based Feature Selection (CFS) is then used to select relevant genes for classification. The model is validated using 10-fold cross-validation and results are compared with several classification techniques.

1.1 Research Questions

This thesis will address the following questions related to gene expression classification for prediction of cancer based on microarray datasets.

- How to identify an ideal set of genes that play a significant role in predicting a certain kind of cancer?

- How to classify normal and cancer samples with high accuracy?

- How to build a classification model that is independent and works best for a wide range of datasets?

1.2 Motivation

Large number of genes relative to number of samples is an important feature of microarray data. With an increase in high dimensionality of gene-space, it's computation complexity is increased which further results in decreased accuracy of classification. Over fitting is also one of the major issues due to increased dimensionality. This is worsened by the small sample size of gene expression data. Thus, due to these issues there is a need of introducing such novel classification methods which are able to perform efficiently and increase prediction accuracy while dealing with the increased dimensionality and small sample size of the gene expression data.

Microarrays has received the most attention in the area of cancer research. According to World Health Organization (WHO), 24 million new cancer cases by 2035 and an annual death rate of 14.5 million due to cancer could be experienced by the world [9]. Thus for the successful diagnosis and cure of the cancer, a reliable and accurate classification system is required. Another challenge is the presence of irrelevant attributes or genes in cancer datasets. Almost all types of dataset contains irrelevant attributes but in case of cancer dataset, the number of relevant genes are very less when compared to the total number of genes in the dataset. These irrelevant genes interfere with the ability relevant gene attributes to discriminate between cancerous and non-cancerous cells. Therefore, a good feature selection method is required which can filter out the subset of genes that plays role in cancer.

The main idea in this research work is to find a combination of feature selection and classification method to improve the performance of cancer classification for microarray dataset. Our approach built on ensemble is highly effective in terms of accuracy and usability. We used a CFS with forward search based approach for selecting the optimal subset of genes. Then, an ensemble based classifier is used to build a model on the basis of selected genes. We used four benchmark cancer datasets i.e. ‘Leukemia’, ‘Colon’, ‘Prostate’ and ‘Breast Cancer’ dataset for the testing and comparison performance of our proposed approach with existing methodologies in literature.

1.3 Thesis Organization

Chapter 2 discusses all the feature selection methods and machine learning techniques present in literature for classification of microarray gene expression data. Chapter 3 provides detail of proposed methodology and all the steps involved along with the description of benchmark datasets. Chapter 4 discusses the experimental setup and the results obtained after implementing our proposed approach along with the comparison with multiple classification techniques. Chapter 5 summarizes the overall performance of this research followed by chapter **??** which gives an insight of the future work that can be done in this particular area.

Chapter 2

Literature Review

This chapter briefly describes the literature that includes microarrays and application of microarrays and then provides a detailed overview of microarray based cancer classification followed by a summary of all the recent published work related to cancer classification.

2.1 Microarrays

Researchers have been facing a great number of problems in the field of genomics due to the increased the volume of genomics data. The techniques of data analysis, modeling and result interpretation are not enough to cater the needs of such a large dataset. Microarrays, on the other hand, have played a significant in this area. This is because multiple DNA sequences have been arranged in orderly fashion to perform data analysis easily [10]. They are considered as a ground breaking technology, as it facilitates the study of thousands of genes simultaneously from complex nucleic acid samples. Moreover, are also used in the identification of changes that are associated with diseases and drug discovery at a genetic level. [11].

It is very important to analyze this data it can lead to the discovery of unknown knowledge which can be validated though multiple experiments. Multiple topics are included under the analysis of microarray gene expression data. They can be used to identify:

- Genes that can be Co-expressed [12]
- Group of genes that exhibit similar expression patterns [13]

- Group of genes having high discrimination rate to differentiate multiple biological samples.
- Behavior of genes under different stress conditions.

This data can also be used to identify and classify cancerous cells [14], which will help in finding a proper treatment for cancer and develop better drugs [15]. Previously, multiple non-molecular characteristics i.e. Types of tumor tissues, pathological characteristics and clinical phase have been used to diagnose complex genetic diseases but now gene expression data is being used to identify cancerous cell using multiple data mining and machine learning algorithms.

2.2 Microarray Cancer Classification

Microarray data presents a great challenge to computational techniques because the nature of this data has high-dimension with small sample sizes. This means that the process of classifying the microarray data is divided into two main steps. First step is implementing a gene selection technique (to reduce dimensionality), and the second step is classification. A lot of studies have been made in past years for the analysis of microarray data to identify diseased and normal samples. The study of Golub et al [16] for classifying different types of leukemia cancer is considered to be one of the pioneer work in this area. After that many subsequent studies both supervised and unsupervised on microarray expression data is performed. In this thesis, we mainly focus on the supervised learning where class labels are known beforehand.

Microarray data contains large number of genes when compared to the number of samples. This feature of microarrays make is very useful. The accuracy in classification of cancerous cells is decreased because of increased computational complexity caused by high dimensionality in gene space. Therefore, due to these issues it is necessary to reduce the high dimensionality of gene space or gene selection. This can help to improve the accuracy of classification, reduce classification complexity and make it easier to identify the relevant genes only. Several methods such as *Partial Least squares* [17], *Principle Component Analysis* [18], *Correlation Based Feature* Selection [19], *Neighborhood Analysis* [20] have been developed in past in order to choose an optimal subset of predictor genes.

Secondly, in experimentations on microarrays, the correct classification of data is also a critical step to predict classes accurately. Multiple machine learning algorithms have been used to evaluate microarray data which include Support Vector Machines (SVM) [21], Artificial Neural Networks (ANN) [22], K Nearest Neighbors (KNN) [23], Random Forest (RF) [24], and evolutionary techniques such as Genetic Algorithms [25] , Genetic Bee Colony (GBC) algorithm [26]. Some comparative studies have also been conducted, like [27] [28] [29]. Rest of the chapter discusses the recent researches for microarray cancer classification in detail.

Rama and Swati [30] presented a hybrid approach to classify microarray data. It consists of three phases feature extraction followed by feature selection and then classification. To extract features, Principal Component Analysis (PCA) was applied. Genetic algorithm (GA) was then used for the selection of an optimal subset from the extracted features. The selected features by GA were then given to Probabilistic Neural Networks (PNN) as input. GA was also used in that phase to optimize the topology of PNN by selecting minimum number of samples. Results were tested on three different cancer datasets Colon, DLBCL and Leukemia and accuracies of 95.83%, 96.67% and 95.83% respectively were achieved.

In 2015, Devi Arockia [31] tried to classify microarray cancer data using SVM and Mutual Information (MI) based gene selection approach. MI is a filter based approach which utilizes the probability distribution of genes to find correlations. For classification of samples, SVM with four kernels was used which are linear, radial, polynomial and quadratic. SVM with linear kernel worked best among other classifiers and gave accuracy of 67.7% for Colon and 97.7% for lymphoma dataset.

Hala et al. proposed a hybrid gene selection algorithm in [26]. This algorithm is known as the Genetic Bee Colony (GBC) algorithm. It combines the advantages and features of Genetic Algorithm (GA) with Artificial Bee Colony (ABC) algorithm. For the preprocessing step, Max Relevance Min Redundancy (mRMR) was used to filter redundant or noisy genes. Results were tested on three microarray datasets which included lung cancer, colon and leukemia. For the comparison of results, following techniques were used ,mRMR with Ant Bee Colony (ABC) algorithm, mRMR combined with Genetic Algorithms and mRMR with Particle Swarm Optimization (PSO). The GBC algorithm showed excellent performance and proved a promising approach by achieving the maximum classification accuracy with minimum average number of selected genes.

Sara Haddou [32] tried multiple feature selections techniques with KNN and SVM classifiers to measure the performance for leukemia, colon and prostate cancer datasets. The feature selection techniques used by them includes ReleifF, t-statistics, Fisher and Signal to Noise Ratio (SNR). Result shows that SNR outperformed for all three cancers with SVM as classifier.

A study by Alagukumar [33] used Associative Classification algorithm to classify gene expression data. They formulated a classification algorithm based on association rule mining. The proposed approach comprised of four phases: Gene filtering, Discretization, Class Association Rules and Results Prediction. Genes that were different from each other were found in Gene Filtering phase using t-test. Genes having p-value of 0.5 or less were selected which removed all non-significant genes from the data. All the continous values were discretized into discrete values in Phase 2. It also substituted the gene expression values with the interval containing it. Closed frequent itemset was used to generate association rules that were used in the classification model. The last phase used a scoring function to predict class. Breast cancer gene expression data was used to test and compare results of this technique with other algorithms such as Decision Trees, LDA and SVM. Leave one out cross validation method was used to evaluate the the performance of the techniques which achieved

90% accuracy. The results of this techniques had a better accuracy as compared to other algorithms.

Another pattern classification approach presented by Ricardo uses Principal Component Analysis for dimensionality reduction [34]. Small factor set were created from the actual number of variables using PCA. Each factor in the factor set was a linear representation of the actual variables. This means that PCA actually created a different and compact dataset from the original data of same data while keeping the same number of features. After this, logistic regression was used to select a subset of features. Then, PCA was again applied on the new dataset obtained from logistic regression. This generated a new set of principal components which was provided to classifier. To test the approach, a total of eight different classifiers were used which are: Support Vector Machine (SVM) with Linear and Radial basis functions, Extreme Learning Machine (ELM), Lattice Neural Network with Dendritic Processing (LNNDP), Bayesian Net (BN), Naive Bayes, Multi-Layer Perceptron (MLP) and Radial Basis Function Neural Networks (RBF). For Leukemia, best accuracy was achieved by MLP (94.42%) whereas for lymphoma, 91.60% was also the highest accuracy achieved by MLP.

An ensemble based approach was proposed by Sara Tarek et al [35] which used 5 individually trained base classifiers. Each classifier implemented 3NN (Nearest Neighbor) technique and had a separate feature selection technique to select the training feature set. BAHSIC feature selection algorithm was used in three classifiers with varying number of genes to select i.e. 50, 5 and 25 respectively. EVD gene selection algorithm was used in the fourth classifier. It used an automated algorithm for selecting genes. For Breast Cancer, Colon and Leukemia datasets, 5127, 49 and 224 genes were selected respectively. SVD entropy was used by the fifth classifier while selecting 1236, 240, 187 genes for Breast cancer, Colon and Leukemia datasets. This selection of genes was done automatically by the algorithm. For error estimation, "Bolstered Re substitution Error" (BRE) was used. Predictions from all classifiers were combined using Majority voting scheme. Experimentation on Breast Cancer, Colon and Leukemia datasets showed that proposed technique gave good results. There is no measure of accuracy of results. ROC was used to evaluate the performance of system.

Hanaa Salem [36] performed gene expression classification for cancer by combining Information Gain (IG) and Standard Genetic Algorithms (SGA). Firstly, a subset of significant features from microarray dataset was selected by IG. Then in the next stage GA was used to reduce the features selected by IG. For the classification task, Genetic Programming (GP) was used to predict the classes. GP is a branch of GA. The difference is that, in GP, individuals has tree structure whereas in case of GA, the individuals are string-structured. Experiments were performed using 7 cancer microarray datasets and compare with other techniques in the literature. Results showed that the algorithm performed differently for each dataset but the overall performance of classification was improved. For colon , leukemia and prostate cancer, the accuracies achieved were 85.48%, 97.06% and 100% respectively.

A hybrid method proposed by Lingyun et al [37] to find informative genes by combining IG with SVM. This approach used high efficiency of filter methods combined with excellent performance of wrapper methods to achieve better classification accuracy. IG first selects a subset of genes followed by SVM to further eliminate redundant genes. SVM was also used to perform the classification task. The results were compared with other methods relief etc but IG +SVM outperformed and gave accuracies of 90.32% for Colon Cancer, 96.08% for Prostate cancer and 98.61% for Leukemia.

Zhong et al [38] presented an approach for distance based feature selection for binary classification. Bhattacharya distance was used to rank genes on the basis of distance between two classes for each gene. Finally, each subset of gene was evaluated with SVM, which led to the identification of subset of discriminative genes that gave minimal classification error rate. Results were compared with Supervised Weighted

Kernel Clustering SVM (SWKC/SVM) and SVM with Recursive Feature Elimination (SVM-RFE). The proposed approach performed better than the compared techniques and gave accuracy of 90.5% for colon cancer and 96.9% for leukemia dataset.

2.3 Summary of Related Work

The section provides a summarized view of all the related work done for microarray cancer classification. Table 2.1 includes the feature selection machine learning techniques and datasets used in each work along with achieved accuracies.

Table 2.1: Summary of Related Work

	Feature Selection	Classifier	Dataset Used	Accuracy
Rama and Swati (2014) [30]	Principal Component Analysis	Probabilistic Neural Network	Leukemia	95.83%
			Colon	95.83%
			DLBCL	96.67%
Devi Arockia (2015) [31]	Mutual Information (MI)	SVM	Lymphoma	97.7%
			Colon	67.7%
Sara Haddou Bouazza (2015) [32]	ReliefF	SVM	Leukemia	96%
			Colon	82%
	Signal to Noise Ratio (SNR)	SVM	Leukemia	97%
			Colon	85%

Alagukumar (2016) [33]	t-statistic	Associative Classification	Breast	90%
Ricardo Ocampo Vega (2016) [34]	Principal Component Analysis	Multi Layer Perceptron	Leukemia	94.42%
			Lymphoma	91.60%
Sara Tarek (2017) [35]	Backward Elimination Hilbert-Schmidt Independence Criterion (BAHSIC) -Extreme Value Distribution based gene selection (EVD) -Singular Value Decomposition Entropy gene selection (SVDEntropy)	Ensemble with 5 base classifiers each with 3-NN	Leukemia	(Area Under Curve) AUC=0.99
			Colon	AUC=1.00
			Breast	AUC=1.00
Hanaa Salem (2017) [36]	IG + GA	Genetic Programming	Leukemia	97.06%
			Colon	85.48%
			Prostate	100%
Lingyun (2017) [37]	IG + SVM	SVM	Leukemia	98.61%
			Colon	90.32%
			Prostate	96.08%
Wenyan Zhong (2017) [38]	Bhattacharya Distance	SVM	Leukemia	96.9%
			Colon	90.5%

To summarize this, different classification and feature selection algorithms have been studied and experimented in the past to classify gene expression data for predicting cancer disease. However, efficient methods are still needed to improve the accuracy of cancer prediction. Moreover, different algorithms works best for different datasets, a more generalized classification approach is needed that performs better independent of dataset.

Chapter 3

Proposed Methodology

The goal of this study to effectively classify an unknown gene sample using Ensemble and compare it with other classification techniques. To reduce the dimensionality of dataset, three feature selection methods are used namely Fast Correlation based Filter (FCBF) , CFS(Correlation with Forward Search) , IG(Information Gain). Four benchmark cancer datasets are used to evaluate the performance of our proposed method. Each dataset is first preprocessed by log transformation or filtering out noisy data. Then, it is split into train and test set with 70%:30% ratio. After that, feature selection is applied to train set to find subset of informative genes which helps in classification of diseased and normal samples. Test samples with selected genes are then classified using Ensemble, SVM , SVM radial, KNN , Random Forest and Logistic Regression. 10-fold cross validation technique is used to validate classifier performance. Fig 3.1 provides an overview of the proposed methodology followed in this research. This chapter includes four main sections:

- Datasets
- Preprocessing
- Feature Selection
- Classification Techniques

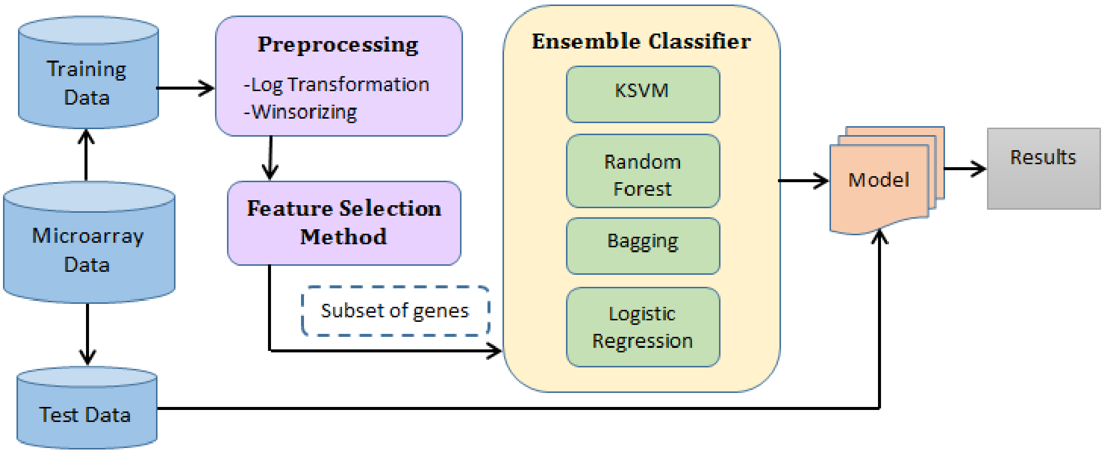


Figure 3.1: Proposed Methodology

3.1 Datasets

Four benchmark microarray cancer datasets are used in this study. 1. Colon Cancer 2. Prostate Cancer 3. Breast Cancer 4. Leukemia Cancer

3.1.1 Colon Cancer

This dataset presented by Alon et al. consists of 62 samples collected from patients of colon cancer. Out the total , there are 20 normal samples and 40 tumor samples. Each sample has expression patterns of 2000 genes. [39]

Table 3.1: Format of Colon Cancer Dataset

	Gene ID	**Sample 1**	**Sample 2**	**....**	**Sample 62**
1	H55933	8589.4163	9164.2537	....	6246.4487
2	R39465	5468.2409	6719.5295	....	7823.5341
3	R85482	4064.9357	3718.1589	....	3975.5643
4	U14973	1997.8929	1997.8929	....	2002.6131
....	....	....	....	....	....
2000	H80240	2773.4212	2793.3875	....	1714.6312

3.1.2 Prostate Cancer

Gene Expression dataset containing 6033 genes for 102 samples was presented by Singh et al [40]. All the samples are labeled as "healthy" or "cancer". These 102 samples include 50 healthy samples and 52 patients of prostate cancer. Table 3.2 shows the format of prostate cancer dataset.

Table 3.2: Format of Prostate Cancer Dataset

	Gene Accession No	**Sample 1**	**Sample 2**	**....**	**Sample 102**
1	AFFXMurIL2_at	-9	-2	....	-1
2	AFFXMurIL10_at	1	1	....	0
3	AFFXMurIL4_at	15	4	....	5
4	AFFXBioB5_at	-3	-5	....	-4
....	....	....	....	....	....
6033	AFFXCreX5_at	0	0	....	-9

3.1.3 Breast Cancer

The Breast cancer dataset used in this study is taken from the research of van't Veer et al. [41]. This dataset consists of 4948 genes expressed over 78 samples taken from patients with lymph node negative. Out of 78, 34 samples were collected from patients who had grown distant metastases within 5 years, and 44 samples from patients who were free from the disease after at least a period of 5 years.

Table 3.3: Format of Breast Cancer Dataset

	Gene ID	**Sample 1**	**Sample 2**	**....**	**Sample 78**
1	Contig45645_RC	-0.125	-0.27	....	-0.382
2	Contig44916_RC	0.07	0.123	....	0.064
3	D25272	-0.006	0.056	....	-0.033
4	J00129	-0.575	-0.499	....	-0.873
....	....	....	....	....	....
4948	Contig29982_RC	-0.575	-0.402	....	-0.474

3.1.4 Leukemia Cancer

Leukemia is a cancer of blood and bone marrow. The leukemia dataset of Golub et al. [16] consists of 7129 features (genes), and is categorized into two classes: Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). The total dataset has 72 bone marrow samples, 25 AML samples and 47 ALL samples. The level of expression of the gene in the microarray experiment is presented in the first column. The second column CALL defines whether the expression value is due to the gene or due to noise. It might be any of the three values: presence, marginal or absence, represented by P, M and A according to the signal.

Table 3.4: Format of Leukemia Cancer Dataset

	Accession No	**Sample 1**	**CALL**	**Sample 2**	**CALL**	**....**	**Sample 72**
1	AA28102_at	181	A	484	A	....	118
2	AB000114_at	72	A	61	A	....	16
3	AB000115_att	281	A	118	A	....	197
4	AB000381_s_at	29	P	38	A	....	50
....	....	....	....	....	....	....	....
7129	AB000449_at	57	P	274	P	....	311

3.2 Preprocessing

Microarray gene expression data are noisy and of high dimensionality. Therefore, an essential preliminary step when performing microarray data analysis is to preprocess the data i.e. to carry out transformations of data to eradicate the measurements that are non-significant and with low quality, so that the analysis is performed with 'clean' data[42]. As part of pre-processing, following steps are performed for each dataset:

For colon cancer dataset, logarithmic transformation to base 10 is applied followed by filtration of genes as part of preprocessing .

Leukemia data is preprocessed via logarithmic transformation to base 10 and feature selection as suggested by Dettling [43]. Genes with CALL values labeled as (A) or Absent are eliminated.

To remove the outliers generated due to background noise from gene expression data, data is winsorized between minimum value around 100 and maximum value around 16000. This is beacause imaging equipment in microarray experiment cannot measure values higher than 16,000 and values which are lower than 20 are due to background noise. [44][45].

3.3 Feature Selection

Feature selection is a process of selecting subset of relevant features which helps in classification. It is applied to microarray data to select the most important or informative genes that play significant role in the prediction of disease. It not only reduces the dimensionality of microarray dataset but also filter out irrelevant genes that affect the classification process. This will lead to improved classification accuracy and also reduces the risk of over fitting. There are two main kinds of feature selection methods, filters and wrappers. Filter approaches select feature subsets by applying various statistical tests on the data. Wrapper approach however, selects feature subset by training a model and uses cross-validation to calculate the score of feature subsets. In this thesis, we used three filter methods FCBF, CFS and Information Gain. The filter methods are selected over methods because they are more robust and less prone to over fitting as compared to wrapper methods.

3.3.1 FCBF

FCBF (Fast Correlation Based Filter) [46] is a quick correlation based feature selection method. It uses backward search strategy to find the optimal subset of features and symmetrical uncertainty (SU) to find the correlation or redundancy among features. The algorithm starts with all the features, and in each step a feature is removed until there are no features left to eliminate. Symmetrical Uncertainty (SU) is a normalized form of information gain which uses entropy and conditional entropy to measure redundancy of features. If X and Y are two random variables, H(X) is the entropy of X and H(X|Y) is the conditional entropy of X given variable Y, then SU(X,Y) is:

$$SU(X,Y) = 2\left[\frac{H(X) - H(X|Y)}{H(X) + H(Y)}\right] \tag{3.1}$$

The SU = 1 indicates that one feature can be completely predicted by other's feature value and 0 value shows that two features are entirely independent. For calculating SU values, the features must be nominal and if continuous, their values must be discretized.

3.3.2 CFS

Correlation based feature selection(CFS) is a filter based technique for evaluating the subsets of features on the basis of following hypothesis "A good feature subset has features which are highly correlated to the output class and uncorrelated to each other" [47]. Irrelevant features are ignored because they will have low correlation with the class. However, redundant features are removed as they will be highly correlated with other features.

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{3.2}$$

where M_S is the merit of a feature subset S containing k features, $\overline{r_{cf}}$ represents the mean of feature-class correlation, and $\overline{r_{ff}}$ is the mean of correlation between feature to feature. The numerator of equation 3.2 represents the extent to which a set of features can predict a class whereas the denominator shows the level of redundancy among the features. The implementation of CFS used in this thesis incorporates forward search as heuristic search strategy. In forward search, the algorithm starts with an empty set of features, and in each step a feature is added till addition of further features doesn't yield higher results . The resultant feature set is constructed in a greedy manner however, it also considers some of the interdependencies between features [48].

3.3.3 Information Gain

Information gain (IG) describes how much âĂIJinformationâĂİ a feature gives us about the class. So, the features that perfectly partition should have maximum value of information gain whereas features that are not related should give no information. Mathematically, it can be calculated as the decrease in *entropy* or *uncertainty* after a dataset is split on an attribute:

$$Gain\ (T,\ X) = Entropy(T) - Entropy(T, X) \tag{3.3}$$

Entropy(T) represents the uncertainty involved in predicting the value of a random variable, whereas *Entropy(T, X)* denotes the uncertainty based on the known variable X.

3.4 Classification Techniques

Classification is a data mining technique used to classify each sample into predefined classes. A model is first trained using a training dataset. Then, this trained model will predict classes for the test samples. In microarray cancer classification, classifier is used to distinguish between different cancer types or "healthy" or "cancerous" samples. Our proposed approach uses an ensemble method (combination of classifiers) to predict normal or tumor samples from cancer microarray data. Ensemble helps to build an optimal classification method by utilizing more than one classifier models to enhance accuracy of classification. One of the main reason it performs better than single classifier is because it uses multiple classification models. Ensembles are usually categorized into two major types: Sequential Ensemble Methods, in which base learners are usually created sequentially and there is a dependency among base learners, whereas in Parallel Ensemble Methods, base learners are usually generated in parallel such as random forest. We used a parallel ensemble technique with four base learners KSVM, Random Forest, Bagging and BayesGLM. Before moving to the proposed ensemble design, lets discuss each classifier used in detail.

3.4.1 KSVM

Support vector machine is a machine learning algorithm that helps to find the linear separating hyper plane with widen margin. Input data is viewed as set of two vectors in an n-dimensional space, an SVM will construct a separating hyper-plane, with maximum margin between positive and negative samples . For dataset with large number of features and small number of samples, many such hyperplanes exist. The points that lie closest to this max-margin hyperplane are called the support vectors. The support vectors are used by the classifier to classify test samples.

Since SVM tries to find linear plane, it's not very convenient in cases of most of the dataset with non-linear decision boundaries. It’s variant KSVM helps to develop a function which can convert the feature space for linearity. The functions used are polynomial and radial basis function.

One of the interesting aspect is increasing feature space might run in over fitting problems for other methods but not KSVM. Decision territory is attained using the examples closest to the margin. The use of kernel with regularization resulted in improved performance.

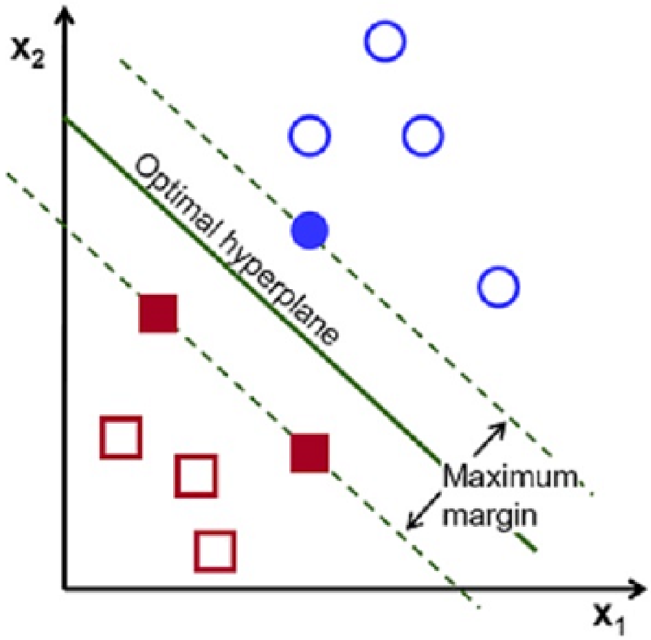


Figure 3.2: Support Vector Machine

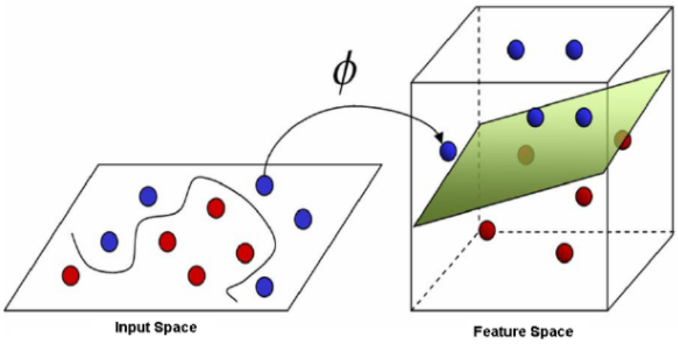


Figure 3.3: Non-Linear SVM with kernel trick

3.4.2 Random Forest

Random forest is an ensemble based classification algorithm which uses decision tree as classifiers [49]. Bootstrap sampling is used to built each classification tree and random subset of the variables are selected at each split for the candidate set. Thus, random forest uses both random variable selection and bagging for tree building. Random forest can achieve both low bias and low variance. Fig 3.4 shows a random forest classifier.

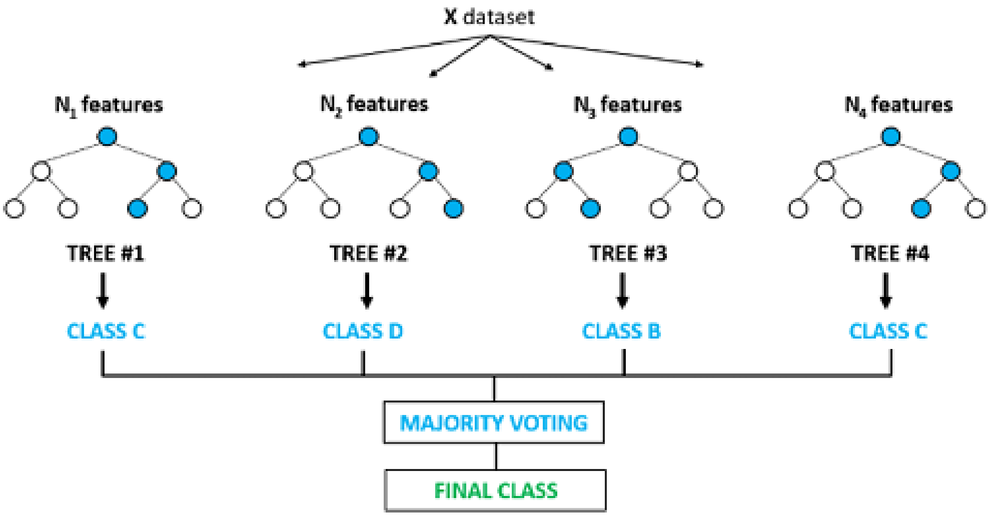


Figure 3.4: Random Forest Classifier

For classification, the performance of random forest is comparable to that of support vector machines. It is an ideal choice for microarray datasets because of the following characteristics:

- Can be used when there are large number of attributes as compared to samples
- Can show good performance in case of noisy variables
- Does not overfit

3.4.3 Bagging

Bagging is one of the variant of ensemble learning method. It runs on an interesting technique in which similar classifiers are trained on small subset of data and then averaging the predictions over all given learners. It is also an example of parallel ensemble and helps in reducing variance [50].

Bagging usually uses bootstrap sampling to obtain subsets of data for training the base classifiers. In bootstrap sampling, m new training sets each of size $n^{"}$ are generated from a given training set D of size n. This is done by sampling from D uniformly and with replacement due to which some observations may be repeated in each set. In case of regression , the outputs from the m models are averaged whereas voting is done in case of plain classification. As discussed above it gives an excellent advantage over single classifiers due to its ability to average votes from multiple decision making classifiers. Fig 3.5 illustrates how bagging classifier works.

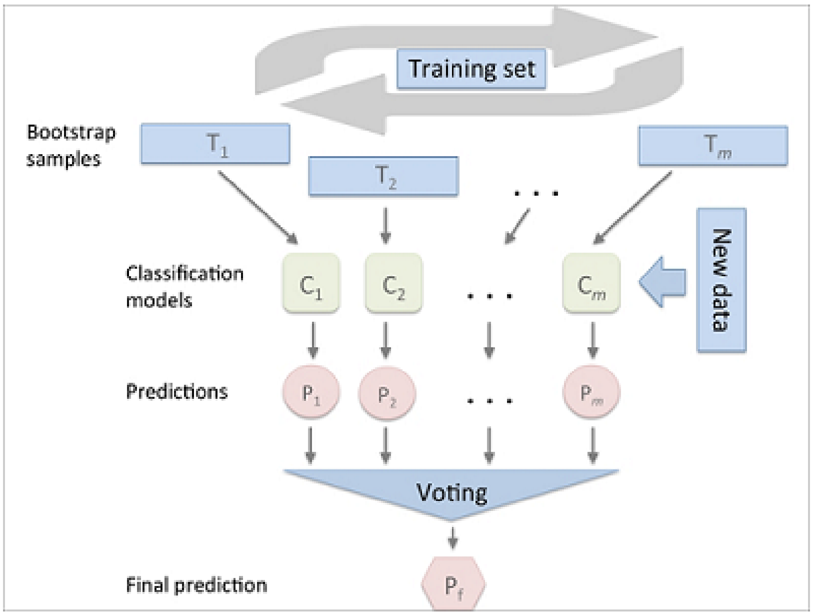


Figure 3.5: Structure of Bagging Ensemble

3.4.4 Bayesian Generalized Linear Model

Bayesian Model is one of the statistical method which used Bayesian theorem for classification [51]. It branches out as the data reveal more information regarding predictions or classification. Bayesian model works on five elements, which are, incorporating previous information, incorporating information with likelihood, utilizing posterior function of coefficient values, creating empirical distribution of most likely values and finally summarizing the empirical distribution. Bayesian Generalized Linear Model is shortcut to Bayesian way of utilizing prior information. It provides an empirical distribution which uses as inference for actual predictions.

$$Y = N(X\beta, \sigma^2) \tag{3.4}$$

Y is a random vector and all data points are distributed over normal distribution. Average of normal distribution is variance σ^2. It is similar to Bayesian model but has two major advantages.

- Priors: It can help one quantify prior information by placing priors on the parameters such as σ.
- Quantifying uncertainty: It can't be found to get an estimate of β as above but instead a complete posterior distribution about how likely different values of β are.

3.4.5 Proposed Ensemble

We have proposed a heterogeneous ensemble with four base classifiers Kernal Support Vector Machine (KSVM) , Random Forest , Bagging and Bayesian Generalized Linear Model (BayesGLM). It is created using Super Learners [52]. Super Learner is defined by a family of classifiers, risk for each classifier, and selection among all classifiers using cross-validation based estimation of risk. The benefit of using Super Learner is that you do not need to decide beforehand which technique to use. You can use multiple classification techniques by incorporating cross validation. In case some classifier do not contribute to the ensemble prediction power, it is automatically removed. The input training set is given to all four classifiers. Each classifier returns the prediction values for all the samples in test set. Cross-validation is used to estimate risk for all models. The weights are automatically assigned. For each dataset, Fig 3.6 shows how proposed ensemble works.

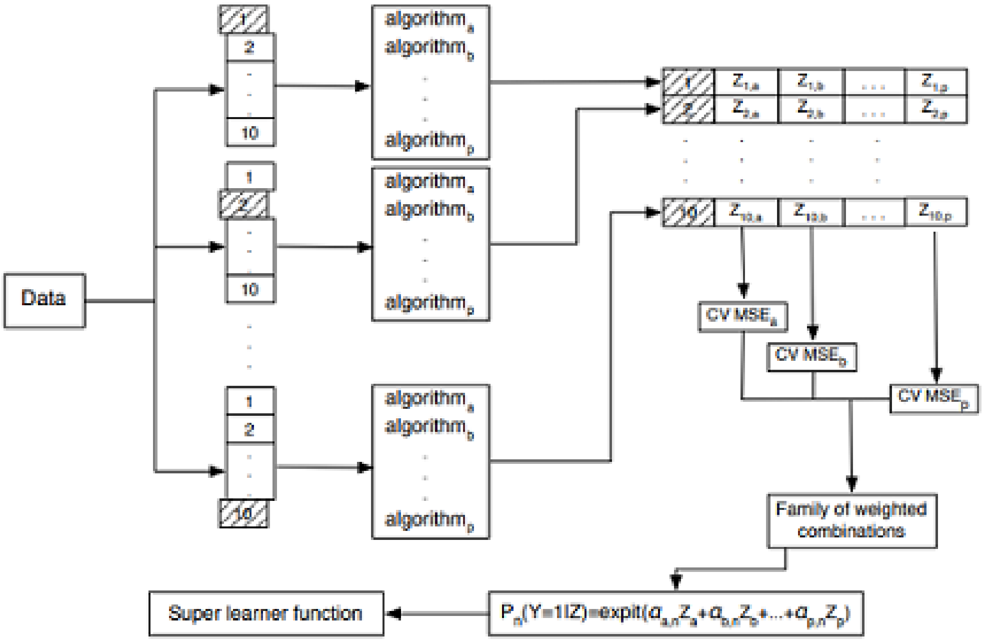


Figure 3.6: Design of Proposed Ensemble

Chapter 4

Implementation and Results

To implement the proposed methodology and other classifiers for comparison, R toolbox is used. For ensemble creation, we used SuperLearner package of R. SuperLearner provides the syntax and structure to combine different prediction algorithms and decide the optimal ensemble on the basis of data. Four benchmark cancer datasets used in this study are first preprocessed. For each of the 4 datasets, the complete dataset was used to generate training and test data randomly. Dataset is distributed by percentage of 70%:30% for training and test dataset respectively. Table 4.1 shows the sample distribution of all datasets for training and test set.

Table 4.1: Training and Test set Sample Distribution

Datasets		**Train**	**Test**	**Total**
Colon	Normal	15	7	22
	Tumor	29	11	40
	Total	44	18	62
Leukemia	ALL	36	11	47
	AML	15	10	25
	Total	51	21	72
Prostate	Healthy	37	12	49
	Cancer	35	18	53
	Total	72	30	102
Breast	DM	24	10	34
	NODM	31	13	44
	Total	55	23	78

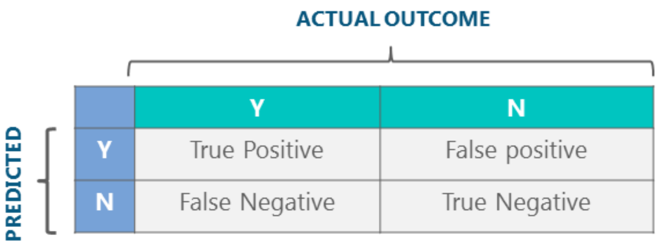
After that Information Gain , FCBF , CFS and ReliefF are applied on the dataset to select relevant features. Each method selects different number of features for each dataset. The details of selected features by all the methods for every dataset will be provided later in this chapter. These selected features are then used to re-create train and test dataset respectively. The newly created train set is then used to build model

4.1 Evaluation Criteria

This section discusses the evaluation metrics used to test and compare the performance of our approach with other classifiers already presented in literature.

4.1.1 Confusion Matrix

Also termed as error matrix. Its is a table which is used to identify or evaluate the performance of any classifier. This table consists of the values that were correctly or incorrectly identified. A simple example of confusion matrix can be explained as in Fig 4.1. 'True Positives' are the values that are correctly identified as positive. 'True Negatives' are the values that are correctly identified as negative. 'False Positives' are incorrectly identified as positive (i.e. the original value is negative but they are identified as positive). 'False Negatives' are incorrectly identified as negative (i.e. originally the values are positive).



ACTUAL OUTCOME

PREDICTED

	Y	N
Y	True Positive	False positive
N	False Negative	True Negative

Figure 4.1: Confusion Matrix Example

4.1.2 Accuracy

Accuracy is defined as a rate of measurement on how closely the predicted values are in accordance with the original/true values of the validation set. It is given as the total number of correctly identified samples out of the total number of samples. Mathematically it is expressed as:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{4.1}$$

where, T_P, T_N, F_P, F_N are *True Positive*, *True Negative*, *False Positive* and *False Negative* values respectively.

4.1.3 Sensitivity

Sensitivity is also known as 'True Positive rate' or 'Recall'. This criteria provides information about the actual number of correct positive values that are correctly predicted by the classifier. In clinical diagnosis, it as defined as the ability of a classifier to correctly identify those with disease. It is defined as:

$$Sensitivity = \frac{T_P}{T_P + F_N} \tag{4.2}$$

4.1.4 Specificity

Specificity is also called 'True Negative Rate'. It provides information about the actual number of correct negative values predicted by the classifier from the original number for negative values in the test data. In clinical diagnosis, it as defined as the ability of a classifier to correctly identify those without disease. Specificity is also defined as:

$$Specificity = \frac{T_N}{T_N + F_P} \tag{4.3}$$

4.2 Evaluation and Results

This sections provides an overview of all the evaluations performed on four microarray cancer datasets. For each dataset, multiple feature selectors and classifiers are used and results are compared based on the criteria in section 4. 10-fold cross validation is used to validate the performance of classifiers.

4.2.1 Colon Cancer Results

Table 4.2 shows the accuracies of all the classifiers with three feature selection methods calculated using K-Fold Cross Validation with K=10. Result shows that CFS outperforms as compared to other feature selection methods and give maximum accuracies with all classification methods. The number of features selected by CFS was 23 but no parameter setting or threshold is required. It selects features automatically. SVM radial, RF, BayesGLM and Proposed Ensemble give same accuracies of 94.4% for this dataset. Whereas, for other feature selection techniques, our Proposed Ensemble also gives comparable results.

Table 4.3 shows the sensitivity and specificity of Proposed Ensemble for Colon Cancer dataset

Table 4.2: % Accuracies of Different Approaches for Colon Cancer Dataset

Classifiers	**FCBF (15)**	**CFS (23)**	**IG (10)**
SVM	72.2	88.9	83.3
SVM (radial)	**83.3**	**94.4**	**94.4**
Random Forest	61.1	**94.4**	89.1
KNN	63.6	79.5	79.5
BayesGLM	77.8	**94.4**	91.7
Proposed Ensemble	**83.3**	**94.4**	**94.4**

Table 4.3: % Sensitivity and Specificity of Proposed Ensemble for Colon Cancer

Measure	FCBF	CFS	IG
Sensitivity	85.7	100	100
Specificity	81.8	90.9	90.9

Overall performance of all the Classifiers for Colon Cancer Dataset using FCBF, CFS and IG as feature selection techniques is shown in Fig 4.2. SVM radial and Proposed Ensemble gives accuracy of 94.4% with both IG and CFS, whereas random forest gives the same accuracy of 94.4% with CFS only.

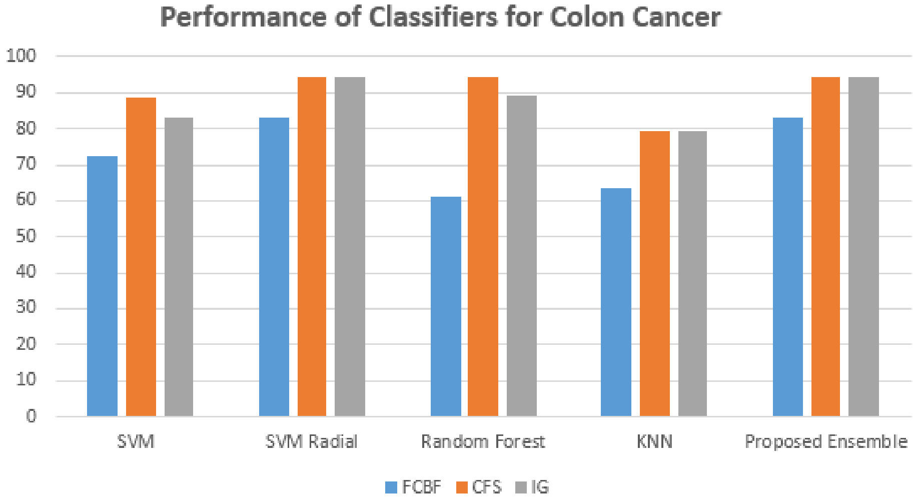


Figure 4.2: Performance of Classifiers for Colon Cancer Dataset using FCBF, CFS and IG as feature selection

4.2.2 Prostate Cancer Results

For prostate cancer dataset, our proposed Ensemble gives the maximum accuracy of 100% for all three feature selection techniques. For CFS and IG, random forest also gives same results of 100% accuracy.

Table 4.4 shows the accuracies of all the classifiers with three feature selection methods calculated using K-Fold Cross Validation with K=10.

Table 4.5 provides an idea that all tumor and healthy samples are correctly identified by our proposed approach.

Table 4.4: % Accuracies of Different Approaches for Prostate Cancer Dataset

Classifiers	**FCBF (15)**	**CFS (23)**	**IG (10)**
SVM	76.6	80	50
SVM (radial)	76.6	83.3	53.3
Random Forest	96.7	**100**	**100**
KNN	62.5	68.1	62.5
BayesGLM	70	86.7	50
Ensemble	**100**	**100**	**100**

Table 4.5: % Sensitivity and Specificity of Proposed Ensemble for Prostate Cancer

Measure	FCBF	CFS	IG
Sensitivity	100	100	100
Specificity	100	100	100

Fig 4.3 shows the performance of all the Classifiers for Prostate Cancer Dataset using FCBF, CFS and IG as feature selection techniques. Proposed Ensemble gives accuracy of 100% with FCBF, IG and CFS, whereas random forest gives the same

accuracy of 100% with CFS and IG only.

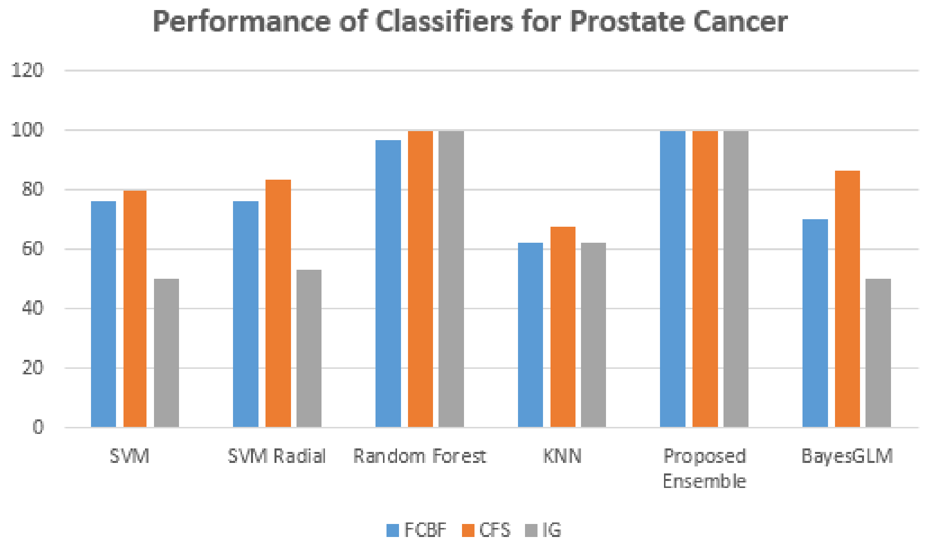


Figure 4.3: Performance of Classifiers for Prostate Cancer Dataset using FCBF, CFS and IG as feature selection

4.2.3 Leukemia Cancer Results

In case of Leukemia cancer dataset, multiple classifiers give maximum accuracy of 100% with CFS technique. For other feature selection techniques, proposed ensemble gives better results as compared to other classifiers. Table 4.6 gives the comparison of accuracies for all techniques.

Table 4.6: % Accuracies of Different Approaches for Leukemia Cancer Dataset

Classifiers	**FCBF (15)**	**CFS (46)**	**IG (15)**
SVM	85.7	90.4	94.1
SVM (radial)	80.9	**100**	94.1
Random Forest	80.9	95.2	92.4
KNN	70.5	98	90.1
BayesGLM	80.9	**100**	**95.2**
Ensemble	**81.5**	**100**	**95.2**

Table 4.7: % Sensitivity and Specificity of Proposed Ensemble for Leukemia Cancer

Measure	FCBF	CFS	IG
Sensitivity	100	100	100
Specificity	60.1	100	87.5

Fig 4.4 shows the graphical representation of performance of all classifiers for Leukemia Cancer Dataset using FCBF, CFS and IG as feature selection techniques. Proposed Ensemble performs almost same as bayesGLM for all three feature selection techniques. In case of FCBF, accuracy of proposed ensemble is slightly higher. SVM radial also gives maximum accuracy when feature selection is done by CFS.

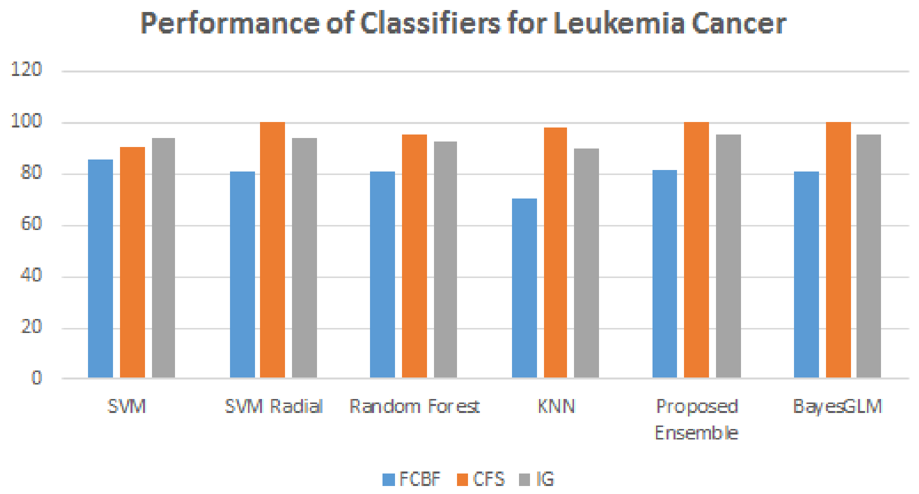


Figure 4.4: Performance of Classifiers for Leukemia Cancer Dataset using FCBF, CFS and IG as feature selection

4.2.4 Breast Cancer Results

Detailed results for classification of breast cancer is given in Table 4.8. KNN gives the best accuracy of 65.4% in case of FCBF. For other two techniques, CFS and IG, our proposed approach gives maximum accuracy of 91.3% and 82.6% respectively which is higher than any other classifier. Table 4.9 shows that in case of FCBF, our proposed technique doesn't correctly identify the cancer samples. However, for CFS and IG, tumor identification is much better.

Table 4.8: % Accuracies of Different Approaches for Breast Cancer Dataset

Classifiers	FCBF	CFS	IG
SVM	60.8	82.6	73.9
SVM (radial)	56.5	82.6	69.5
Random Forest	65.2	86.9	78.2
KNN	**65.4**	81.8	74.5
BayesGLM	60.8	82.6	73.9
Ensemble	60.8	**91.3**	**82.6**

Table 4.9: % Sensitivity and Specificity of Proposed Ensemble for Breast Cancer

Measure	FCBF	CFS	IG
Sensitivity	40	80	80
Specificity	76.9	100	84.6

Fig 4.4 shows the a graph for the performance of all classifiers for Leukemia Cancer Dataset using FCBF, CFS and IG as feature selection techniques. Proposed Ensemble outperforms in case of IG and CFS whereas KNN has maximum accuracy for FCBF.

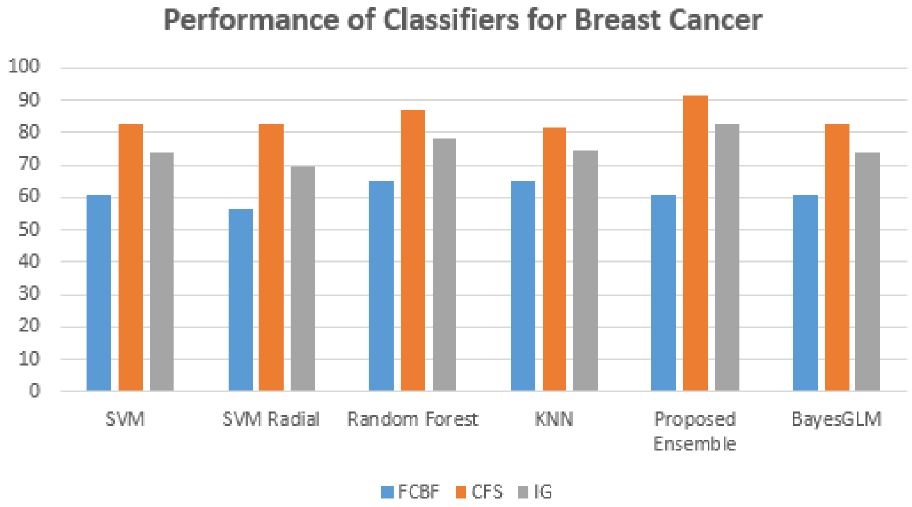


Figure 4.5: Performance of Classifiers for Breast Cancer Dataset using FCBF, CFS and IG as feature selection

4.2.5 Performance Analysis of Proposed Ensemble for all datasets

Comparison of results for our proposed technique with other classifiers for all datasets has been shown in Fig 4.6. It can be concluded from the graph that that our proposed ensemble selects the best classifier or combination of classifiers for every dataset, thus giving maximum accuracy in every case.

Random Forest works best in case of Colon and Prostate dataset while having low accuracy for other two datasets. SVM radial gives higher accuracy for Colon and Leukemia datasets whereas for Prostate and Breast cancer dataset, the results are not a s good as other techniques. Similarly BayesGLM, performs best for Colon and Leukemia dataset. For other two datasets, it doesn't give best accuracy. KNN has relatively low performance for all the datasets.

From the comparison graph, it can be easily summarized that no single classifier works best for all the four cancer datasets. Our approach combined the power of all the classifiers and give maximum results in all cases.

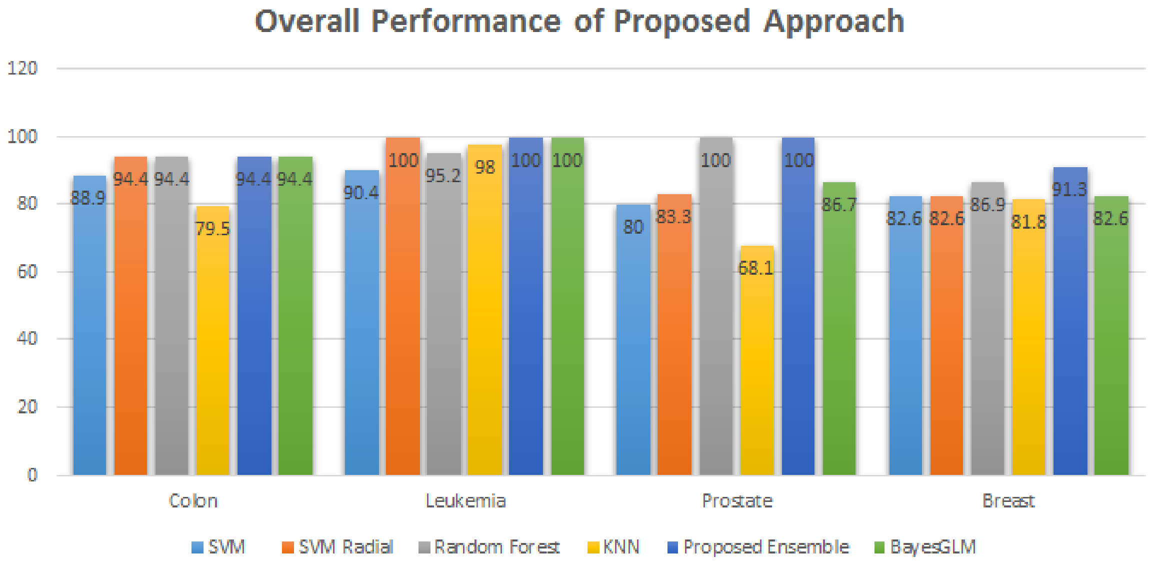


Figure 4.6: Overall Performance of Proposed Approach for all Datasets

4.2.6 Accuracy Comparison with other techniques in Literature

Table 4.10 shows the comparison of our proposed method with already published techniques. Our approach gives best results for Leukemia, prostate and breast cancer datasets. However, for colon dataset, the reported accuracy in literature is better. But our method gives comparable accuracy for colon also. To summarize the results, we can say that this ensemble technique works best for a range of microarray datasets and not specific to a particular dataset.

Table 4.10: Accuracy Comparison of with Other Techniques

Techniques	**Colon Cancer**	**Leukemia Cancer**	**Prostate Cancer**	**Breast Cancer**
Proposed Ensemble	94.4%	**100%**	**100%**	**91.3%**
Sara Haddou [32]	85%	97%	-	-
Rama [30]	**95.83%**	95.83%	-	-
Hanaa [36]	85.48%	97.06%	100%	-
Lingyun [37]	90.32%	98.61%	96.08%	-
Zhong [38]	90.5%	96.9%	-	-

Chapter 5

Conclusion

This research work proposes an ensemble based system to improve cancer classification prediction for microarray datasets. It uses CFS with forward search as feature selection. Instead of using a single classifier for different cancer datasets, a combination of best performing classifiers from literature is used to improve the classification performance. We applied the proposed technique along with other classifiers like SVM , KNN random forest, and BayesGLM to compare performance on four microarray datasets: colon, leukemia, prostate and breast cancer

To answer our first research question, we used a Correlation based feature selection (CFS) method to identify the subset of genes that helps in prediction of cancer. Section 3.3 shows how this feature selection work whereas the details of selected genes are provided in section 4.2.

For the efficient prediction of cancer, we used an ensemble based approach described in section 3.4 Results in section 4.2 show that our proposed approach gives the best performance of almost 100% for all the datasets and is comparatively better than simple classifiers or the techniques already discussed in literature.

It can also be concluded from the results that this approach gives best performance independent of dataset. This is due to the fact that it always selects the best performing algorithm or combination of algorithms according to data as explained in sub section 3.4.5.

5.1 Contributions

This thesis contributed to microarray based gene expression cancer classification in the following way. A unique combination of feature selection and classification algorithm is identified to improve the accuracy of cancer prediction. An ensemble based approach is designed to give better classification results independent of dataset.

5.2 Limitations

As for now, this research work does not does not take in account different types of cancers (multi-class problem). So, as an extension to this work, we will test our proposed approach for multi-class cancer datasets. Also, we tested this with only few feature selection methods and only four datasets, using other datasets with large size provides fair comparison of its performance.

5.3 Future Work

The results of this study shows that different variations of ensembles can be used to achieve better performance for the classification of a microarray cancer dataset. However, preprocessing and feature selection plays a major role in the stability and performance of classification model. Following are the suggested future directions:

- Different preprocessing techniques can be tried to improve the performance of classification.
- Since, the benchmark cancer datasets used in this study are small in terms of number of samples. The proposed approach can be tested for large cancer datasets.
- We only tried three feature selection techniques with the proposed ensemble.

Other feature selection techniques can be experimented to improve accuracy of classification.

- The problem we addressed in this thesis is binary classification problem. However, this work can be extended for multi class datasets.

References

[1] "Outline of detection method of genes by dna microarrays," http://www.3d-gene.com/en/about/chip/chi_003.html.

[2] V. Trevino, F. Falciani, and H. A. Barrera-Saldaña, "Dna microarrays: a powerful genomic tool for biomedical and clinical research," *Molecular Medicine*, vol. 13, no. 9-10, p. 527, 2007.

[3] H. J. Hong, W. S. Koom, and W.-G. Koh, "Cell microarray technologies for high-throughput cell-based biosensors," *Sensors*, vol. 17, no. 6, p. 1293, 2017.

[4] B. Wojtas, A. Pfeifer, M. Oczko-Wojciechowska, J. Krajewska, A. Czarniecka, A. Kukulska, M. Eszlinger, T. Musholt, T. Stokowy, M. Swierniak *et al.*, "Gene expression (mrna) markers for differentiating between malignant and benign follicular thyroid tumours," *International journal of molecular sciences*, vol. 18, no. 6, p. 1184, 2017.

[5] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[6] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[7] E. Pashaei, M. Ozen, and N. Aydin, "Gene selection and classification approach for microarray data based on random forest ranking and bbha," in *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*. IEEE, 2016, pp. 308–311.

[8] A. Gelman, A. Jakulin, M. G. Pittau, Y.-S. Su *et al.*, “A weakly informative default prior distribution for logistic and other regression models,” *The Annals of Applied Statistics*, vol. 2, no. 4, pp. 1360–1383, 2008.

[9] “Inca national cancer institute josÃĩ alencar gomes da silva (2016),” http://www.inca.gov.br/estimativa/2016/index.asp?ID=2.

[10] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart, “High density synthetic oligonucleotide arrays,” *Nature genetics*, vol. 21, no. 1s, p. 20, 1999.

[11] “Microarrays,” http://www.premierbiosoft.com/tech_notes/microarray.html.

[12] D. Ucar, I. Neuhaus, P. Ross-MacDonald, C. Tilford, S. Parthasarathy, N. Siemers, and R.-R. Ji, “Construction of a reference gene association network from multiple profiling data: application to data analysis,” *Bioinformatics*, vol. 23, no. 20, pp. 2716–2724, 2007.

[13] M. Dettling and P. Bühlmann, “Finding predictive gene groups from microarray data,” *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 106–131, 2004.

[14] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson *et al.*, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” *Nature medicine*, vol. 7, no. 6, p. 673, 2001.

[15] U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews *et al.*, “A gene expression database for the molecular pharmacology of cancer,” *Nature genetics*, vol. 24, no. 3, p. 236, 2000.

[16] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, “Molecular classifi-

cation of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[17] S. Student and K. Fujarewicz, "Stable feature selection and classification algorithms for multiclass microarray data," *Biology direct*, vol. 7, no. 1, p. 33, 2012.

[18] D. V. Nguyen and D. M. Rocke, "Classification of acute leukemia based on dna microarray gene expressions using partial least squares," in *Methods of Microarray Data Analysis.* Springer, 2002, pp. 109–124.

[19] A. Castaño, F. Fernández-Navarro, C. Hervás-Martínez, and P. A. Gutiérrez, "Neuro-logistic models based on evolutionary generalized radial basis function for the microarray gene expression classification problem," *Neural processing letters*, vol. 34, no. 2, p. 117, 2011.

[20] M. Kumar, N. K. Rath, A. Swain, and S. K. Rath, "Feature selection and classification of microarray data using mapreduce based anova and k-nearest neighbor," *Procedia Computer Science*, vol. 54, pp. 301–310, 2015.

[21] F. Chu and L. Wang, "Applications of support vector machines to cancer classification with microarray data," *International journal of neural systems*, vol. 15, no. 06, pp. 475–484, 2005.

[22] H. T. Huynh, J.-J. Kim, and Y. Won, "Dna microarray classification with compact single hidden-layer feedforward neural networks," in *fbit.* IEEE, 1899, pp. 193–198.

[23] R. Parry, W. Jones, T. Stokes, J. Phan, R. Moffitt, H. Fang, L. Shi, A. Oberthuer, M. Fischer, W. Tong *et al.*, "k-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction," *The pharmacogenomics journal*, vol. 10, no. 4, p. 292, 2010.

[24] K. Moorthy and M. S. Mohamad, “Random forest for gene selection and microarray data classification,” *Bioinformation*, vol. 7, no. 3, p. 142, 2011.

[25] J. M. Diaz, R. C. Pinon, and G. Solano, “Lung cancer classification using genetic algorithm to optimize prediction models,” in *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on.* IEEE, 2014, pp. 1–6.

[26] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, “Genetic bee colony (gbc) algorithm: A new gene selection method for microarray cancer classification,” *Computational biology and chemistry*, vol. 56, pp. 49–60, 2015.

[27] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis,” *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2004.

[28] A. M. Mahmoud, B. A. Maher, E.-S. M. El-Horbaty, and A. B. M. Salem, “Analysis of machine learning techniques for gene selection and classification of microarray data,” in *The 6th International Conference on Information Technology*, 2013.

[29] Z. M. Hira and D. F. Gillies, “A review of feature selection and feature extraction methods applied on microarray data,” *Advances in bioinformatics*, vol. 2015, 2015.

[30] R. S. Sreepada, S. Vipsita, and P. Mohapatra, “An efficient approach for classification of gene expression microarray data,” in *Emerging Applications of Information Technology (EAIT), 2014 Fourth International Conference of.* IEEE, 2014, pp. 344–348.

[31] C. D. A. Vanitha, D. Devaraj, and M. Venkatesulu, “Gene expression data classification using support vector machine and mutual information-based gene selection,” *procedia computer science*, vol. 47, pp. 13–21, 2015.

[32] S. H. Bouazza, N. Hamdi, A. Zeroual, and K. Auhmani, “Gene-expression-based cancer classification through feature selection with knn and svm classifiers,” in *Intelligent Systems and Computer Vision (ISCV), 2015*. IEEE, 2015, pp. 1–6.

[33] S. Alagukumar and R. Lawrance, “Classification of microarray gene expression data using associative classification,” in *Computing Technologies and Intelligent Data Engineering (ICCTIDE), International Conference on*. IEEE, 2016, pp. 1–8.

[34] R. Ocampo-Vega, G. Sanchez-Ante, M. A. de Luna, R. Vega, L. E. Falcón-Morales, and H. Sossa, “Improving pattern classification of dna microarray data by using pca and logistic regression,” *Intelligent Data Analysis*, vol. 20, no. s1, pp. S53–S67, 2016.

[35] S. Tarek, R. A. Elwahab, and M. Shoman, “Gene expression based cancer classification,” *Egyptian Informatics Journal*, vol. 18, no. 3, pp. 151–159, 2017.

[36] H. Salem, G. Attiya, and N. El-Fishawy, “Classification of human cancer diseases by gene expression profiles,” *Applied Soft Computing*, vol. 50, pp. 124–134, 2017.

[37] L. Gao, M. Ye, X. Lu, and D. Huang, “Hybrid method based on information gain and support vector machine for gene selection in cancer classification,” *Genomics, proteomics & bioinformatics*, vol. 15, no. 6, pp. 389–395, 2017.

[38] Z. Wenyan, L. Xuewen, and W. Jingjing, “Feature selection for cancer classification using microarray gene expression data,” *Biostat 03 Biometrics Open Acc J*, vol. 1, no. 2, p. 555557, 2017.

[39] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[40] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie *et al.*, “Gene expression correlates of clinical prostate cancer behavior,” *Cancer cell*, vol. 1, no. 2, pp. 203–209, 2002.

[41] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton, A. T. Witteveen *et al.*, “Gene expression profiling predicts clinical outcome of breast cancer,” *nature*, vol. 415, no. 6871, p. 530, 2002.

[42] F. Rafii, M. A. Kbir, and B. D. Rossi, “Data preprocessing and reducing for microarray data exploration and analysis,” *International Journal of Computer Applications*, vol. 132, no. 16, pp. 20–26, 2015.

[43] M. Dettling, “Bagboosting for tumor classification with gene expression data,” *Bioinformatics*, vol. 20, no. 18, pp. 3583–3593, 2004.

[44] K. Yang, Z. Cai, J. Li, and G. Lin, “A stable gene selection in microarray data analysis,” *BMC bioinformatics*, vol. 7, no. 1, p. 228, 2006.

[45] A. A. Antipova, P. Tamayo, and T. R. Golub, “A strategy for oligonucleotide microarray probe reduction,” *Genome Biology*, vol. 3, no. 12, pp. research0073–1, 2002.

[46] L. Yu and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” *Journal of machine learning research*, vol. 5, no. Oct, pp. 1205–1224, 2004.

[47] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. Mayer, and H. W. Mewes, "Gene selection from microarray data for cancer classificationâĂŤa machine learning approach," *Computational biology and chemistry*, vol. 29, no. 1, pp. 37–46, 2005.

[48] K. Michalak and H. Kwaśnicka, "Correlation-based feature selection strategy in classification problems," *International Journal of Applied Mathematics and Computer Science*, vol. 16, pp. 503–511, 2006.

[49] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[50] L. Brieman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[51] S. L. Zeger and M. R. Karim, "Generalized linear models with random effects; a gibbs sampling approach," *Journal of the American statistical association*, vol. 86, no. 413, pp. 79–86, 1991.

[52] M. J. Van der Laan, E. C. Polley, and A. E. Hubbard, "Super learner," *Statistical applications in genetics and molecular biology*, vol. 6, no. 1, 2007.