# MICROARRAY BASED CANCER CLASSIFICATION USING SIGNAL PROCESSING AND NEURAL NETWORK

A Thesis submitted to Goa University for the award of Degree of

**DOCTOR OF PHILOSOPHY**

**IN**

**ELECTRONICS**


by

**Mrs. SUPRIYA PATIL**

Guide

**Dr. G. M. NAIK**

Professor and H. O. D., Department of Electronics,

Goa University, Goa.

Co Guide

**Dr. K. R. Pai**

Ex-Professor and H.O.D, E.T.C Department,

Padre Conceicao College of Engineering., Verna, Goa.

March, 2018

**Department of Electronics**
**Goa University**

# CERTIFICATE

This is to certify that the thesis titled **"Microarray based Cancer Classification using Signal Processing and Neural Network"** submitted by Mrs. Supriya Ashok Patil for the award of degree of Doctor of Philosophy in Electronics, is based on original and independent work carried out by her during the period of the study under my supervision. The thesis wholly or in part has not been submitted for any other degree or diploma in any university or Institute.

**Place : Goa University** **Prof. G. M. Naik**
**Date : March 2018** **Research Guide.**

**Department of Electronics**
**Goa University**

# DECLARATION

I hereby declare that the present thesis entitled **"Microarray based Cancer Classification using Signal Processing and Neural Network"** is my contribution and same has not been submitted on any occasion for any other degree or diploma of this or any other university/institute. To the best of my knowledge, the proposed study is the comprehensive work in the area mentioned. The literature related to the problem investigated has been cited. Due acknowledgements have been made wherever facilities and suggestions have been availed of. Entire work embodied in this doctoral thesis has been carried out by me at the Padre Conceicao College of Engineering, Verna, Goa and Department of Electronics, Goa University, Goa, under the supervision of Prof. G. M. Naik.

**Mrs. Supriya Patil**
**Research scholar.**

# Acknowledgements

This work is possible due to the efforts of many. I am indebted to each one of them!

I express my profound gratitude to my guide, Prof. G. M. Naik, H.O.D. , Electronics Department, Goa University. He has made time for me during his busiest schedules and generously spent time in understanding the problem at hand and suggesting innovative methods to get to a solution. The confidence he expressed in me, was much more than what even I had in myself; and it has been my greatest motivation in keeping me going through the toughest times in the course of my research work.

I also extend my heartfelt gratitude to Prof. K. R. Pai, Ex- H.O.D. of E.T.C. Department of P. C. College of Engineering, Goa. He has been my mentor and my greatest support in all my technical requirements. Without his backing, this work would have been an impossible task to accomplish! I am thankful for all the time Prof. K. R. Pai slated for me during which he corrected and suggested many methods, giving a different perspective of how we could approach the problem more efficiently. Many of the ideas proposed in this work have been the fruit of such meticulous, deliberate analysis and discussions with him, for which I am deeply indebted!

I am thankful to Dr. R. S. Gad for his support and guidance during the course of my doctoral work.

I would like to thank Prof. Sanjeev C. Ghadi, Professor, Department of Biotechnology, Goa University and Mr. Pravinkumar M. K., Research Scholar, Department of Zoology for their timely guidance in correlating the proposed study with Genomics.

I would also like to express my thanks to Goa University for having this Ph.D. program. A special mention of thanks to the administrative staff handling the Ph.D.

section, for always being obliging. I also thank the Electronics Department of Goa University for extending their support willingly to ensure smooth conduct of the presentations and helping with other logistics whenever required.

I am grateful to the management and Principal of Padre Conceicao College of Engineering for permitting me to pursue my doctoral studies. I am grateful to my departmental colleagues, who have eagerly extended their support in whatever way possible to make my journey less taxing.

I would like to specially thank Prof. Geetalaxmi. K. for the various insights and correction provided with regards to paper writing and submissions. I am also grateful to Dr. Jayalaxmi Devate who has selflessly guided me with the structuring of my thesis appropriately. My thanks are also due to Ms. Rohini Korti, who has helped me immensely with Latex. I sincerely thank our Lab Assistants, Ms. Suzette Dourado and Ms Veena Gomes, who have been ever ready to fix any glitches at the work station and for all the other technical support provided. I am grateful to Ms. Merlyn D'souza for the grammar check and the read back service provided at the earliest despite of her busy schedule.

I owe the strength and perseverance it took to complete this work to my family- my mother, husband and children! No words would be enough to express my gratitude and appreciation to my husband for making sure I get quality time to complete this work by taking up many of our responsibilities single handedly. I am proud of my children for showing maturity beyond their age and giving me confidence that they would be fine even with less attention from my side. I cannot thank my parents and in-laws enough for their silent backing during this work.

And most importantly, I thank the almighty God who blessed me, been with me, guided me to the successful and fulfilling completion of my Ph.D. work.

**Mrs. Supriya Patil.**

# Dedication

Dedicated with Extreme Gratitude and Affection to

**My Mother (Mrs. Kusum Ashok Patil)**

The rock solid support behind me.

# Abstract

Cancer is considered as one of the most dreadful diseases, accounting for a major share in the incidence of death at the global level and the number of cancer cases are burgeoning at an alarming rate. The transformation in cancer triggering or cancer quashing genes leads to cancer. An accurate screening of the cancer subtype helps in administering the appropriate therapy and speedy recovery, ensuing considerable increase in the survival rate of cancer patients. Customary procedures for cancer screening are predominantly influenced by the skill of the oncologist which at best is subjective. Moreover, cancer screening using biomarkers fails occasionally. Therefore, it is necessary to develop a method to improve the accuracy of cancer classification using small number of genes. Microarray technology facilitates the simultaneous examination of all types of gene mutations in human body for disease identification. Further, it automizes the process of disease identification. One of the important utilization of Microarray technology is cancer recognition.

Main objective of this research work is to design an efficient system for microarray gene expression based cancer classification with optimum number of genes, particularly at higher malignancy levels at which genes of various cancer classes are less distinctly expressed. The gene expression data is obtained as a result of microarray experiment, image processing and quantification process. In the proposed work, the classification of Grade III and Grade IV Glioma is implemented using GDS1975, GDS1976, GDS1815 and GDS1816 microarray datasets. The dimension reduction of these Glioma datasets is accomplished by using feature selection method followed by feature extraction method. Three different gene selection methods applied are Thresholding method, Ratio method and Fusion of Thresholding and Ratio method.

Thresholding method is used to select the genes with consistent intensity variation within the chosen values of threshold. Ratio method is used to choose the genes with small values of maximum to minimum gene intensity ratio which are more appropriate for classification. The Fusion of Thresholding and Ratio method helps to select the genes with diminished ratio within the best performing threshold range. The Fusion of Thresholding and Ratio method provides smaller and more suitable gene subset for cancer classification as compared to Thresholding and Ratio method applied individually. The features of the data obtained by feature selection method are extracted with the help of Discrete Wavelet Transform (DWT). The performance of Thresholding method, Ratio method and Fusion of Thresholding and Ratio method in combination with DWT based feature extraction is compared with the help of various classification algorithms namely, Resilient Back Propagation (RPROP), Levenberg Marquardt (LM), Conjugate Gradient and Stacked Autoencoder (SAEN).

The proposed system outperforms the existing techniques used for classification of Glioma datasets giving 100% classification accuracy with only five genes. Moreover, mutations in the obtained optimal gene subset is found to be directly or indirectly linked with the occurrence of Glioma. Further, testing of this optimal gene subset for Brain tumor dataset GDS1962 at various level of malignancies delivers 100% classification accuracy.

# List of Abbreviations and Symbols

**Abbreviations**

| | |
|---|---|
| $A$ | Adenine |
| $ANKRD17$ | Ankyrin Repeat Domain 17 |
| $ANN$ | Artificial Neural Network |
| $C$ | Cytosine |
| $cDNA$ | Complementary DNA |
| $CFS$ | Correlation based Feature Selection |
| $CGFR$ | Conjugate Gradient with Fletcher-Reeves Update |
| $CGPB$ | Conjugate Gradient with Powell-Beale Restarts |
| $CGPR$ | Conjugate Gradient with Polak-Ribire Update |
| $CWT$ | Continuous Wavelet Transform |
| $DCT$ | Discrete Cosine Transform |
| $DNA$ | Deoxynucleic Acid |
| $EBPA$ | Error Back Propagation Algorithm |
| $EGSG$ | Ensemble Gene Selection by Grouping |
| $FSVM$ | Fuzzy Support Vector Machine |

| | |
|---|---|
| $FT$ | Fourier Transform |
| $G$ | Guanine |
| $GA$ | Genetic Algorithm |
| $GADP$ | Genetic Algorithm with Dynamic Parameter Setting |
| $GR$ | Gain Ratio |
| $HPF$ | High Pass Filter |
| $IG$ | Information Gain |
| $LM$ | Levenberg Marquardt |
| $LPF$ | Low Pass Filter |
| $MF$ | Multivariate Filter |
| $MF-GE$ | Genetic Ensemble with Multiple Filter |
| $MFMW$ | Multiple Filter Multiple Wrapper |
| $MORF4L2$ | Mortality Factor 4 Like 2 |
| $MRMR$ | Minimum Redundancy Maximum Relevance |
| $mRNA$ | Messenger Ribonucleic Acid |
| $MSE$ | Mean Square Error |
| $PCA$ | Principal Component Analysis |
| $PCR$ | Polymerase Chain Reaction |
| $PKB3$ | Protein Kinase B3 |
| $PSNR$ | Peak Signal to Noise Ratio |
| $RF$ | Random Forest |

| | |
|---|---|
| $RPROP$ | Resilient Back Propagation |
| $SAEN$ | Stacked Autoencoder |
| $SRP14$ | Signal Recognition Particle 14 |
| $STFT$ | Short Time Fourier Transform |
| $SVM-RFE$ | Support Vector Machine- Recursive Feature Elimination |
| $T$ | Thymine |
| $TS$ | T-statistics |
| $UF$ | Univariate Filter |
| $ZNF550$ | Zinc Finger Protein 550 |
| DWT | Discrete Wavelet Transform |

**Symbols**

| | |
|---|---|
| $Db$ | Daubecies wavelet |
| $\alpha(s)$ | Positive constant |
| $\beta$ | Scale parameter for thresholding |
| $\Delta v_{kj}$ | Change in weight $v_{kj}$ |
| $\Delta$ | Weight control parameter |
| $\mu$ | Fusion coefficient |
| $\psi((b-si)/sc)$ | Mother wavelet |
| $\sigma$ | Standard Deviation of Noise in an image |
| $A(s)$ | DCT coefficients |
| $ai$ | $i^{th}$ input to neural network |

| | |
|---|---|
| $B$ | Identity matrix |
| $b$ | Gene number |
| $Bior$ | Bio-orthogonal wavelet |
| $bk$ | Error vector |
| $c$ | Learning constant |
| $Coif$ | Coiflet wavelet |
| $D$ | Input to Autoencoder |
| $d$ | Direction of steepest descent |
| $DS2$ | Down sampling by two |
| $EC$ | Energy in the current gradient |
| $ek$ | Expected output of $k^{th}$ output layer neuron |
| $EP$ | Energy in the previous gradient |
| $F$ | New search direction |
| $g$ | Multiplicative factor |
| $h$ | Previous search direction |
| $HH1$ | Wavelet detailed diagonal coefficients at decomposition level 1 |
| $HL1$ | Wavelet detailed horizontal coefficients at decomposition level 1 |
| $I$ | Reference microarray image |
| $J_k$ | Jacobian matrix |
| $K$ | De-noised microarray image |
| $l$ | Wavelet decomposition level |

| | |
|---|---|
| $LH1$ | Wavelet detailed vertical coefficients at decomposition level 1 |
| $LL1$ | Wavelet approximation coefficients at decomposition level 1 |
| $M$ | Thresholded wavelet detailed coefficients |
| $m \times e$ | Number of image pixels |
| $Max(I)$ | Maximum intensity in an image matrix |
| $N$ | Number of wavelet coefficients |
| $P$ | Length of the signal |
| $pl(si)$ | Detailed wavelet coefficients |
| $ql(si)$ | Approximate wavelet coefficients |
| $r(n)$ | Impulse response of low pass filter |
| $Rbio$ | Reverse bio-orthogonal wavelet |
| $s(n)$ | Impulse response of High pass filter |
| $Sym$ | Symlet wavelet |
| $thd$ | Threshold |
| $U$ | Weight vector of encoder |
| $UP2$ | Up sampling by two |
| $V$ | Weight vector of decoder |
| $W$ | Window function of Short Time Fourier Transform |
| $wd$ | Wavelet detailed coefficients before thresholding |
| $Y$ | Transformed input at the output of hidden layer of an autoencoder |
| $Y$' | Reconstructed input |

| | |
|---|---|
| $Y(\tau, W)$ | Short Time Fourier Transform Coefficient at frequency f |
| $y(b)$ | Gene expression data sample |
| $Y(f)$ | Fourier Transform coefficient at frequency f |
| $Y(sc, si)$ | CWT coefficient at scale parameter sc and shift parameter si |
| ${\sigma_x}^2$ | Noisy image intensity variance |
| ${\sigma_x}^2$ | Reference image intensity variance |
| $z_k$ | Actual output of $k^{th}$ neuron in the output layer |
| $z_k$' | Derivative of actual output of $k^{th}$ neuron in output layer neuron |
| $j_k$ | Number of wavelet coefficients |
| $u_{ji}$ | Weight that connects $i^{th}$ input to $j^{th}$ neuron in the hidden layer |
| $v_{kj}$ | Weight that connects output of $j^{th}$ neuron in the hidden layer to $k^{th}$ neuron in the output layer |
| $y_j$ | Output of $j^{th}$ hidden layer neuron |
| $y_k$' | Derivative of output of $j^{th}$ hidden layer neuron with respect to $v_{kj}$ |

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Based on the data published by World Health Organization, cancer is the second leading reason of death universally. In 2015, about 8.8 million deaths were caused by cancer. Further, in next two decades, about 70% increase is anticipated in cancer cases [1].

Therefore it is of utmost importance to improve the techniques to avert, detect and treat cancer. There are about trillions of cells in a human body viz., skin cells, nerve cells etc. A gene is an orderly arrangement of nucleotides (Adenine (A), Thymine (T), Guanine (G) and Cytosine (C)) inside the nucleus of the cell. Genes control the lifecycle of cells via division, growth and death of the cells. Moreover, it provides the direction for the cells to generate proteins that are essential for proper functioning of the human body. For example, red blood cells produce hemoglobin required to carry oxygen from lungs to other body parts and carbon dioxide from other body parts to lungs while, skin cells like Melanocytes produce melanin that gives a color to the skin. Due to various reasons when the genes start mutating, the life cycle of cells gets disturbed. The mutations in the gene may be germinal (hereditary) or somatic (acquired). The germinal mutations are passed from generations to generations while, somatic mutations are due to life style habits such as smoking, diet, excessive exposure to the radiations, carcinogenic materials, obesity etc. Genetic mutations lead to

uncontrolled growth of cells which, in turn, affect the amount of the protein produced by the cells and causes cancer. In other words, cancer develops when certain genes (cancer causing or cancer suppressing) in the human body start mutating. Because of its awful impact on human being, it has become one of life′s major threats [2].

There are number of types of cancer such as Brain cancer, Breast cancer, Lung cancer etc. and every cancer has number of sub-types. As an example the Brain tumor sub-types are shown in the Figure 1.1

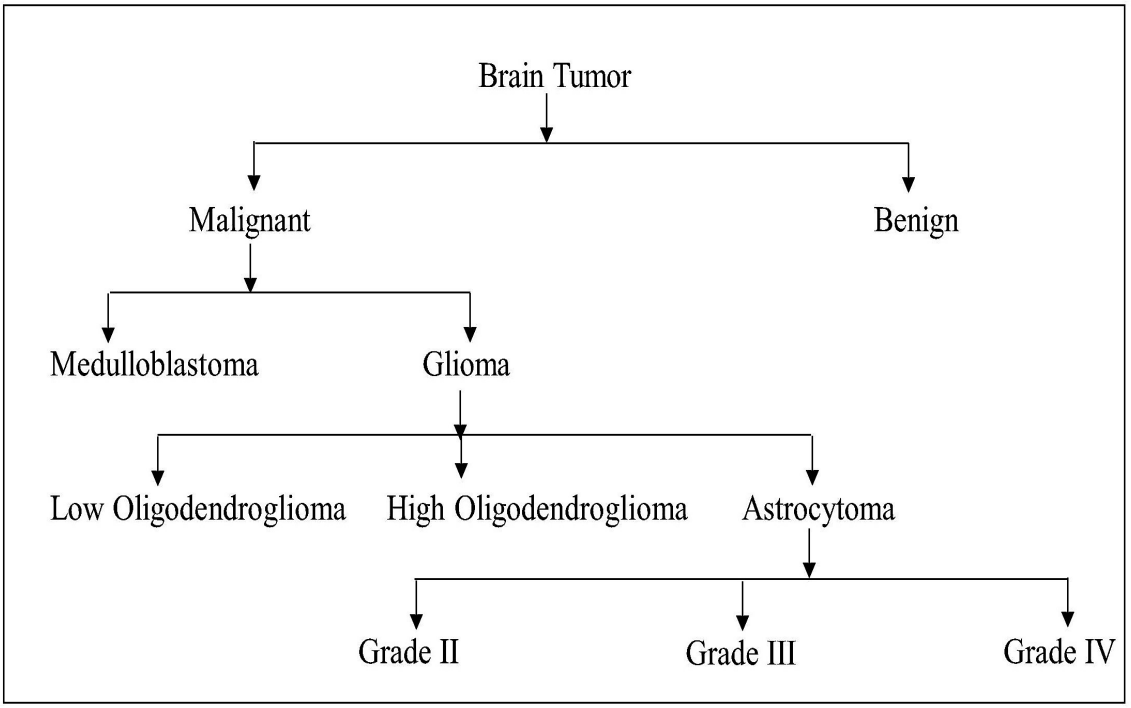


Figure 1.1: Sub-types of Brain tumor

The cancerous and non-cancerous samples of Brain tumor have gene intensities wide apart from each other, making it easy to differentiate between them.

Figure 1.2 demonstrates the variations in average intensity of 30 genes for Malignant and Benign Brain tumors [3].

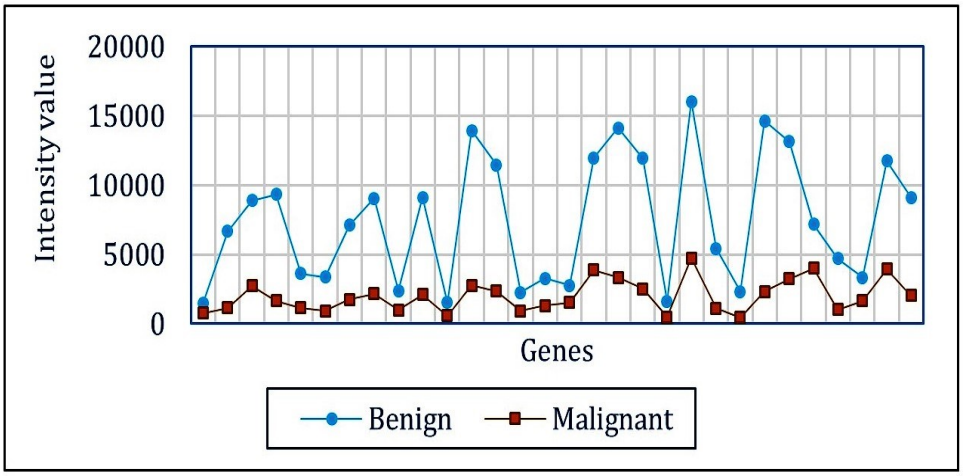


Figure 1.2: Variation in average intensity of 30 genes for Malignant and Benign Brain tumor.

However, with increase in the level of malignancy, genes become less differentially expressed, making the classification an uphill battle. Grade III and grade IV Glioma Brain tumors are the sub-types of the Brain tumor at higher malignancy level. Consequently, it is a real challenge to singularize them. Figure 1.3 presents the variation in average intensity of 30 genes for Glioma Grade III and Grade IV samples [4].

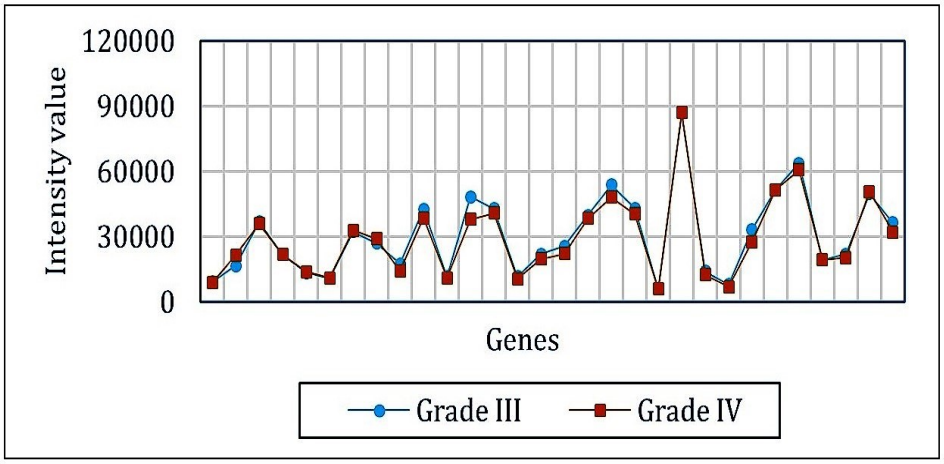


Figure 1.3: Variation in average intensity of 30 genes for Glioma Grade III and Grade IV samples.

Correct recognition of the sub-type of a cancer is crucial in determining the prognosis and planning of the treatment. The survival rate of the cancer patients can be enhanced by precise diagnosis of cancer sub-type. Therefore cancer classification has become one of the most important research areas in the biomedical field. However, conventional techniques of cancer diagnosis depend largely on the experience and skill of the physician. Genetic mutations being the basis of occurrence of a disease, detection of a disease can be efficiently accomplished by monitoring genetic mutations. To facilitate the monitoring of large number of genes (about 25,000) in a human body at once, microarray technology is most efficient technique.

## 1.1 Microarray technology

Introduction of Microarray technology in 1990 has rendered it possible to assess and analyze the variations in expression levels of entire genome in a single experiment. It caters varieties of applications in bio-medical field like gene discovery, disease diagnosis, drug discovery, human identification etc. Cancer prediction and recognition is one of the most important applications of Microarray technology. It automizes the process of cancer identification and assists in accurate cancer diagnosis [5], [6].

Microarray technology has four major steps:

1. Chip fabrication
2. Experiment
3. Image processing
4. Data analysis

The block diagram of Microarray technology is shown in the Figure 1.4.

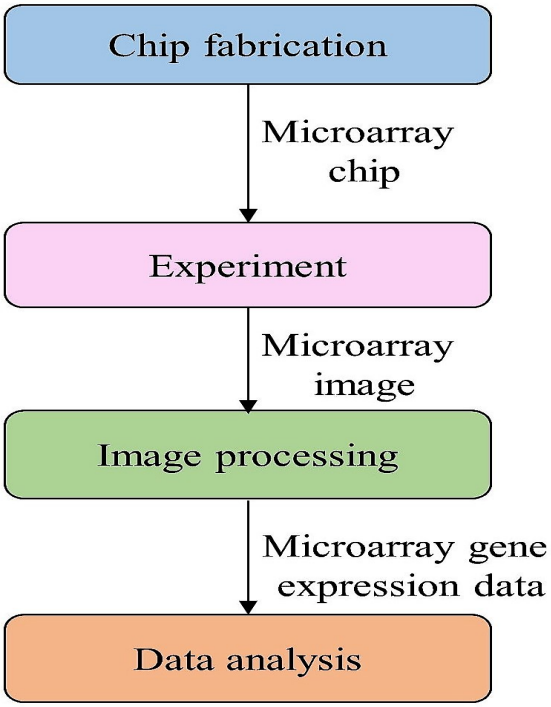


Figure 1.4: Block diagram of Microarray technology

The various steps in the Microarray technology are explained below.

### 1.1.1 Chip fabrication

Microarray technology makes use of a chip that may be either Glass Deoxynucleic Acid (DNA) microarray or Oligonucleotide microarray. These microarrays differ from each other with respect to the length of fragments of DNA to be attached to the chip, the way in which the DNA fragments are to be printed onto the microarray chip and the format of image to be generated. The fabrication of Glass DNA microarray and Oligonucleotide microarray is explained below:

#### Glass DNA microarray

Glass DNA microarrays were first fabricated at Stanford University by Patrick Brown and his teammates in 1990. Glass DNA microarrays normally use the glass slide with

some specific characteristics like best mechanical stability, good resistance against solvents etc. The steps in the fabrication of Glass DNA microarray are described below:

1. The required genes are collected from the public repositories, public databases or from corresponding institutions.

2. The glass base is coated with non-florescent material for proper attachment of the DNA fragments obtained from the genes onto the glass slide.

3. DNA fragments are purified, amplified and finally attached onto the glass slide in an orderly pattern.

4. The attachment of the DNAs is carried out with the help of inkjet, robotic or contact printing.

5. Finally, for a proper binding of DNA fragments the microarray chip is cooled at room temperature and exposed to ultraviolet light to reduce the effect of the background intensity on the microarray data.

Alternatively, complementary DNA (cDNA) or Polymerase Chain Reaction (PCR) product can also be used instead of DNA fragments. A typical Glass DNA Microarray contains 10,000 to 20,000 DNA strands in an area of 3.6 $cm^2$ with the length of the DNA being about 17-25 mers (orderly arrangement of nucleotides). The typical spot size for Glass DNA microarray is 10 $\mu$m. The fabrication of Glass DNA microarrays does not require any specialized equipment. They are cheaper and offer greater detection sensitivity. However, it requires more manpower for synthesis, purification and storage of DNA solutions prior to chip fabrication [5], [6], [7], [8].

Figure 1.5 demonstrates the fabrication of Glass DNA microarray and a typical Glass DNA microarray chip is presented in the Figure 1.6.

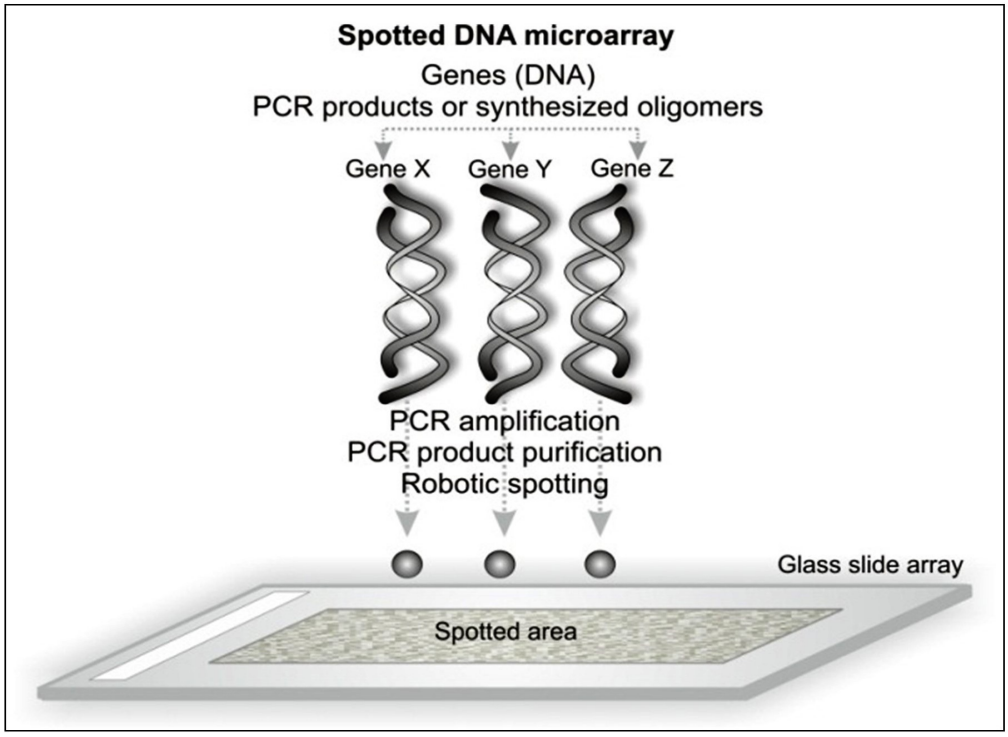


Figure 1.5: Fabrication of Glass DNA microarray.

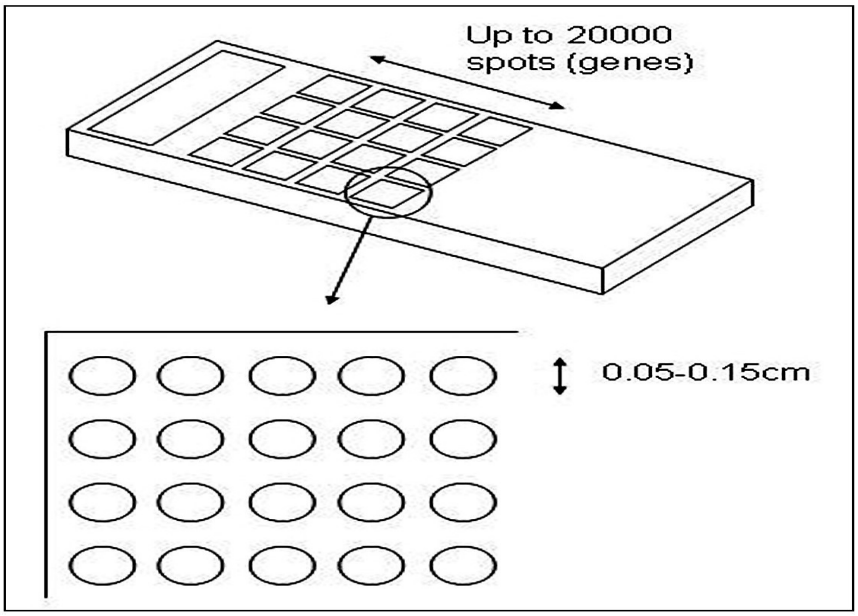


Figure 1.6: Glass DNA microarray.

**In Situ Oligonucleotide DNA microarray**

In Situ Oligonucleotide microarrays were first fabricated by Stephen Fodor and teammates in 1991. First the strands of DNA are chemically synthesized and subsequently the nucleotides (A, C, G and T) of the corresponding DNA strand are attached one by one to the chip using photolithography process. As an example, one of the step of fabrication assuming the requirement of Thymine in the first position for some of the spots on the quartz crystal is explained below [5], [6], [7], [8].

1. For attachment of Thymine, the spot positions that require Thymine nucleotide in the first position are identified.
2. The spot position that does not require Thymine nucleotide in the first position are protected with the help of mask.
3. The solution that contains Thymine nucleotide is poured on the quartz chip.
4. The quartz crystal is washed in order to remove the residuals.
5. Above procedure is repeated until synthesis of complete DNA strand.

Oligonucleotide DNA microarray typically contains about 50,000 DNA strands in an area of 1.28 $cm^2$ with the length of DNA strands being about 25 mers. In Situ Oligonucleotide microarrays offer more fabrication speed and increased reproducibility as compared to Glass DNA microarrays. However, the requirement of specialized and costly equipment for hybridization, label staining, washing and quantization processes, makes the In Situ Oligonucleotide microarrays expensive. Further, it has decreased detection sensitivity [5], [6].

The fabrication process of In Situ DNA microarray and In Situ DNA microarray chip are respectively presented in Figure 1.7 and Figure 1.8 [9].

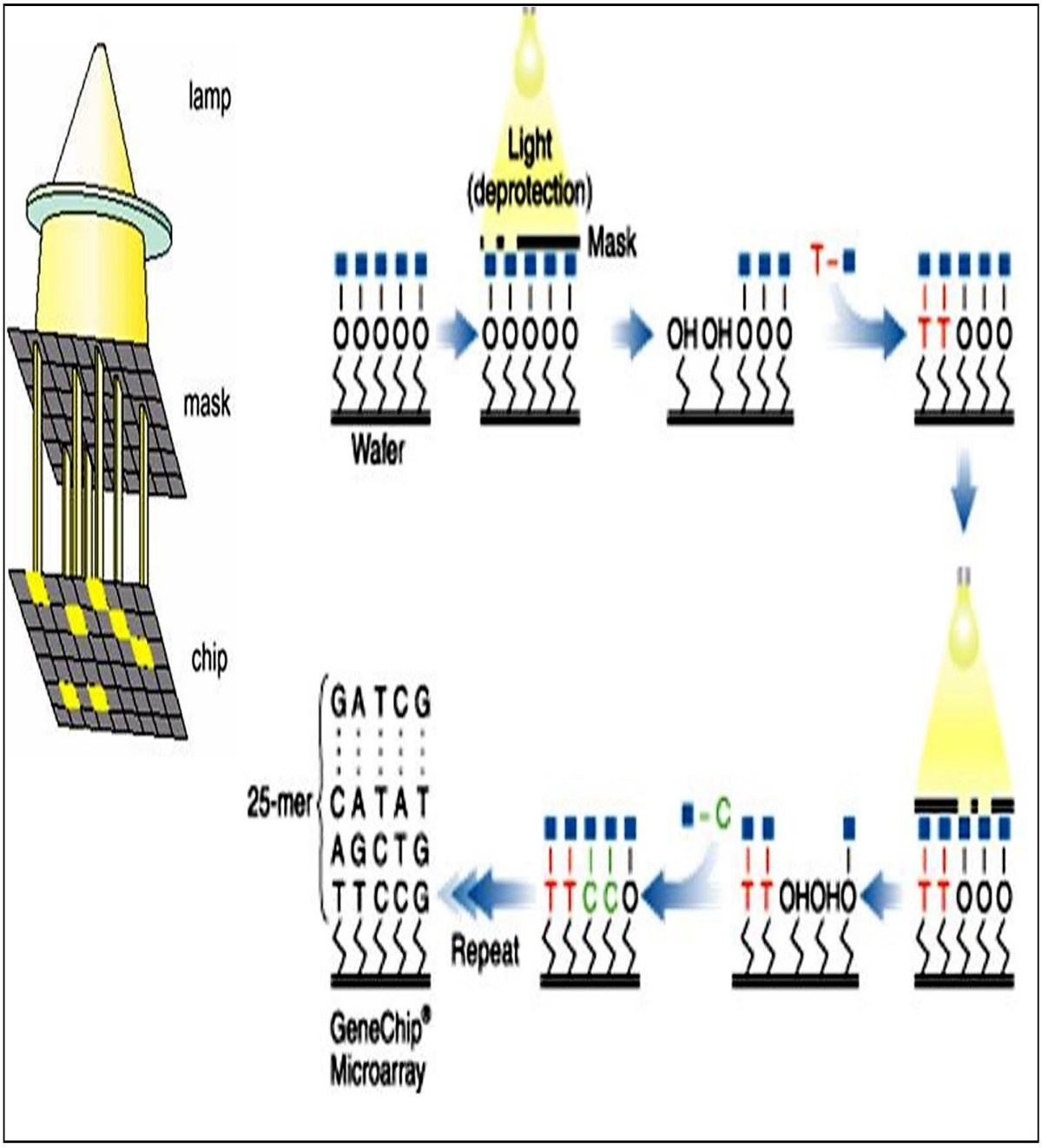


Figure 1.7: Fabrication of In Situ Oligonucleotide microarray.

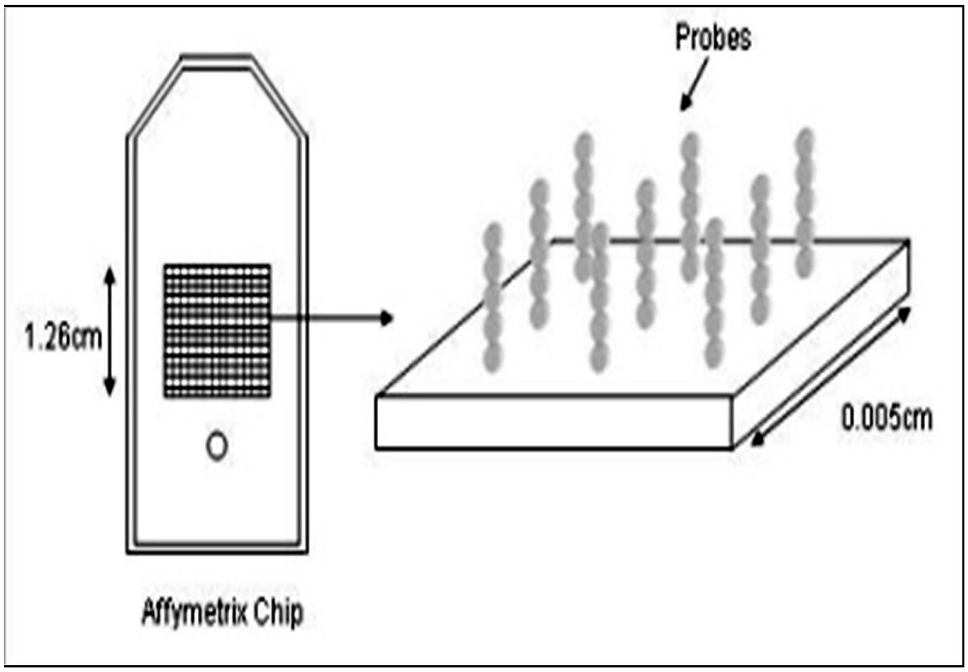


Figure 1.8: In Situ Oligonucleotide microarray

A typical microarray chip is presented in Figure 1.9.

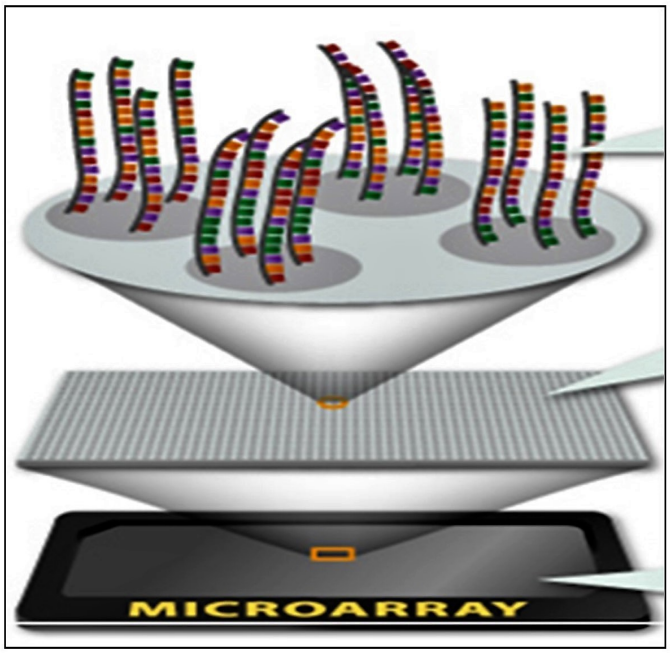


Figure 1.9: Microarray chip

Microarray chip consists of thousands of spots on it, with each spot corresponding to one gene and comprises of multiple copies of the DNA strands of the particular gene.

Some of the commercial microarray chip manufacturers are Affermetrix, Applied Biosystems, Agilent Technologies, Illumina, Arrayjet etc.

### 1.1.2 Experiment

The microarray experiment [2] involves the following steps:

1. The Messenger Ribonucleic Acid (mRNA) are extracted from the cancerous cell and the normal cell.
2. The mRNA are amplified and transformed into cDNA with the help of reverse transcriptase enzyme.
3. The obtained cDNA of cancerous and noncancerous cells are labelled with different color dyes which is normally fluorescent dyes. Usually cancerous sample is labelled with red color while, non-cancerous sample is labelled with green color.
4. These dyes are allowed to hybridize onto microarray chip. In the process of hybridization, the single strands of cDNA from the dyes get attached to its complementary target cDNA.
5. The microarray chip is washed to remove the un-hybridized cDNA.
6. After incubation the microarray chip is scanned with laser (green and red color) at appropriate wavelength.
7. The microarray image in .tiff form is obtained by detection of emitted spectra from microarray image and used for further processing.

The diagrammatic view of a typical microarray experiment is demonstrated in Figure 1.10

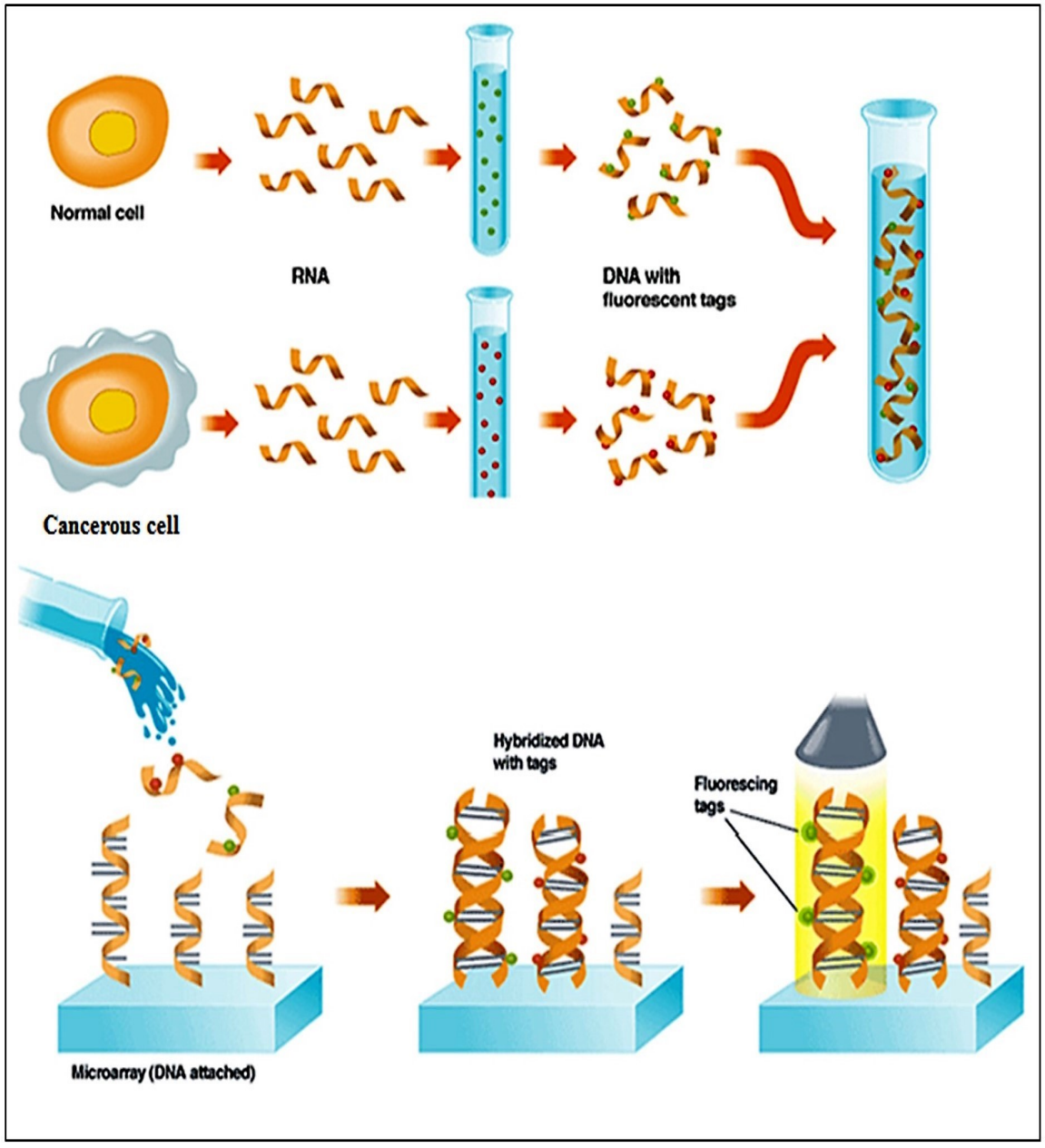


Figure 1.10: Microarray experiment.

Microarray image contains green, red, yellow and grey colored spots [2]. An ideal microarray image generated as a result of microarray experiment is shown in Figure 1.11.

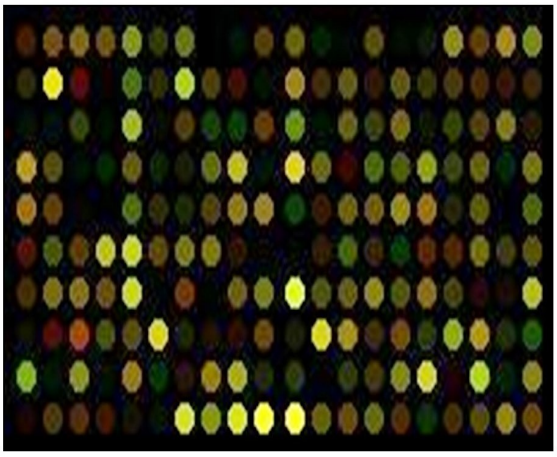

Figure 1.11: Ideal microarray image

The significance of color of the spot (gene) in the generated microarray image is given below:

1. Yellow- The gene is expressed similarly in non-cancerous and cancerous samples.
2. Red- The gene is expressed more (down regulated) in cancerous samples.
3. Green - The gene is expressed more (upregulated) in non-cancerous samples.
4. Grey - The gene is neither expressed in non-cancerous nor in cancerous samples.

### 1.1.3 Image processing

During the course of microarray experiment a number of errors are introduced in the microarray image. As a result, microarray image gets corrupted with variety of noises. These noises appear in the microarray image during process of chip fabrication, treatment of the glass slide, amplification of mRNA, scanning, detection and digitization of microarray chip [10], [11]. The practical microarray image also contains misaligned, irregularly shaped spots. Moreover, the spot intensity is affected by the back ground intensity. The practical microarray image is shown in Figure 1.12.

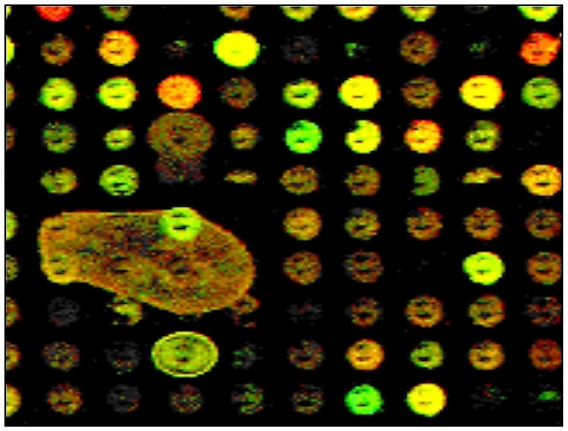

Figure 1.12: Practical microarray image

Hence it is required to process the practical microarray image to obtain insightful results for cancer classification. The microarray image processing comprises of de-noising, gridding, segmentation and quantification [12], [13], [14]. The block diagram for microarray image processing is presented in Figure 1.13.

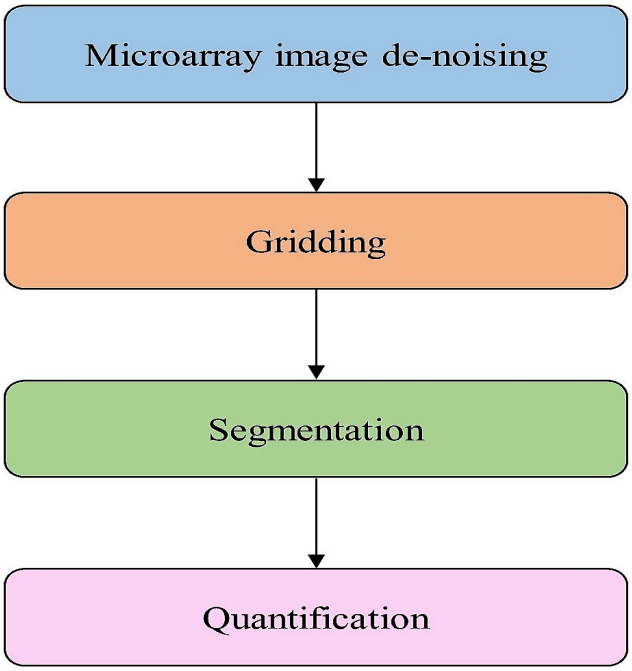


Figure 1.13: Block diagram for microarray image processing.

The proper alignment and detection of the spots on the microarray image is achieved by using gridding and with the help of segmentation the spots are separated from the background. In the process of quantification, either the median of spot intensities or logarithmically transformed ratio of red intensity to green intensity is assigned to the particular spot of corresponding gene.

### 1.1.4 Data analysis

Microarray gene expression data gets generated as a result of microarray image processing. Microarray gene expression data consists of few high dimensional samples. It includes a matrix of the data, in which, each column of data represents one sample and each row represents intensity of the spot corresponding to a specific gene [5], [6]. Figure 1.14 shows microarray gene expression data.

| B | C | D | E | F | G |
|---|---|---|---|---|---|
| IDENTIFIE | Sample1 | Sample2 | Sample3 | Sample4 | Sample5 |
| DDR1 | 29844.3 | 17265.6 | 25947.9 | 29054.5 | 30286.4 |
| RFC2 | 1011.66 | 793.229 | 880.637 | 1000.8 | 687.285 |
| HSPA6 | 689.391 | 584.338 | 862.301 | 842.576 | 798.224 |
| PAX8 | 5198.42 | 4367.41 | 4867.59 | 6974.65 | 8020.01 |
| GUCA1A | 269.881 | 538.245 | 294.926 | 428.498 | 404.549 |
| UBA7 | 2033.8 | 1342.17 | 1036.07 | 1890.96 | 2420.76 |
| THRA | 986.082 | 1087.83 | 979.293 | 1055.92 | 1398.94 |
| PTPN21 | 149.57 | 117.612 | 93.0947 | 80.6684 | 248.498 |
| CCL5 | 44.1338 | 42.1265 | 28.9823 | 71.3078 | 67.2669 |
| CYP2E1 | 527.588 | 421.173 | 494.55 | 1010.7 | 945.524 |
| EPHB3 | 878.971 | 1466.81 | 1226.65 | 861.631 | 1446.3 |
| ESRRA | 1471.37 | 1670.97 | 1690.59 | 3758.83 | 4079.39 |
| CYP2A6 | 1679.75 | 1098.15 | 1028.79 | 1668.09 | 2573.38 |

Figure 1.14: Microarray gene expression data.

The gene expression data can be analyzed as per requirement of the application. Usually, for cancer classification, the high dimension of microarray data is reduced using feature selection and (or) feature extraction methods and subsequently used for classification.

## 1.2 Organization of the thesis

The organization of the remaining part of the thesis is mentioned below,

Chapter 2

Literature review and scope of the present work: This chapter comprises of review of the research papers associated with processing of the gene expression data for cancer classification. It mainly includes review of the research papers pertaining to feature selection, feature extraction and classification of gene expression data for cancer classification. The objectives and overview of implementation of proposed method are explained towards the end of the chapter.

Chapter 3

Image de-noising: This chapter presents the different sources noise introduced in microarray image during the experiment. It includes the details of pixel domain (Median filter) and transform domain (DWT) microarray image de-noising methods. The results of implementation of Median filter and DWT based de-noising for Hard and Soft thresholding (Visushrink, Bayesshrink and Normalshrink) are presented towards the end of chapter.

Chapter 4

Feature selection: In this chapter, a brief overview of filter, wrapper, embedded and hybrid feature selection methods is presented. Further, it includes the feature selection methods namely, Thresholding method, Ratio method and Fusion of Threshold-

ing and ratio method utilized in the proposed work for the implementation of cancer classification based on microarray gene expression data.

Chapter 5

Feature extraction: This chapter includes the details of feature extraction methods viz., Principal Component Analysis (PCA), Discrete Cosine Transform (DCT), Fourier Transform (FT) and Short Time Fourier Transform (STFT). Further, it includes DWT based feature extraction method utilized in proposed work. The motivation for using DWT based feature extraction and the various ways of selection of mother wavelet suitable for a particular application are explained towards the end of chapter.

Chapter 6

Classification Algorithms: This chapter presents the advantages of using Artificial Neural Network (ANN) for classification of non-linear data. It includes details of Error Back Propagation Algorithm (EBPA). Further, it includes details of RPROP algorithm, LM, Conjugate gradient algorithms and SAEN algorithm utilized in the proposed work. The advantages of SAEN algorithm are demonstrated towards the end of chapter.

Chapter 7

Result analysis and conclusion: This chapter demonstrates the results of classification of microarray gene expression based cancer classification for GDS1962 dataset that comprises of Brain tumor samples at different levels of malignancy. These results are obtained with and without using feature selection/extraction methods (DCT and DWT) in combination with classification algorithms namely, RPROP, Conjugate Gradient and LM. It includes the results of classification of Glioma Grade III and Grade IV implemented utilizing feature selection (Thresholding method, Ratio method and

hybrid of Thresholding and Ratio method) and DWT based feature extraction combined with RPROP, LM, Conjugate Gradient and SAEN classification algorithms. Further, the results of testing of optimal gene subset for GDS1962 dataset at every level of malignancy are demonstrated. The conclusion and the scope of the future work are presented towards the end of the chapter.

# Chapter 2

# Literature Review and Scope of the Proposed Work

A detailed review of the research papers related to cancer classification using microarray gene expression data is reported in this chapter.

## 2.1 Feature selection methods

### 2.1.1 Filter methods

J. G. Znang and H. W. Deng [15] applied Bayes Error Filter for feature selection together with K Nearest Neighbor (KNN) and Support Vector Machine (SVM) classifiers. The proposed method is implemented for Colon cancer (62 samples, 2000 genes and 2 classes), Diffused Large B Cell Lymphoma (DLBCL) (77 samples, 6285 genes and 2 classes), Leukemia (38 samples, 3051 genes and 2 classes), Prostate (102 samples, 6033 genes and 2 classes) and Lymphoma (47 samples, 4026 genes and 2 classes) datasets. The classification accuracy of 90.32% with 12 genes for Colon cancer and 92.21% with 6 genes for DLBCL datasets is obtained using KNN classifier. Linear SVM classifier delivers the classification accuracy of 100% for Leukemia dataset with 2 genes, 96.08% for Prostate dataset with 13 genes and 100% for Lymphoma

dataset with 3 genes.

Q. Shen, Z. Mei et al. [16] developed an evolutionary method namely, Simultaneous Sample Particle Swarm Optimization (SSPSO) for simultaneous selection of samples and features combined with SVM classifier for Bipolar Disorder (61 samples, 22283 genes and 2 classes), Glioma (85 samples, 22645 genes and 2 classes) and Sarcoma (54 samples, 22283 genes and 2 classes) datasets. The performance of SSPSO using 400 top ranked genes selected by t-test is compared with Naive SVM without feature selection and PSO-SVM using 400 top ranked genes selected by t-test. SSPSO outperforms Naive SVM and PSO-SVM and attains the classification accuracy of 98% for Bipolar Disorder (18 genes and 34 samples), 96.34% for Glioma (41 genes and 43 samples) and 99.48% for Sarcoma (22 genes and 31 samples) datasets. This method is disadvantageous in terms of usage of limited number of samples for classification.

D. Mishra and B. Sahu [17] implemented classification of Leukemia dataset (72 samples, 7129 genes and 2 classes) using Signal to Noise Ratio (SNR) feature selection method. In the first approach, SNR is applied to the clustered genes and classification is implemented for top 5 genes using SVM and KNN classifiers. In the second approach, SNR is calculated for entire gene set and classification is implemented for top 20 genes using SVM and KNN classifiers. The usage of K-means and SNR with SVM as well as KNN algorithm delivers a classification accuracy of 99% for Leukemia dataset.

B. Chandra and M. Gupta [18] suggested a non-iterative Effective Range based Feature Selection (ERGS) method in combination with Naive Bayes (NB) and SVM classifier for Acute Lymphoblastic Leukemia/ Acute Myeloid Leukemia (ALL/AML) (72 samples, 7192 genes and 2 classes), Colon cancer (62 samples, 2000 genes and 2 classes), Lung cancer (181 samples, 12533 genes and 2 classes), Mixed Lineage Leukemia (MLL) (102 samples, 12600 genes and 2 classes) and DLBCL (96 samples, 4026 genes and 2 classes) datasets. The performance of ERGS is compared with Relief Feature (ReliefF), Minimum Redundancy Maximum Relevance-F Test Distance

Multiplicative (MRMR-FDM), MRMR-F Test Similarity Quotient (MRMR-FSQ), T-Statistics (TS), Information Gain (IG) and Chi Square Statistics method. ERGS together with NB classifier delivers classification accuracy of 98.61% for ALL dataset, 94.79% for DLBCL, 94.79% for Prostate and 94.44% for MLL dataset. ERGS combined with SVM classifier delivers classification accuracy of 82.26% for Colon cancer and 98.34% for Lung cancer dataset. These results are obtained using 10 genes selected by ERGS for every dataset.

A. Sharma, S. Imoto et al. [19] implemented Successive Feature Selection (SFS) method and block reduction for the selection of the genes together with Linear discriminant analysis (LDA), Nearest Centroid, NB and KNN classifiers. This method is implemented for Small Round Blood Cell Tumor (SRBCT) (83 samples, 2308 genes and 4 classes), MLL (72 samples, 12582 genes and 3 classes) and Prostate cancer (102 samples, 12600 genes and 2 classes) datasets. The classification accuracy of 100% is achieved for all three datasets using Nearest Centroid classifier and four genes selected by SFS method.

M. Mandal and A. Mukhopadhyay [20] proposed a method in which, the maximum relevant and minimum redundant gene subset is selected using MRMR technique and its performance is compared with Mutual Information Quotient and Mutual Information Difference methods for Prostate cancer (102 samples, 12533 genes and 2 classes), Childhood ALL (110 samples, 8280 genes and 2 classes), Ovarian cancer (253 samples, 15154 genes and 2 classes) and ALL/AML (72 samples, 7192 genes and 2 classes) datasets. The classification accuracy of 96% for Prostate cancer, 88% Childhood ALL, 99.8% for Ovarian cancer and 100% for ALL/AML datasets is obtained using 100 genes for every dataset.

M. Hajiloo, B. Damavandi et al. [21] implemented the Breast cancer (696 samples) classification using EIGENSTRAT population stratification correction method followed by mean difference feature selection and KNN classifier and classification accuracy of 60.25% is obtained for Breast cancer dataset.

J. C. Rajapakse and P. A. Mundra [22] presented F-score and Kruskal Wallis (KW) - score to select the features and Pareto Font Analysis (PFA) to evaluate the selected gene subsets. The technique is implemented for Global Cancer Map (GCM) (198 samples, 14122 genes and 14 classes), MLL (72 samples, 10930 genes and 3 classes), Ross National Cancer Institute (NCI) (58 samples, 5643 genes and 8 classes), Staunton NCI (58 samples, 3144 genes and 8 classes), Lung cancer (203 samples, 5345 genes and 5 classes) and 11 Tumor (174 samples, 9700 genes and 11 classes) datasets. The performance of F-Score, F-PFA, KW Score, KW-PFA and Gene Dominant and Gene Dormant Indices (GDI) is compared for the cancer datasets. A classification accuracy of 70.80% for Ross NCI (386 genes) dataset is obtained using F-PFA. KW-PFA delivers classification accuracy of 76.65 % for GCM (327 genes) dataset. GDI delivers 99.65% for MLL (30 genes) dataset. KW-Score delivers 93.73% (397 genes) for 11 Tumor and 95.59% (189 genes) for Lung cancer dataset.

Heba Abusamra [23] presented eight feature selection method such as IG, Twoing rule (TR), Sum Minority (SM), Max Minority (MM), Gini Index (GI), Sum Variances (SV), TS and one dimensional SVM along with KNN, SVM and Random Forest (RF) classifiers for Glioma datasets (85 samples, 22283 genes and 2 classes) generated by Freije and Phillip. Without feature selection SVM performed better giving 91.89% and 86.73 % classification accuracy for Phillip and Freije datasets, respectively. With 20 genes selected using SM, GI, IG and TS gene selection methods, SVM classifier performed best with 91.89% and 88.77% , respectively, for Phillip and Freije datasets. As a result of usage of 20 most frequently appearing genes from all eight feature selection methods, classification accuracy of 94.59% and 84.69 % is obtained for Phillip (linear SVM, radial SVM, sigmoidal SVM and KNN) and Freije datasets (sigmoidal SVM), respectively. As common genes are not found in shortlisted gene subsets of both the datasets, the performance is evaluated using both the subsets of genes and classification accuracy of 94.59% and 90.81% is obtained for Phillip and Freije datasets, respectively using linear SVM, radial SVM and sigmoidal SVM .

S. Tarek, R.A. Elwahab et al. [24] implemented ensemble classification for Colon cancer (62 samples, 2000 genes and 14 classes), Leukemia (72 samples, 3751 genes and 3 classes) and Breast cancer (78 samples, 22481 genes and 8 classes) datasets. The feature selection is implemented by using Extreme Value Distribution (EVD) based gene selection, Backward Elimination Hillbert Schmidt Criterion (BAHSIC) and Singular Value Decomposition Entropy (SVDE) gene selection while classification is implemented using neural network algorithms. A classification accuracy of 72.42% for Colon cancer (5 genes), 72.64% for Leukemia (5 genes) and 62.24% for Breast cancer (5 genes) datasets is obtained using BHASIC. A classification accuracy of 89.68% for Colon cancer (49 genes), 98.61% for Leukemia (224 genes) and 100% for Breast cancer (5727 genes) datasets is obtained using EVD. SVDE based ensemble classifier delivers classification accuracy of 90.47% for Colon cancer (240 genes), 97.36% for Leukemia (187 genes) and 98.28% for Breast cancer (1236 genes) datasets.

W. Zhong, X. Lu et al. [25] proposed a method to perform feature selection using Bhattacharyya distance and classification using SVM classifier for Colon cancer (62 samples, 2000 genes and 2 classes) and Leukemia (72 samples, 3571 genes and 2 classes) datasets. For Colon cancer classification accuracy of 90.5% is obtained using Bhattacharyya distance combined with SVM classifier (7 genes). For Leukemia classification accuracy of 97.42% is obtained using SVM- Recursive Feature Elimination (SVM-RFE) (10 genes).

### 2.1.2 Wrapper methods

L. Yu and M. E. Berens [26] proposed gene selection using Sample Weighing (SW) for Colon cancer (62 samples, 2000 genes and 2 classes), Leukemia (72 samples, 7129 genes and 2 classes), Lung cancer (181 samples, 12533 genes and 2 classes) and Prostate cancer (102 samples, 6034 genes and 2 classes) datasets. The performance of SVM-RFE, SW-SVM, Ensemble-SVM, ReliefF, Ensemble ReliefF and SW ReliefF algorithms is compared. SVM-RFE algorithm delivers 80.3% (10 genes) and

97.2% (100 genes) classification accuracy for Colon cancer and Leukemia datasets, respectively, while Ensemble ReliefF and SW SVM-RFE algorithm delivers 93.4% (10 genes) classification accuracy for Prostate cancer. SW-ReliefF delivers 98.8% (200 genes) classification accuracy for Lung cancer dataset.

Q. Liu, Z. Zhao et al. [27] suggested the Fuzzy logic based feature methods such as Feature Selection based on Clustering (FS-Cluster), Feature selection based on Sample Selection (FS-SSM) along with SVM and KNN classifier for Multiple Myeloma (105 samples, 7129 genes and 2 classes), Acute Leukemia (72 samples, 7129 genes and 2 classes), Colon cancer (62 samples, 2000 genes and 2 classes), DLBCL (77 samples, 7129 genes and 2 classes) and Prostate cancer (102 samples, 12600 genes and 2 classes) datasets. For Myeloma dataset the combination of all suggested feature selection methods together with SVM gives 100% classification accuracy. For Leukemia and DLBCL datasets the FS-SSM combined with SVM algorithm delivers 94.6% and 86.1% classification accuracy, respectively. For Colon and Prostate cancer datasets FS-SSM together with KNN algorithm delivers 84.5% and 93.3% classification accuracy, respectively.

### 2.1.3 Embedded methods

S. Niijima and Y. Okuno [28] proposed feature selection based on Laplacian LDA (LLDA) for Leukemia (38 samples, 7129 genes and 2 classes), Colon cancer (62 samples, 2000 genes and 2 classes), Medulloblastoma (60 samples, 7129 genes and 2 classes), Breast cancer (76 samples, 4918 genes and 2 classes) , Lung cancer (86 samples, 7129 genes and 2 classes), MLL(57 samples,12582 genes and 3 classes) and SR-BCT (63 samples, 2308 genes and 4 classes) datasets. The performance of LLDA-RFE is compared with Laplacian Score, SVDE and Fisher Score (FS). LLDA-RFE delivers 99.4% (50 genes), 65.9% (20 genes), 67.2% (100 genes) and 96.2% (100 genes) classification accuracy for Leukemia, Medulloblastoma, Breast cancer and MLL datasets, respectively. FS delivers 64.9% (100 genes) and 97.4% (100 genes) classification ac-

curacy for Lung cancer and SRBCT datasets, respectively. SVDE delivers 88.5% (50 genes) classification accuracy for Colon cancer dataset.

S. Maldonado, R. Weber et al. [29] implemented Kernal Penalised SVM (KP-SVM) for the classification of Diabetes (768 samples, 8 genes and 2 classes), Wisconsin Breast cancer (569 samples, 30 genes and 2 classes), Colorectal (62 samples, 2000 genes and 2 classes) and Lymphoma (96 samples, 4026 genes and 2 classes) datasets. For these datasets the performance of KP-SVM is compared with SVM and SVM-RFE in combination with Fisher Score and Concave feature selection methods. KP-SVM is proved to be beneficial giving 76.74% (5 genes), 97.55% (15 genes), 96.57% (20 genes) and 99.73% (8 genes) for Diabetes, Wisconsion Breast, Colorectal and Lymphoma datasets, respectively.

A. Anaissi, M. Goyal et al. [30] introduced Balanced Iterative RF (BIRF) algorithm for Childhood Leukemia (110 samples, 22678 genes and 3 classes), NCI (61 samples, 5244 genes and 82 classes), Colon cancer (72 samples, 2000 genes and 2 classes) and Lung cancer (181 samples, 12533 genes and 2 classes) datasets. The performance of BIRF is compared with various methods such as SVM-RFE, RF, NB classifiers. BIRF is proved to be advantageous in terms of providing classification accuracy of 96% (19 genes), 97% (57 genes), 99.83% (100 genes) and 99.83 % (112 genes) for Colon cancer, Lung cancer, Leukemia and NCI datasets, respectively.

### 2.1.4 Hybrid methods

Y. Leung and Y. Hung [31] proposed Multiple Filter and Multiple Wrapper method (MFMW) for cancer classification. The performance of various feature selection methods along with various algorithms is compared. This method is implemented for Leukemia (38 samples, 7129 genes and 2 classes), Breast cancer (49 samples, 6817 genes and 2 classes), Colon cancer (62 samples, 6500 genes and 2 classes), Lymphoma (77 samples, 6817 genes and 2 classes), Prostate cancer (102 samples, 12600 genes and 2 classes) and Lung cancer (181 samples, 12600 genes and 2 classes) datasets. With

MFMW model, the maximum classification accuracy obtained is 100%for Leukemia datasets with IG, C4.5 classifier and 4 genes, 100% for Breast cancer with Mutual Information and KNN classifier, 100% for Lymphoma with Multi object Evolutionary algorithm, Weighing Votes and 6 genes, 98.04% for Prostate cancer with TS, SVM and 6 genes and 98.34% for Lung cancer dataset with Redundancy based Filter, C4.5 Classifier and 6 genes.

C. Lee, Y. Leu [32] suggested combination of Genetic Algorithm with Dynamic Parameter Setting for feature selection, Chi Square based homogeneity test in combination with SVM classifier. The proposed method is implemented for Colon cancer (62 samples, 2000 genes and 2 classes), SRBCT (83 samples, 2308 genes and 4 classes), Breast cancer (22 samples, 3226 genes and 2 classes), ALL/AML (72 samples, 7129 genes and 2 classes), DLBCL (47 samples, 4026 genes and 2 classes) and GCM (198 samples, 16306 genes and 14 classes) datasets. With this method classification accuracy of 100% (8 genes), 100% (8 genes), 100% (5 genes), 100% (6 genes) and 87.04% (26 genes) for Colon cancer, SRBCT, ALL/AML , DLBCL and GCM datasets, respectively, are obtained.

M. Hajiloo, H. R. Rabiee et al. [33] implemented Fuzzy SVM (FSVM) for classification of Leukemia (72 samples, 6817 genes and 2 classes), Prostate cancer (102 samples, 12600 genes and 2 classes) and Colon cancer (62 samples, 2000 genes and 2 classes) datasets. The performance of FSVM without feature selection, FSVM with SVM-RFE, FSVM with SNR and FSVM with SVM-REF is compared. FSVM with SVM-REF delivers the best performance with classification accuracy of 98.57% (10 genes) and 96.77% (50 genes) for Leukemia and Colon cancer datasets, respectively. FSVM with SNR delivers 95.18 % (5 genes) classification accuracy for Prostate cancer dataset.

### 2.1.5 Ensemble methods

P. Yang, B. B. Zhou et al. [34] developed the Genetic Ensemble System with Multi filter (MF-GE) for the classification of Leukemia (72 samples, 7129 genes and 2 classes),

Colon cancer (62 samples, 2000 genes and 2 classes), Liver cancer (157 samples, 20983 genes and 2 classes) and MLL (72 samples, 12582 genes and 3 classes) datasets. The performance of various feature selection methods like Gain Ratio (GR), Genetic Algorithm (GA)/KNN, GE, MF-GE combined with the classifiers such as C4.5, RF, 3-NN, 7 NN and NB is compared. Combination of MF-GE and NB algorithm results into classification accuracy of 96.27% and 91.50% for Leukemia and MLL datasets while, the classification accuracy of 77.01% is obtained for Colon cancer dataset using combination of MF-MG and 3-NN algorithm. MF-GE together with Majority Voting delivers 93.80% classification accuracy for Liver cancer dataset.

H. Liu, L. Liu et al. [35] suggested Ensemble Gene Selection by Grouping (EGSG) method for Breast cancer (97 samples, 24481 genes and 2 classes), CNS cancer (7129 samples, 60 genes and 2 classes), Colon cancer (62 samples, 6000 genes and 2 classes), Leukemia (72 samples, 7129 genes and 2 classes) and Prostate cancer (102 samples, 12600 genes and 2 classes) datasets. The classification accuracy obtained by EGSG is compared with classification accuracy obtained by MRMR, Fast Correlation Based Filter and Ensemble Classification with Random Partitioning in combination with 3-NN and NB classifiers. NB classifier outperforms with classification accuracy of 100% , 100% , 93.55% , 100% and 98.02% for Breast, CNS, Colon, Leukemia and Prostate cancer datasets with 30 genes.

## 2.2 Feature extraction methods

S. Li, C. Liao et al. [36] proposed the combination of DWT and SVM for classification of Colon (72 samples, 7129 genes and 2 classes) and ALL/AML (62 samples, 2000 genes and 2 classes). The approximation as well as thresholded detailed coefficients are used for classification. Thresholding is implemented using Maximum Modulus method and accuracy obtained by wavelets such as Daubecies (Db1 and Db8), Coiflets (Coif1 and Coif3), Symlet (Sym2 and Sym15) and Bio-orthogonal (Bior1.1 and Bior2.6) is compared. For ALL/AML dataset 100% classification is

achieved using Bior2.6 (level 3) and 100 genes while, for Colon cancer dataset 93.55 % classification accuracy is obtained using Db8 (level 4) and 250 genes.

R. Mahapatra, B. Majhi et al. [37] performed feature extraction using DCT and PCA while classification using Functional Link Neural Network (FLNN) classifiers for Lung cancer ( 197 samples, 581 genes and 2 classes) and Breast cancer datasets ( 981 samples,1213 genes and 2 classes). The FLNN classifier with Chebyshev Expansion delivers better accuracy for both the datasets. Using PCA, the classification accuracy of 83 % (24 genes) is obtained for Breast cancer and Usage of DCT gives the classification accuracy of 90% (187 genes) for Lung cancer dataset.

M. Vimaladevi and B. Kalaavathi [38] implemented classification of SRBCT (83 samples, 2308 genes and 4 classes), Leukemia (124 samples, 7129 genes and 2 classes) and Lymphoma (98 samples and 4 classes) datasets using hybrid combination of GA and EBPA algorithm. The performance of this method is compared with Error Back Propagation Algorithm (EBPA). The hybrid combination of GA and EBPA gives better performance as compared to EBPA, with classification accuracy of 85.65% (2 genes) for Lymphoma, 89.33% (3 genes) for Leukemia and 91.7% (4 genes) for SRBCT datasets.

## 2.3 Fusion of feature selection and extraction methods

Z. Zainuddin and O. Pauline [39] made use of Improved Wavelet Neural Network (WNN) together with conditional TS for classification of SRBCT (63 samples, 2308 genes, 4 classes), Leukemia (72 samples, 7129 genes and 2 classes), Glioma (50 samples, 12625 genes and 2 classes) and CNS (40 samples, 7129 genes and 2 classes) datasets. The comparison of accuracy obtained using different wavelets such as Mexican hat, Gaussian wavelet, Morlet and number of classifiers is presented. The usage of WNN results into the classification accuracy of 98.61% ( Gaussian wavelet) for Leukemia dataset, 95% (Gaussian and Mexican hat) for CNS dataset, 100% (Gaussian, Morlet and Mexican hat) for SRBCT dataset and 92% (Gaussian wavelet) for

Glioma dataset.

S. Rashid and G. M. Maruf [40] implemented the classification for Ovarian cancer (225 samples, 15154 genes and 2 classes), ALL/AML (128 samples, 12625 genes and 2 classes) and Pancreatic cancer (181 samples, 6771 genes and 2 classes) datasets using combination of t-test, Db1 wavelet and SVM classifier. The classification accuracy of 99.92% (327 genes), 78.26% (59 genes) and 99.84 % (302 genes) is obtained for Ovarian cancer, Pancreatic cancer and ALL/AML datasets, respectively.

J. Bennet, C.A. Ganaprakasam et al. [41] extracted the features of the genes chosen by Moving Window followed by TS for wavelet coefficient selection and hybrid of KNN, NB and SVM classifier. The proposed method is implemented for Breast cancer (97 samples, 22481 genes and 2 classes), Colon cancer (62 samples, 1909 genes and 2 classes), Ovarian cancer (253 samples, 15154 genes and 2 classes), CNS cancer (60 samples, 7129 genes and 2 classes) and Leukemia (72 samples, 7129 genes and 2 classes) datasets. Various wavelets used for analysis are Db7, Sym2, Bior2.2 and Reverse Bio-orthogonal (Rbio2.2). The classification accuracy achieved with this method is 100% (512 window size, Db7 wavelet at level 4 and 24 genes) for Breast cancer, 100% (512 window size, Db7 wavelet at level 1 and 4 genes) for Colon cancer, 100% (64 window size, Sym2, Bior2.2, Rbio2.2 wavelets at level 2 and 237 genes) for Ovarian cancer, 100% (64 window size, Rbio2.2 wavelet at level 1 and 14 genes) for CNS cancer and 100% (64 window size, Rbio2.2 wavelet at level 3 and 14 genes) for Leukemia dataset.

Many a times the cancer marker genes are used for screening of the cancer but there are evidences of failure of this method [42].

## 2.4 Scope of the proposed work

Researchers implemented plenteous ways for diminishing the size of the microarray data, however, there are many open prospects for further improvement in terms of achieving 100% classification accuracy for less number of genes [43]. Therefore the

proposed work is implemented with the following research objectives.

### 2.4.1 Research objectives

1. To improve the accuracy and speed of cancer classification by using various feature selection, feature extraction techniques and neural network classifiers.
2. To compare performance of various feature selection techniques and arrive at an efficient method for cancer classification.
3. To compare performance of various classification algorithms namely, RPROP, Conjugate Gradient, LM, SAEN and suggest the most suitable algorithm for cancer classification.

### 2.4.2 Proposed system

To accomplish the research objectives, the proposed system is designed to obtain 100% classification accuracy with optimum number of genes. The Wrapper, Embedded, Hybrid and Ensemble methods are complex and also computationally expensive as compared to filter methods. Therefore, in the proposed system the selection of features is implemented using computationally efficient filter methods such as Thresholding method and Ratio method individually as well as conjointly (hybrid method). The implementation of the feature selection methods is followed by the transformation of the signal using DWT. The performance of Thresholding method, Ratio method and fusion of Thresholding and Ratio method along with DWT is compared for various classification algorithms such as RPROP, LM, Conjugate Gradient and SAEN. The proposed system is implemented using MATLAB R2017a software.

#### Databases

In the proposed work, classification is implemented for Glioma Grade III and Grade IV datasets- GDS1975, GDS1976, GDS1815 and GDS1816 from Gene Expression Omnibus Database [44], [4], [45]. Originally, Phillip and Freije obtained the datasets

using Affymetrix Human Genome U133 Array. GDS1975, GDS1976 datasets contain 26 samples of Glioma Grade III and 59 samples of Glioma Grade IV while GDS1815, GDS1816 contain 24 samples of Glioma Grade III and 76 samples of Glioma Grade IV. Every dataset contains 22283 features. From every dataset 70% samples are used for training and 15% are used for testing and validation each.

**System flow chart**

Before using the datasets the duplicate features are eliminated by averaging and the data is normalized to have zero mean and unit standard deviation. Figure 2.1 shows the system flow chart for individual Glioma dataset.

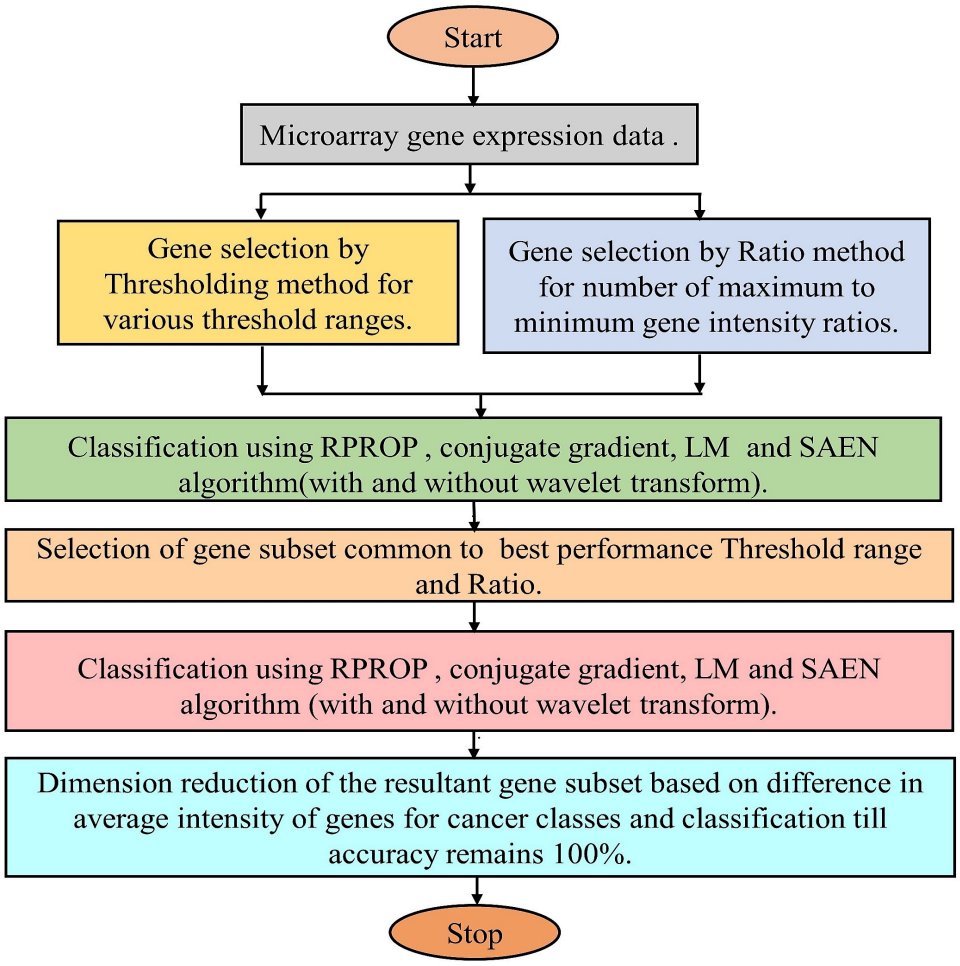


Figure 2.1: System flow chart.

Finally, to derive the identical gene subset across Glioma datasets, considering optimal gene subsets, the genes common to all Glioma datasets are mined. The classification of Grade III and Grade IV Glioma is implemented using obtained common gene subset. The performance of this common gene subset is tested for Brain tumor dataset GDS1962 at various malignancy levels.

# Chapter 3

# Image De-Noising

Practical microarray image is degraded due to various noises viz., sample preparation noise, scanning noise and hybridization noise. De-noising of the microarray image is one of the major steps to attain accurate gene expression data. Various types of noises present in the practical microarray image are described in the following section.

## 3.1 Types of the noises in microarray image

1. Sample preparation noise

   The sample preparation noise appears in the microarray image due to amplification of mRNA in the process of conversion of mRNA to cDNA and chemical processes employed during the course of sample preparation.

2. Scanning noise

   Noise during scanning comes into picture as a result of photon noise, electron noise in the equipment involved in the scanning. Further, it may also appear due to laser light leakage, dust particles, laser light reflection, quantization in the digitization process etc.

3. Hybridization noise

   Hybridization noise in the microarray image is due to variation in binding of the

target molecules and cross hybridization.

Usually, the hybridization noise is more significant as compared to sample preparation and scanning noise [11],[46],[47], [48].

The diagrammatic representation of the noises introduced in the microarray image during the microarray experiment is presented in Figure 3.1.

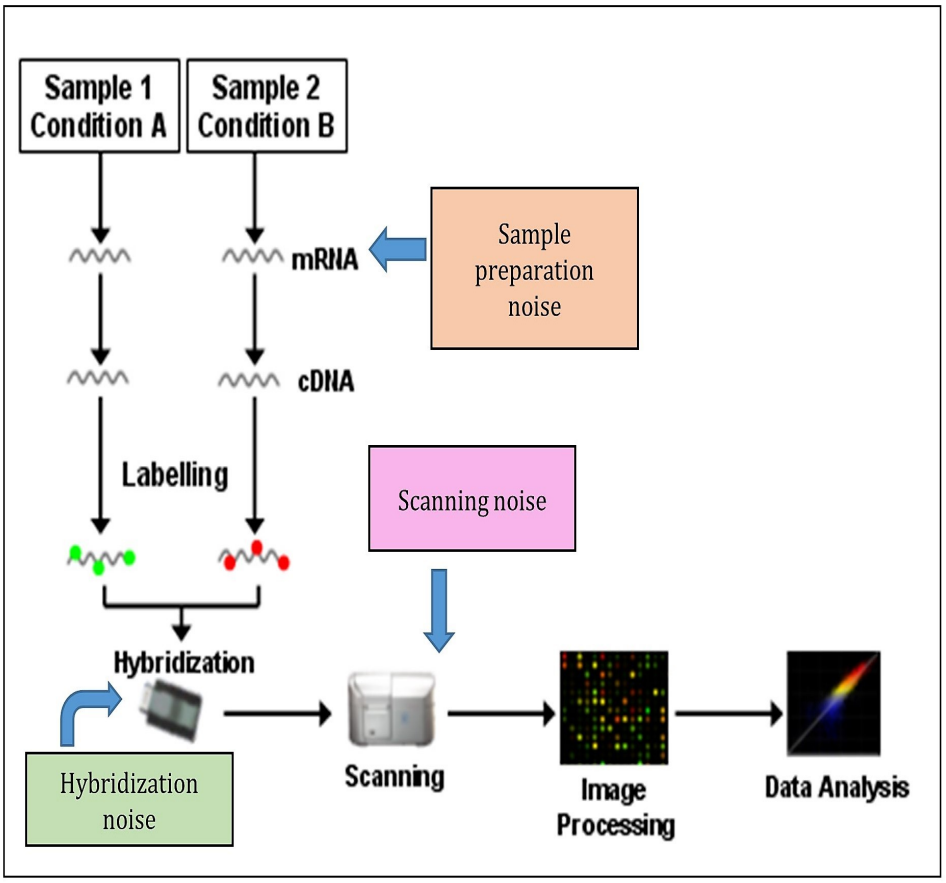


Figure 3.1: Types of noises introduced in the microarray image during microarray experiment.

## 3.2 Methods of noise reduction

Microarray image de-noising is usually accomplished by,

1. Precise adjustment of the fluorescent machine, fabrication machine and stan-

dardization of the experimental conditions.

2. Design of an appropriate filter.

Improvement in the experimental conditions diminishes the noise in the microarray image to some extent. Therefore, the most effective way to reduce the noise is by designing appropriate filter such as pixel/spatial domain filter [49], [50], [51] or transform domain filters [36], [52], [53].

The block diagram for microarray image de-noising is shown in the Figure 3.2

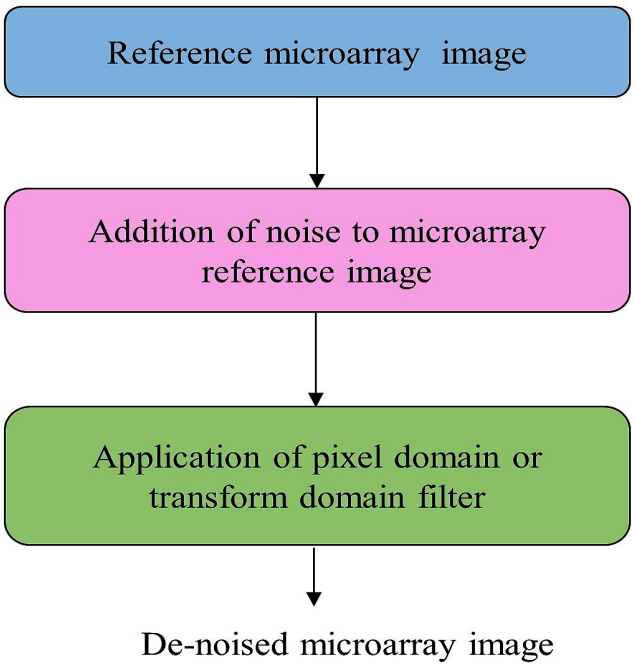


Figure 3.2: Block diagram for microarray image de-noising

The steps in the microarray image de-noising are described below:

### 3.2.1 Generation of microarray reference image

To measure the effectiveness of the designed filter, it is essential to generate the reference microarray image with the help of microarray image simulators available on the internet.

### 3.2.2 Addition of the noise

The hybridization noise at higher intensity of the spots and sample preparation noise in the image appears to be closer to a Poisson noise while the hybridization noise that appears at low intensity is complex in nature [1]. According to Central Limit Theorem Gaussian noise is more appropriate model for representation of overall noise in the image. Therefore Gaussian noise (with zero mean and known variance) and Poisson′s noise are introduced in the microarray image [54].

### 3.2.3 Application of pixel domain or transform domain filter

Pixel and transform domain filters are most widely used for microarray image de-noising explained in the following sections.

#### Pixel domain filter

To de-noise the microarray image using pixel domain filter, the filter template is made to move over the image from point to point. The central pixel of the mask is altered in number of ways for different filters.

Median filter is a non-linear pixel domain filter. In this case, the pixels in the filter template of an image portion are arranged based on some ranking method [54]. Median filter performs better for salt and pepper noise [54]. The steps in implementation of microarray image de-noising using Median filter are described below:

1. The Median filter mask is placed at the first location of the microarray image.
2. The pixels of an image of corresponding mask are arranged in an increasing order.
3. The median of these pixel values is computed.
4. The central pixel is replaced with the median of the pixel values in the image part.

5. The mask moved from pixel to pixel.

6. The above mentioned procedure is repeated for each pixel in the image.

**Transform domain filter**

Wavelet transform is one of the most extensively recognized transform domain filter for image de-noising. DWT is an indispensable tool for image de-noising due to its ability to provide multiresolution analysis, ease of selection of necessary wavelet basis function and higher energy compaction [55], [56]. The block diagram of DWT based image de-noising is demonstrated in the Figure 3.3

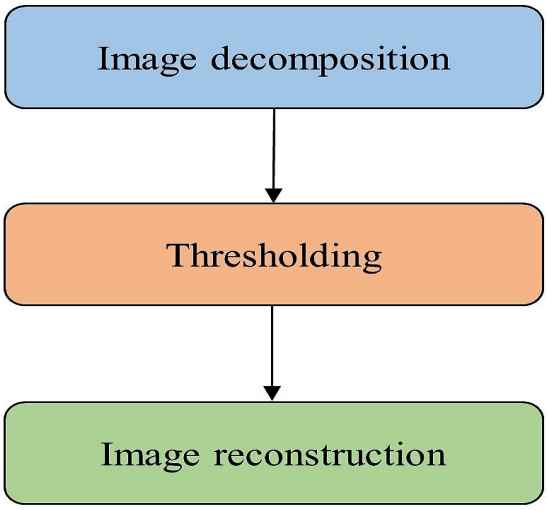


Figure 3.3: Block diagram of DWT based image de-noising

**Image decomposition** Image decomposition, deconstructs the noisy image to level N using DWT. It is accomplished by application of DWT using two separate single dimensional transforms [56]. The major steps in the image decomposition using DWT are listed below:

1. Initially, low pass filter (LPF) and high pass filter (HPF) are applied to an

image along the x-dimension of an image. To get rid of the redundant wavelet coefficients, the resultant coefficients are down sampled by a factor of two (DS2). The coefficients of low pass filtered image are stored on the left half section of an image matrix while the coefficients of the high pass filtered image are stored on the right half section of an image matrix.

2. Subsequently, LPF and HPF is applied to an image obtained in step1 along the y-dimension of an image. The resultant coefficients are down sampled by a factor of two. Finally, image is divided into four bands LL1, HL1, LH1 and HH1 at the end of first level of decomposition.

Figure 3.4 and Figure 3.5 demonstrates the process of microarray image decomposition at level 1.

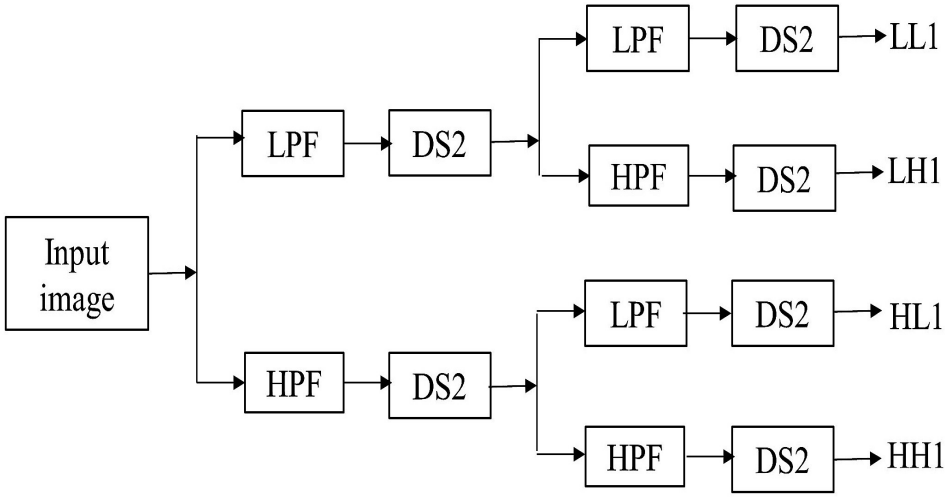


Figure 3.4: Process of applying dwt to an image

In the image,

LL1- signifies wavelet approximation coefficients at decomposition level 1

LH1- signifies wavelet detailed vertical coefficients at decomposition level 1

HL1- signifies wavelet detailed horizontal coefficients at decomposition level 1

HH1 signifies wavelet detailed diagonal coefficients at decomposition level 1.

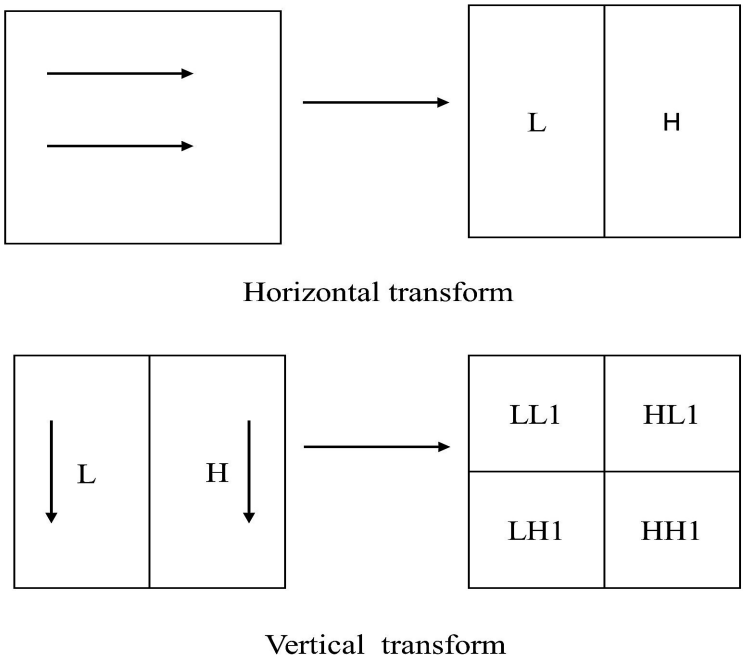


Figure 3.5: Application of wavelet transform to an image

Usually, the wavelet approximation (LPF) coefficients constitute the information signal while the detailed (HPF) coefficients constitute noise and edges in an image [55], [56], [57].

**Thresholding** Thresholding of the detailed wavelet coefficients helps to separate the edges in the microarray image from the noise so as to diminish the noise in the microarray image while retaining the edges. The different methods of thresholding namely, Hard thresholding and Soft thresholding are described below.

In the case of Hard thresholding, the wavelet coefficients smaller than threshold (thd) are made zero while, the coefficients larger than threshold are retained using Equation 3.1,

$$M = \begin{cases} wd, & \text{if } |wd| \geq thd \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

Where,

$M$ = thresholded wavelet detailed coefficients,

$wd$ = wavelet detailed coefficients before thresholding.

In the case of Soft thresholding, the wavelet coefficients less than threshold are made zero and coefficients larger than threshold are altered using Equation 3.2

$$M = \text{sgn}(wd)\max(0, (|wd| - thd)) \tag{3.2}$$

The digramatic representation of Hard and Soft thresholding is demonstrated in Figure 3.6

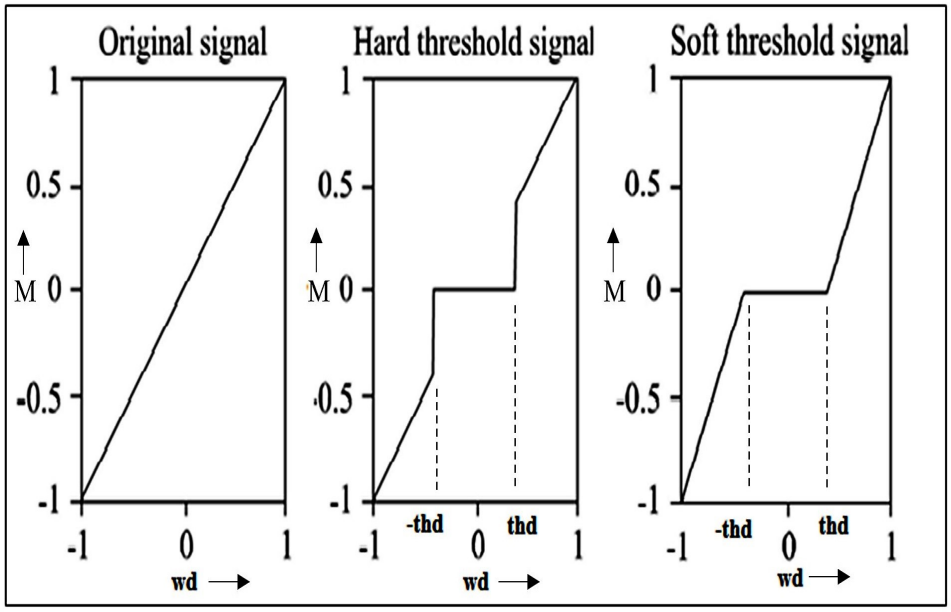


Figure 3.6: Hard and Soft thresholding

Some of the measures of the threshold value for Hard or Soft thresholding are given below, [55], [56], [57]:

1. Visushrink: For Visushrink the 'thd' is defined using Equation 3.3,

$$thd = \sigma\sqrt{2log(me)} \tag{3.3}$$

where,

$\sigma$ = standard deviation of noise in an image

$= median\ (abs(\text{detailed coefficients}))/0.6745$

$m \times e =$ number of image pixels.

Visushrink decides the threshold based on the number of image pixels and leads to smooth image. However, it does not consider the intensity of image pixels while choosing the threshold.

2. Bayesshrink: For Bayesshrink 'thd' is defined using Equation 3.4,

$$thd = \sigma^2/({\sigma_x}^2) \tag{3.4}$$

where,

${\sigma_x}^2 =$ reference image intensity variance $= max((\sigma_y^2 - \sigma^2), 0)$

${\sigma_y}^2 =$ noisy image intensity variance $= (1/2) \sum(\text{horizontal coefficients})^2$

For Small value of $\sigma/\sigma_x$ (signal being stronger than the noise signal) it is required to select small value of $thd/\sigma$ and vice-versa.

3. Normalshrink: It is a level based thresholding method. For Normalshrink 'thd' is defined using Equation 3.5,

$$thd = \beta(\sigma^2/\sigma_y) \tag{3.5}$$

where,

$\beta =$ scale parameter for thresholding $= (\log(\mathrm{j}_k/l))^{0.5}$

$\mathrm{j}_k =$ number of wavelet coefficients

$l =$ wavelet decomposition level.

**Image reconstruction** The de-noised microarray image is reconstructed by up sampling (UP2) the approximate and thresholded detailed wavelet coefficients followed by filtering using reconstruction filters ($\mathrm{LFP}^{-1}$ and $\mathrm{HPF}^{-1}$ ) as shown in Figure 3.7.

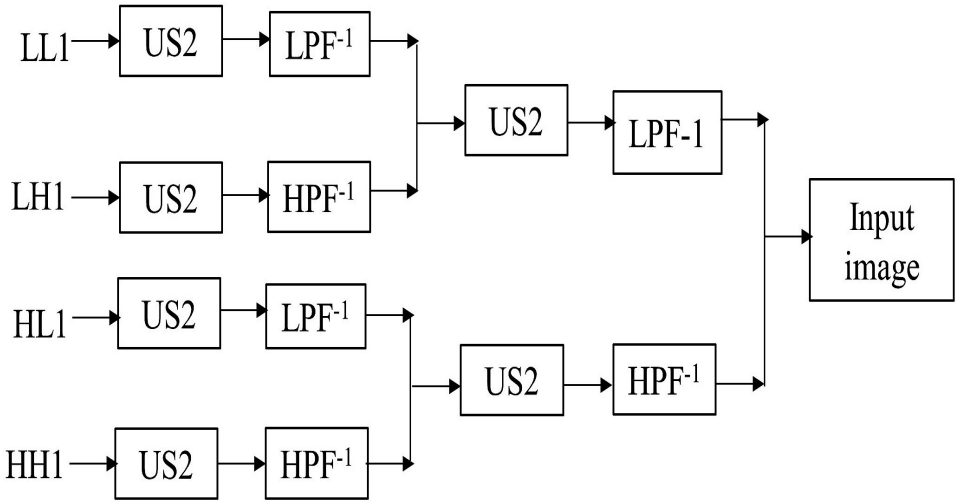


Figure 3.7: Image reconstruction

## 3.3 Quality assessment of de-noised image

The quality of de-noised image is assessed with the parameters like Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR).

1. Mean Square Error: It denotes the error between reference microarray image and noisy image. It is defined using Equation 3.6,

$$MSE = \sum_{j=1}^{e} \sum_{i=1}^{m} (I(i,j) - K(i,j)) \tag{3.6}$$

where,
$K(i,j)$ = De-noised microarray image of size (m, e)
$I(i,j)$ = Reference microarray image of size (m, e).

Lower the value of MSE, better will be the quality of de-noised image. Ideally MSE is expected to be zero.

2. Peak Signal to Noise Ratio: The PSNR of de-noised microarray image is given

using Equation 3.7,

$$PSNR = 10\log_{10}\left(\frac{(max(I))^2}{MSE}\right) \tag{3.7}$$

where,

$\max(I)$ = maximum intensity in the image matrix = 65,535 (16 bit image).

PSNR is usually expressed in decibels. Higher the value of PSNR, better will be the quality of de-noised image.

## 3.4 Results of microarray image de-noising

Microarray image generated using Microarray Scan Simulator [58] is presented in the Figure 3.8

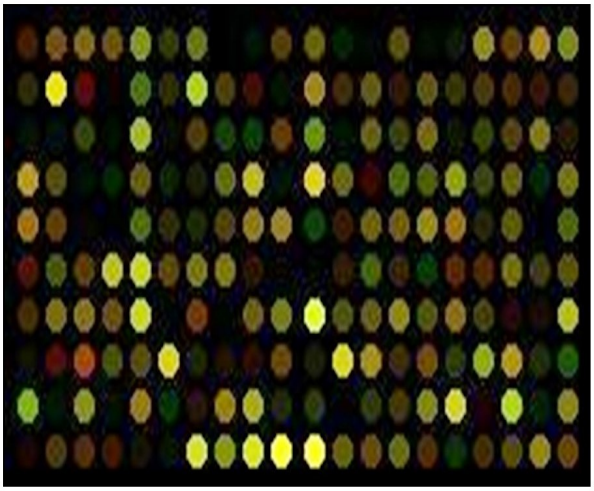

Figure 3.8: Simulated microarray image

The microarray image de-noising is implemented using Image Processing and Wavelet Toolbox of Matlab2017a. Image de-noising is applied individually for the red and green plane of microarray image and upon completion of de-noising process, the image is reformed using de-noised red and green plane of an image.

Figure 3.9 demonstrates the red plane and green plane of the reference image before and after addition of noise.

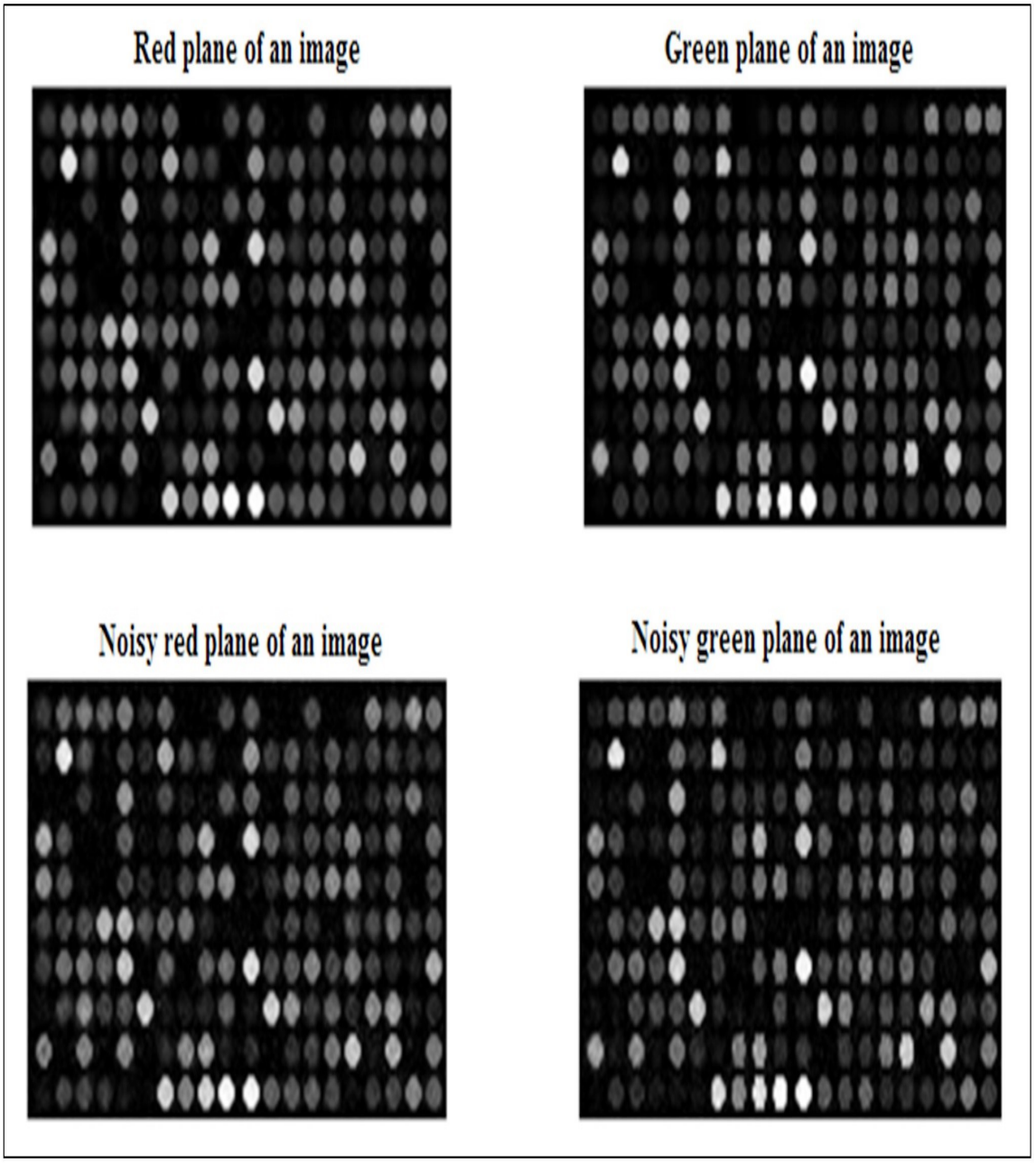


Figure 3.9: Reference and noisy microarray image

The microarray image obtained as a result de-noising using Median filter, wavelet Hard thresholding and Soft thresholding is presented in Figure 3.10.

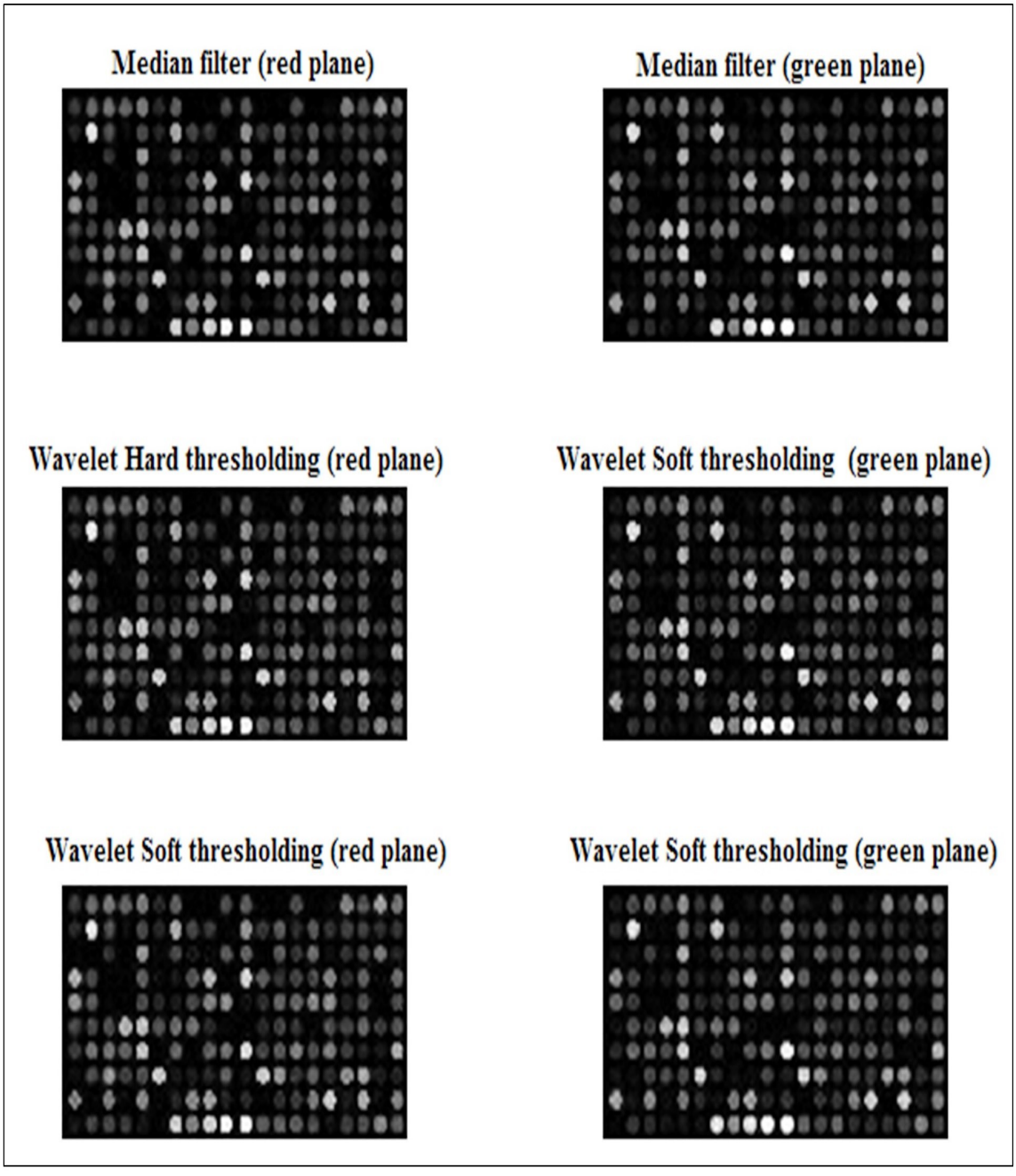


Figure 3.10: Result of microarray image de-noising

The SNR and PSNR (in decibels) values for de-noised microarray image using Median filter are presented in Table 3.1.

Table 3.1: Result of Median filtering

| Sr. No. | Variance | Median filter | |
|---|---|---|---|
| | | MSE | PSNR |
| 1 | 0.5 | 41.5635 | 31.9438 |
| 2 | 2.5 | 42.9435 | 31.8026 |
| 3 | 3.5 | 44.4479 | 31.6529 |
| 4 | 5 | 47.7892 | 31.3377 |
| 5 | 25 | 166.5944 | 25.9151 |
| 6 | 35 | 266.9120 | 23.8673 |

SNR and PSNR (in decibels) values for de-noised microarray image Hard thresholding are presented in Table 3.2.

Table 3.2: Result of Hard thresholding

| Sr. No. | Variance | Visushrink | | Bayesshrink | | Normalshrink | |
|---|---|---|---|---|---|---|---|
| | | MSE | PSNR | MSE | PSNR | MSE | PSNR |
| 1 | 0.5 | 36.9526 | 32.4548 | 26.6915 | 33.8672 | 26.0688 | 33.9715 |
| 2 | 2.5 | 41.8776 | 31.9110 | 32.0815 | 33.0745 | 31.4003 | 33.1615 |
| 3 | 3.5 | 45.8200 | 31.5206 | 37.4338 | 32.3989 | 36.1042 | 32.5559 |
| 4 | 5 | 54.8082 | 30.7423 | 46.6837 | 31.4411 | 45.3549 | 31.5653 |
| 5 | 25 | 273.024 | 23.7688 | 443.944 | 21.6577 | 380.170 | 22.3314 |
| 6 | 35 | 448.583 | 21.6125 | 806.824 | 19.0637 | 666.683 | 19.8902 |

Table 3.3 demonstrates SNR and PSNR (in decibels) values obtained for microarray image de-noising using Soft thresholding and the best of the SNR and PSNR (in

decibels) values obtained as a result of microarray image de-noising are presented in Table 3.4.

Table 3.3: Result of Soft thresholding

| Sr. No. | Variance | Visushrink | | Bayesshrink | | Normalshrink | |
|---|---|---|---|---|---|---|---|
| | | MSE | PSNR | MSE | PSNR | MSE | PSNR |
| 1 | 0.5 | 62.7249 | 30.2864 | 42.2523 | 32.2240 | 39.8600 | 32.3879 |
| 2 | 2.5 | 67.8803 | 29.9108 | 45.4607 | 31.7384 | 42.7701 | 32.0483 |
| 3 | 3.5 | 71.9939 | 29.6445 | 50.3006 | 31.2591 | 46.3786 | 31.6687 |
| 4 | 5 | 81.1672 | 29.1023 | 59.5512 | 30.4623 | 53.1157 | 30.9866 |
| 5 | 25 | 272.291 | 23.7823 | 406.157 | 22.0485 | 276.554 | 23.7169 |
| 6 | 35 | 421.735 | 21.8804 | 717.674 | 19.5780 | 463.101 | 21.4931 |

Table 3.4: Best of the result of microarray image de-noising

| Sr. No | Variance | Filter | MSE | PSNR |
|---|---|---|---|---|
| 1 | 0.5 | Hard thresholding, Normalshrink, level1 | 26.0688 | 33.9715 |
| 2 | 2.5 | Hard thresholding, Normalshrink, level1 | 31.4003 | 33.1615 |
| 3 | 3.5 | Hard thresholding, Normalshrink, level1 | 36.1042 | 33.5559 |
| 4 | 5 | Hard thresholding, Normalshrink, level1 | 45.3549 | 31.5653 |
| 5 | 25 | Median filter | 166.594 | 25.9151 |
| 6 | 35 | Median filter | 266.912 | 23.8673 |

In the case of image de-noising, the Soft thresholding tends to give smooth image while Hard thresholding preserves the edges in the image. On account of large number of edges in the microarray image, Hard thresholding (Normalshrink) outperforms Soft thresholding and Median filter for lower noise variance. At higher noise variance Gaussian noise appears like a salt and pepper noise. As a result, Median filter outperforms Hard and Soft thresholding.

Further, Stationary Wavelet Transform, Complex Wavelet transform can be used for image de-noising to obtain better performance than DWT based image de-noising.

Most of the techniques involved in Microarray technology namely, image gridding, image de-noising, image segmentation and gene expression data processing being independent research topics, the further part of this thesis is dedicated to gene expression data processing.

# Chapter 4

# Feature Selection

The microarray gene expression data usually comprises of considerably small number of samples (approximately 100 to 200) involving multitude of gene expressions values (approximately 50,000 to 60,000) [59]. However, many genes involved in gene expression data may not be useful in the classification of cancer. Therefore, it is imperative that the size of microarray data to be reduced before using it for cancer classification [60], [61]. Feature selection helps in dimension reduction of gene expression data. In this method, based on certain criteria few informative genes are mined from the original microarray data.

## 4.1 Feature selection methods

Different means of feature selection namely, Filter, Wrapper, Embedded, Hybrid and Ensemble [43], [62] method for labelled, unlabelled or semi-labelled data are explained in the following sub-topics.

### 4.1.1 Filter method

In Filter method, every gene is assigned a rank based on the inherent properties of the gene expression values such as correlation, distance, consistency and IG [61],

[63]. The subset of informative genes is selected on the basis of the gene rank and subsequently used for classification as demonstrated in the Figure 4.1.

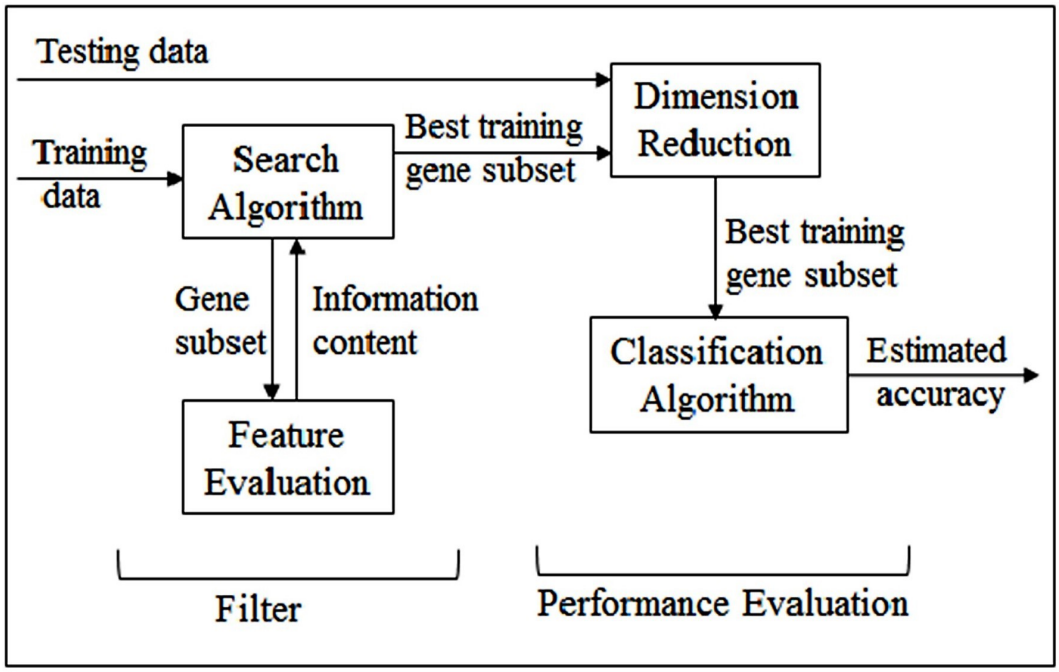


Figure 4.1: Block diagram for the cancer classification based on the Filter method

The filter methods may be univariate or multivariate [61], [43], [64].

1. Univariate filter method (UF)

   Univariate method assigns rank to the individual gene at a time using GR [65], IG [63], Euclidean Distance [66], Correlation based Feature Selection (CFS) [66], [67], Pearson Correlation Coefficient [68], TS [69] etc. Parametric UF method is based on the assumption that input data follows Gaussian or normal distribution. T-test [70], Gamma and Bayesian [71] are few examples of parametric feature selection method. Nonparametric UF method is usually applied, when the data does not follow normal distribution and/or when the number of input samples are small in number and not equal to the classes of input data. Few examples of the non-parametric method are Random Permutations [72], Rank Products [73] and Wilcoxon Rank Sum [74]. While revealing the statistical difference between

the classes of input data, parametric test is more viable than the non-parametric test.

UF methods offer higher speed of feature selection but fails to consider the variation of a particular gene intensity with respect to that of other genes. Moreover, the obtained gene subset may not be optimum as it does not act together with classifier [64].

2. Multivariate filter method (MF)

   While assigning the rank to the features, MF method considers the intensity variation of genes with respect to one another. Most popular methods for multivariate feature selection are MRMR [63], Fast CFS, Markov Blanket Filter [75] etc.

   Due to lack of interaction with the classifier, there is no considerable improvement in gene subset. Since this method involves all the genes in the process of assignment of gene rank, the computational complexity increases which, in turn reduces the speed of feature selection [64].

### 4.1.2 Wrapper method

Wrapper method is an iterative method which, involves the modification of initially generated gene subset and its performance evaluation till the desired classification accuracy is achieved. Feature subset is generated by using methods such as Sequential Backward Elimination, Sequential Forward Selection [43] etc. or chosen randomly using GA [76], Stimulated Annealing [77] etc. This generated gene subset is used for cancer classification. It is a simple method but tends to get stuck into local minimum during the training. Consideration of the variation of the gene intensities with respect to one another and the interaction with the classifier, makes the Wrapper method outperforms the Filter method. However, Wrapper method is computationally more expensive than the Filter method [60], [64]. The block diagram of the

cancer classification based on the Wrapper method is presented in Figure 4.2.

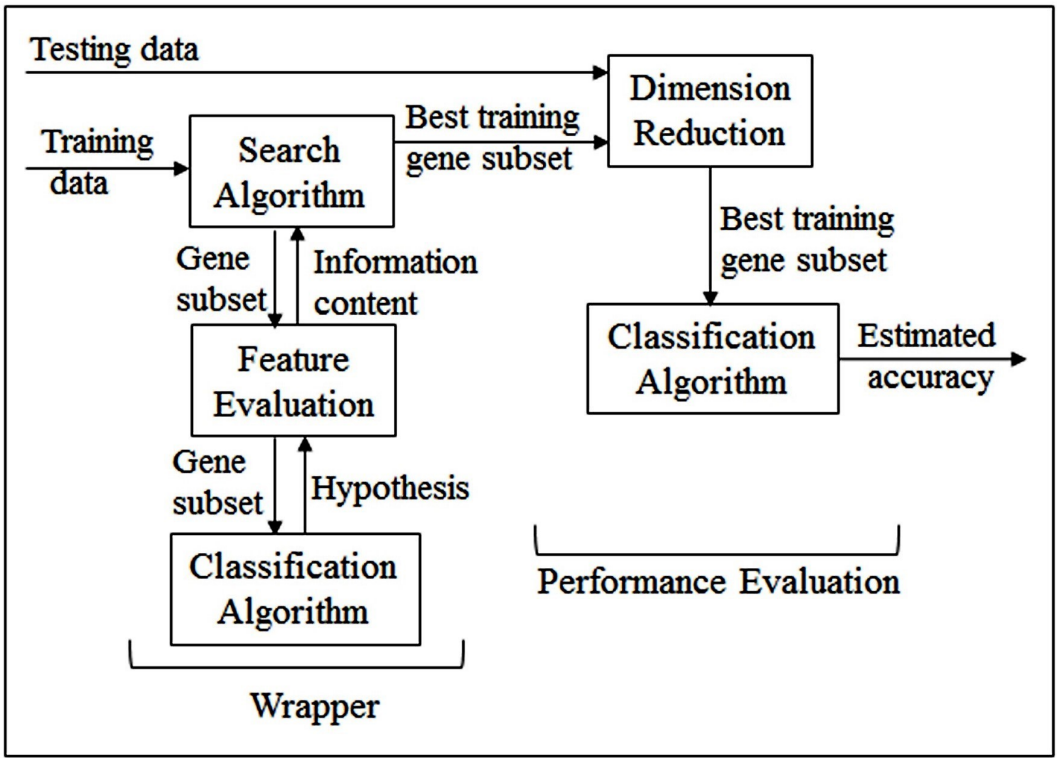


Figure 4.2: Block diagram for the cancer classification based on the Wrapper method

### 4.1.3 Embedded method

In this method, the gene selection is embedded within classification algorithm. With the help of iterative training of algorithm the genes from the initially generated subset that are useful for classification are retrained. Some of the examples of Embedded feature selection are SVM-RFE [78], RF [30] algorithms etc. While selecting the features, Embedded methods consider variation in gene intensities with respect to one another. However, they are specific to the classifier and many a time tend to have over-fitting problem [60], [64]. The block diagram of the cancer classification based on the Embedded method is demonstrated in Figure 4.3

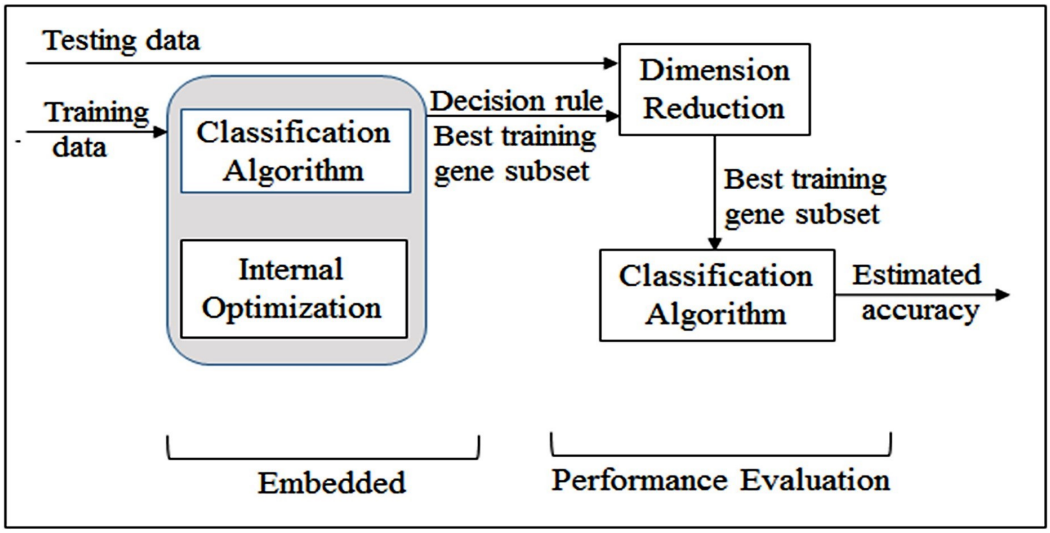


Figure 4.3: Block diagram of the cancer classification based on the Embedded method

### 4.1.4 Hybrid method

Hybrid method is the fusion of the feature selection methods. Usually, a reduced gene subset is obtained using Filter method and subsequently Wrapper method is used to optimize the obtained gene subset. Finally, cancer classification is implemented using the obtained gene subset. Few examples of Hybrid feature selection are MFMW [31], GADP [32] and FSVM [33] It gives better accuracy as compared to Filter method and reduces computational complexity in comparison with Wrapper method. Further, Hybrid method is specific to classification algorithm and has less tendency of over-fitting [43].

### 4.1.5 Ensemble method

Ensemble feature selection method involves selection and evaluation of various gene subsets based on certain criteria. Some of the examples of Ensemble feature selection are MF-GE [34] and EGSG [35]. This method has less tendency of over-fitting. However, it is complex, hard to comprehend and computationally expensive [43], [35]. Block diagram of cancer classification based on Hybrid and Ensemble gene

selection method is presented in Figure 4.4 and Figure 4.5, respectively.

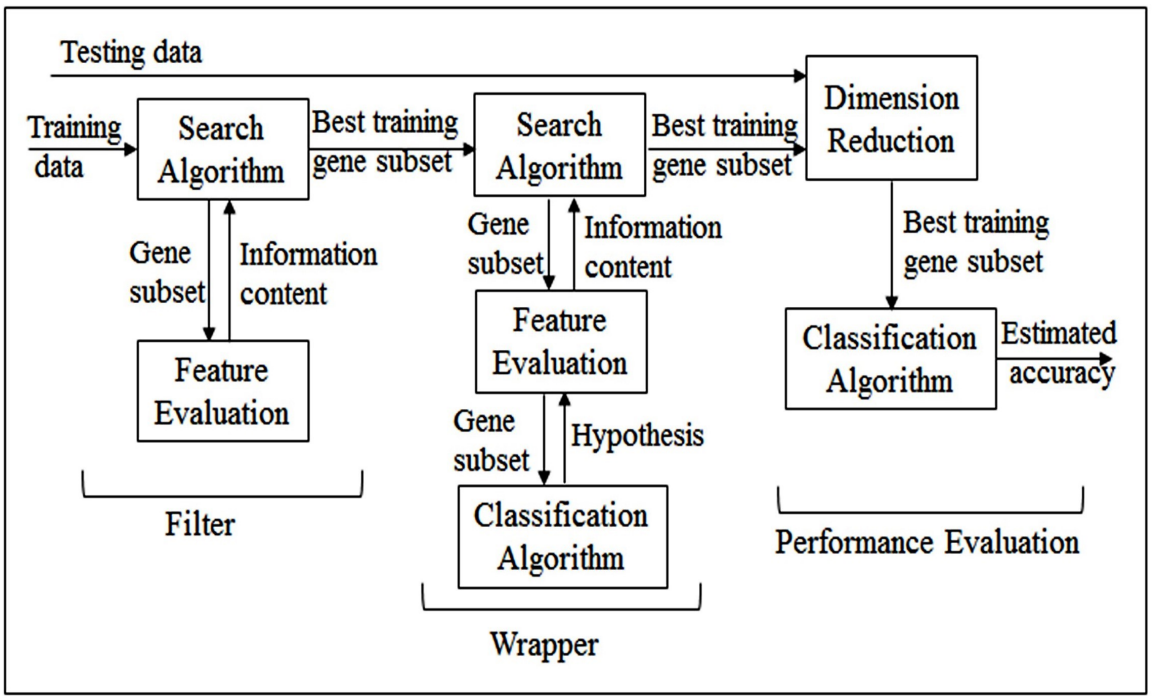


Figure 4.4: Block diagram for the cancer classification based on the Hybrid method

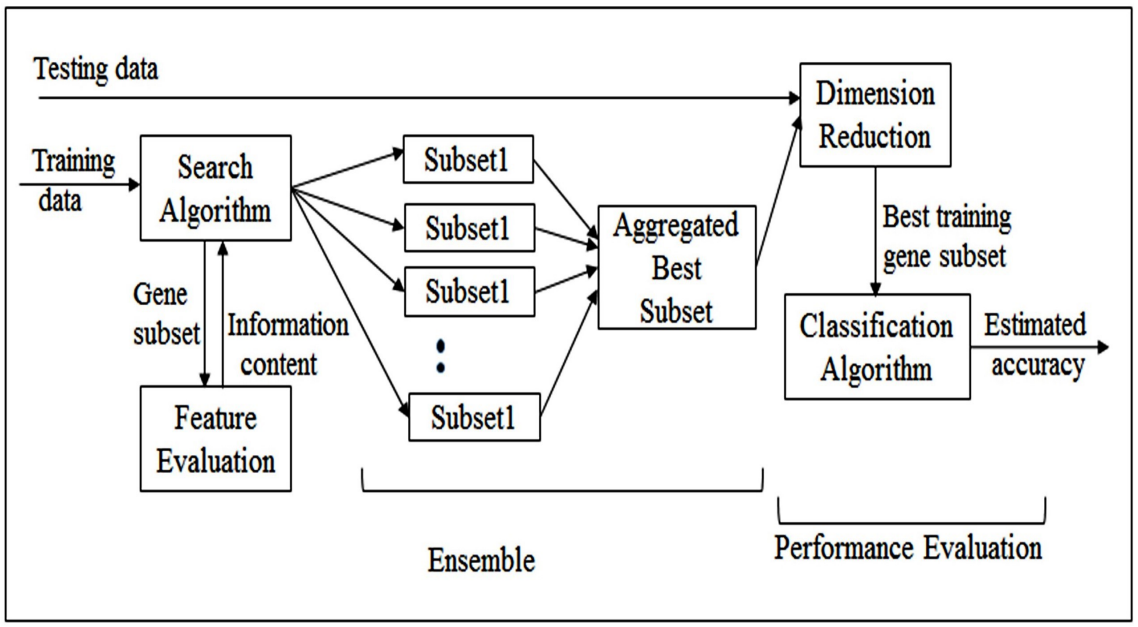


Figure 4.5: Block diagram for the cancer classification based on the Ensemble method

## 4.2 Feature selection methods in the proposed system.

### 4.2.1 Thresholding method

Thresholding method is a simple UF method which, confines the search space by discarding the genes with inconsistent intensity variation within the selected threshold range across the dataset samples. The flow chart for cancer classification based on Thresholding method is illustrated in Figure 4.6.

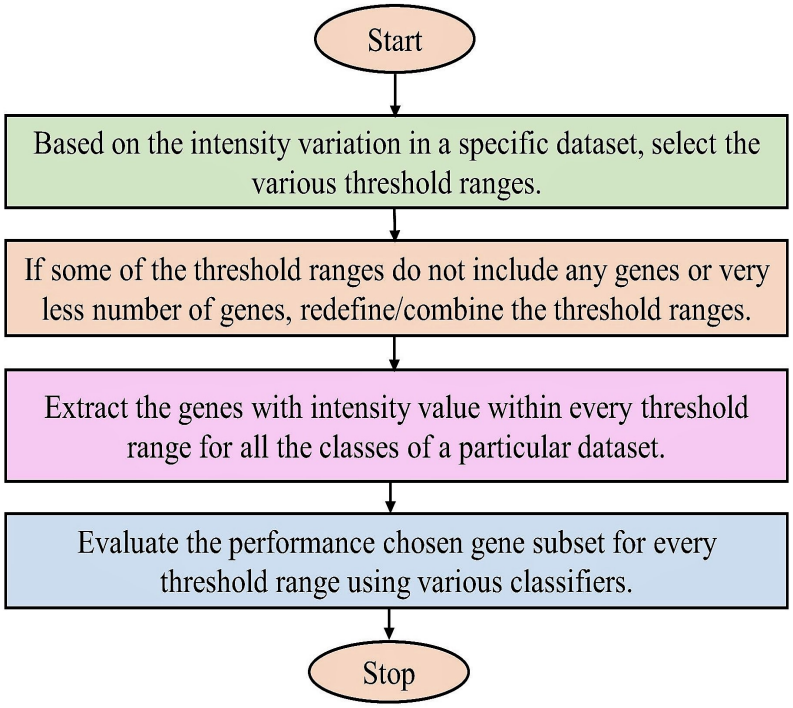


Figure 4.6: Flow chart of cancer classification based on Thresholding method

For the Glioma datasets utilized in the proposed work, the microarray gene expression data is generated from 16 bit microarray image. Therefore the threshold ranges considered as per the standard 1-2-5 sequence are 500-1000, 1000-2000, 2000-5000 and 5000-10000. Since many of the threshold ranges did not include any genes or very less number of genes, two consecutive threshold ranges are combined to form the threshold ranges as THD1 (500, 2000), THD2 (2000, 10000) and THD3 (10000,

100000). The gene intensities below 500 are neglected seeing that they mostly involve noise. The number of gene subsets obtained are same as that of number of threshold ranges for cancer dataset.

### 4.2.2 Ratio method

Ratio method limits the search space by eliminating the genes with inconsistent intensity variation across the dataset samples. The flow chart for the implementation of cancer classification based on Ratio method is demonstrated in Figure 7.14.

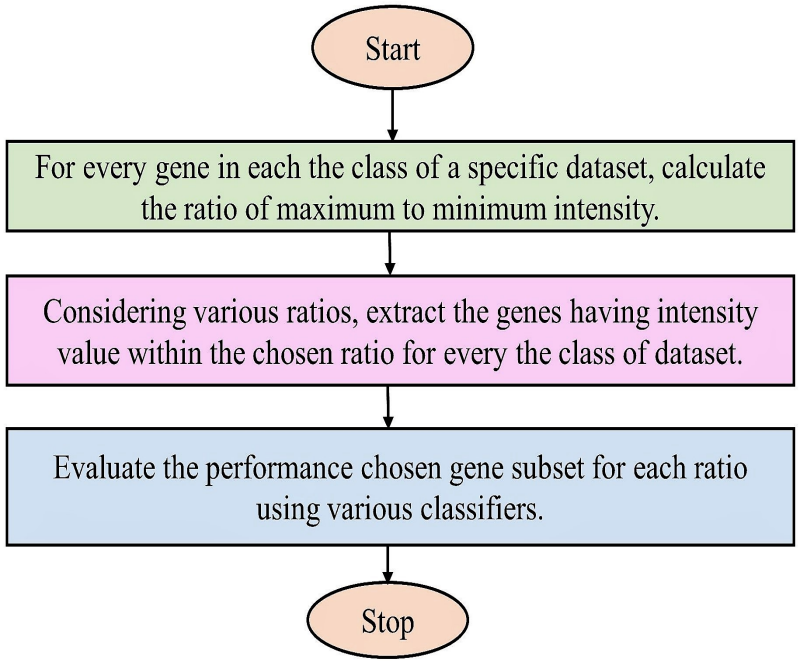


Figure 4.7: Flow chart of cancer classification based on Ratio method

Number of ratios chosen for Glioma datasets utilized in the proposed work are ratio <= 4, ratio <= 3.5, ratio <= 3, ratio <= 2.5.

### 4.2.3 Fusion of Thresholding and Ratio method

The fusion of Threshold and Ratio method is used to pool the advantages of both the methods. It augments the possibility of obtaining subset of genes with less ratio

within the specific range of threshold. The flow chart for the implementation of cancer classification based on Fusion of Threshold and Ratio method is presented in Figure 4.8

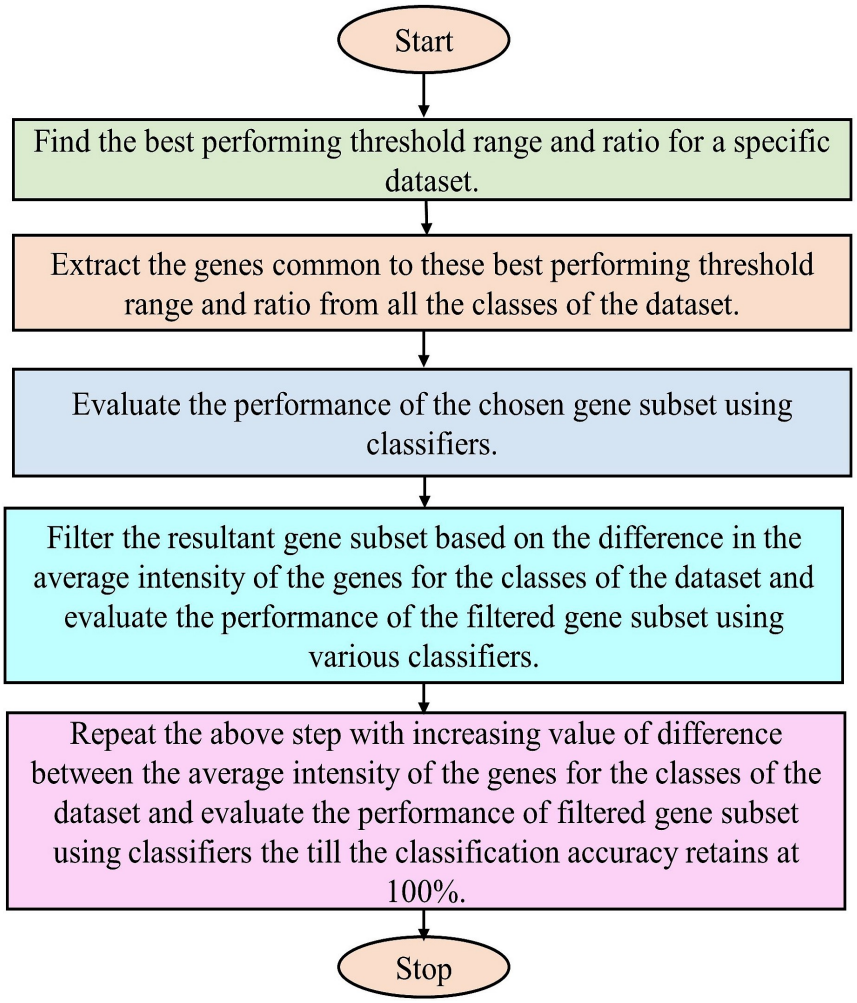


Figure 4.8: Flow Chart of cancer classification based on Fusion of Thresholding and Ratio method

The best performing threshold range and ratio is the one which gives the classification accuracy of 100% using almost all classification algorithms.

# Chapter 5

# Feature Extraction

Feature extraction is used independently or combined with feature selection for diminishing the size of gene expression data. In the process of dimension reduction of the gene expression data, feature extraction transforms it into a different domain, for example frequency domain [60]. Usage of feature extraction leads to substantial amount of dimension reduction for microarray data. However, during the process of feature extraction gene identity is lost.

## 5.1 Feature extraction methods

Feature extraction methods namely, PCA, DCT, FT, STFT and the proposed DWT based method are described in the following subsections.

### 5.1.1 Principal Component Analysis

PCA is a statistical method for diminishing the size of gene expression data [79]. It is an orthogonal transform which, converts the gene expression data to a set of uncorrelated variables known as principal components. Implementation steps of PCA are explained below [80], [81]

1. The matrix of gene expression data is organized in such way that rows signify

the samples and columns signify the variables.

2. The matrix of gene expression data is transformed to have zero mean and unit standard deviation.

3. Covariance matrix of resultant data is computed.

4. Eigen vectors and values of the covariance matrix are calculated.

5. Eigen vectors and values are reorganized to derive new transformed signal.

Reconstructed gene expression data using inverse PCA may not be same as original data. Further, the performance of PCA varies depending on the scaling factor [79].

### 5.1.2 Discrete Cosine Transform

DCT converts a sample of gene expression data into the linear combination of orthogonal cosine basis functions. These weighted basis functions represent the frequency components of a sample of the gene expression data [82]. Subsequently, DCT is applied to all the samples of gene expression data. DCT is applied to a sample of gene expression data with length P using Equation 5.1

$$A(s) = \alpha(s) \sum_{b=0}^{P-1} y(b) \cos\left(\frac{\pi(2b+1)s}{2P}\right); s = 0, 1, 2, ....P-1 \tag{5.1}$$

where,
$A(s)$ = DCT coefficient.
$y(b)$ = gene expression data sample
$b$ = gene number.
and $\alpha(s)$ is defined by Equation 5.2

$$\alpha(s) = \begin{cases} \sqrt{\frac{1}{P}}, & \text{if } s = 0 \\ \sqrt{\frac{2}{P}}, & \text{otherwise} \end{cases} \tag{5.2}$$

For a specific sample, DCT coefficients comprises of DC coefficient (first coefficients) and AC coefficients organized with increase in frequency [82]. DCT has less energy compaction ability and it is computationally less expensive as compared to DWT.

### 5.1.3 Fourier Transform

FT converts the gene expression data samples into complex exponentials of various frequencies, one at a time. The FT of a sample of gene expression data is calculated using Equation 5.3 [83].

$$Y(f) = \int_{-\infty}^{+\infty} y(b) \exp(-2jft\pi)db \tag{5.3}$$

where,

$Y(f)$ = FT coefficient of frequency f.

FT gives precise information about the frequency contents of samples of the gene expression data. However, it does not work well for non-stationary signals (the signals whose frequency content varies with time) [83].

### 5.1.4 Short Time Fourier Transform

In order to efficiently deal with non-stationary signals like microarray data, Short Time Fourier Transform (STFT) is introduced. It separates the individual sample of gene expression data into small portions and considers the separated portions to be stationary. It uses the window function with the size same as that of the every portion of the sample considered [84]. The STFT of a sample of gene expression data is calculated using Equation 5.4.

$$Y(\tau, W) = \int_{-\infty}^{+\infty} y(b)W(b-\tau) \exp(-j2pifb)db \tag{5.4}$$

where,

$Y(\tau, W)$ = STFT coefficient of frequency f

$W$ = window function

In the case of STFT, the window size is considered finite. As a result the band of frequencies that exists in the gene expression data samples are known rather than exact frequencies. Choice of the appropriate window size for a particular application is critical task which, limits the usage of STFT [84] [85].

### 5.1.5 Feature extraction method in proposed work-Wavelet Transform

Wavelet Transform works effectively in detection of the characteristics of data such as break points, disruptions in higher derivatives and self-similarity etc. Hence, in last few decades Wavelet transform has gained immense importance in the field of signal and image processing (for compression, de-noising, enhancement etc.). The wavelet transform efficiently analyzes stationary as well as non- stationary signals [86]. Gene expression data being non-stationary in nature [87], [88] wavelet transform appears to be most suitable transform for its analysis. It is a linear transformation method used to convert the time domain signal to the frequency domain signal. The details of Continuous Wavelet Transform (CWT) and DWT are explained below.

#### Continuous Wavelet Transform

The steps in implementation of CWT are given below

1. At a particular scale, a sample of gene expression data is compared with shifted versions of mother wavelet function and the correlation coefficient between the signal under consideration and the wavelet function is calculated for every value of time shift.

2. The above process is repeated for every value of the scale.

3. Step 2 and Step 3 is repeated for all microarray data samples.

It gives number of wavelet coefficients as a frequency domain representation of samples of gene expression data. The scale parameter is inversely proportional to frequency. The small value of scale parameter (high frequency) represents the coarse

information in the signal while, large value of scale parameter (low frequency) represents the detailed information of the signal. Since CWT is calculated with continuous variation in values of shifting and scaling parameter, the large number of CWT coefficients are generated. The CWT output is redundant in nature [86]. Further, it is computationally expensive.

The CWT of a sample of gene expression data is calculated using Equation 5.5.

$$Y(sc, si) = \left(\frac{1}{\mid sc \mid}\right)^{\frac{1}{2}} \int_{-\infty}^{+\infty} y(b)\psi\Big(\frac{b - si}{sc}\Big)db \tag{5.5}$$

where,

$Y(sc, si)$ = CWT coefficient

$\psi\Big(\frac{b-si}{sc}\Big)$ = mother wavelet

$sc$ = wavelet scale parameter

$si$ = wavelet time shift parameter.

#### Discrete Wavelet Transform

In the case of DWT, non-redundant output is obtained with the use of scaling parameter and the shifting parameter that is varied by the factor of two (2, 4, 6 etc.) [89]. To facilitate the same, DWT has a set of functions known as scaling (LPF) and wavelet function (HPF). According to Mallat′s algorithm, to compute DWT of a sample of the gene expression data, it is successively passed through HPF and LPF. Every time the sample is passed through the filters, approximate and detailed coefficients are generated. In order to diminish the redundancy, the resultant wavelet coefficients are down sampled by two (DS2) [55].This process is repeated for all the samples of gene expression data.

The approximate and detailed coefficients of a sample of gene expression data are given by Equation 5.6 and Equation 5.7 [56]

$$q_l(si) = \sum_{m-2si}^{2si+N-1} r(m - 2si)q_{l+1}(m) \tag{5.6}$$

$$p_l(si) = \sum_{m-2si}^{2si+N-1} s(m-2si)p_{l+1}(m) \tag{5.7}$$

where,

$s(n)$ = impulse response of HPF

$r(n)$ = impulse response of LPF

$N$ = number of wavelet coefficients.

The process of application of DWT to a sample of gene expression data is demonstrated in the Figure 5.1 .

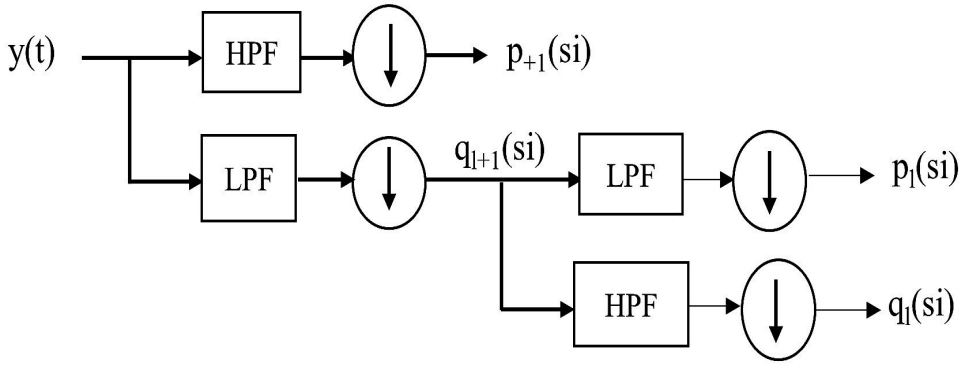


Figure 5.1: DWT process.

The down sampled approximate and/or detailed coefficients are used for cancer classification. Due to its ability of providing multiresolution analysis and localized time-frequency information wavelet transform outperforms the other transforms. Also wavelet transform provides higher energy compaction as compared to other transforms. Varieties of mother wavelets such as Haar, Daubechies, Symlets, Coiflets, Bio-orthogonal, Meyer, Mexican Hat etc. are available [56]. The mother wavelets are different from each other with respect to characteristics such as regularity, symmetry, number of vanishing moments etc [56]. The varieties of available mother wavelet allows the researcher to compare their results using number of wavelets and in turn to choose the mother wavelet suitable for a particular application. There is no one common mother wavelet that is suitable for every applications [90]. The optimal

mother wavelet function suitable for a particular application depends on the nature of variation of the input data under consideration.

To determine such a wavelet, it is necessary to consider one of the following parameter [90].

1. Properties of mother wavelet
2. Degree of matching of the mother wavelet with the signal under consideration
3. Cross correlation between approximate wavelet coefficients and signal under consideration
4. In the case of classifier, classification accuracy
5. MSE

Of all the parameters, degree of matching between mother wavelet and signal under consideration and MSE are the most commonly used measures for the selection of the optimal wavelet [90], [91], [92], [93], [94].

In the proposed work, the wavelets used for the feature extraction of gene expression data are namely, Db2, Db4, Sym2, Sym4, Bior1.3 and Bior2.4.

# Chapter 6

# Classification Algorithms

ANN efficiently processes large volume, non-linear and chaotic data such as gene expression data. Pattern recognition, classification, control and time series prediction are some of the important applications of ANN. An artificial neuron modeled to get all benefits of biological neuron is the basic building block of ANN. It learns with examples and offers several advantages such as

1. Adaptive learning ability
2. Self-organizing nature
3. Faster information processing ability due to parallel processing
4. An ability to handle inconsistent and random information
5. An ability to handle missing and faulty data
6. An ability to process nonlinear data efficiently.

Various ANN classification algorithms namely EBPA, RPROP, Conjugate Gradient, LM, SAEN are explained in the following sub-sections.

## 6.1 Error Back Propagation algorithm

EBPA is one of the most common supervised algorithm for training of the multi-layer neural network which, is applicable only for the continuous neuron. The multilayer neural network has input layer, one or more hidden layers and an output layer. In EBPA, the sigmoid and linear activation function is usually used for hidden and output layer neurons, respectively. The multi-layer neural network with I inputs, one hidden layer, J hidden neurons and K output neurons is as shown in the Figure 6.1.

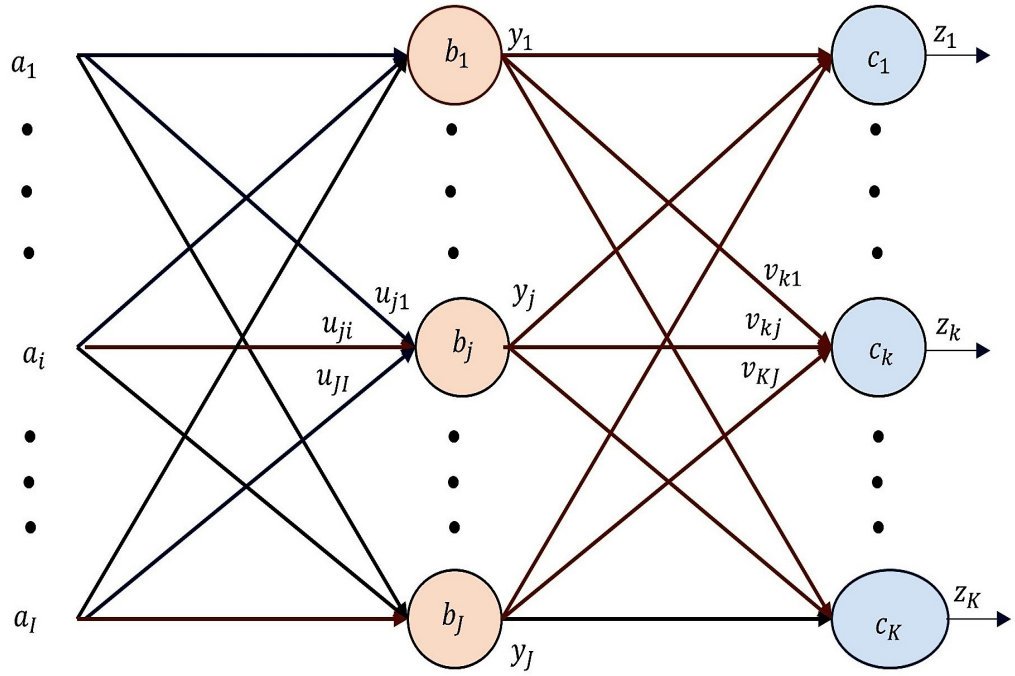


Figure 6.1: Multilayer neural network.

In the case of EBPA, initially, the output of the hidden layer neurons and output layer neurons is computed. Subsequently, the difference between actual and expected output of final layer neurons is back propagated to update the weights of output layer neurons and hidden layer neurons.

The flowchart for EBPA is shown in Figure 6.2

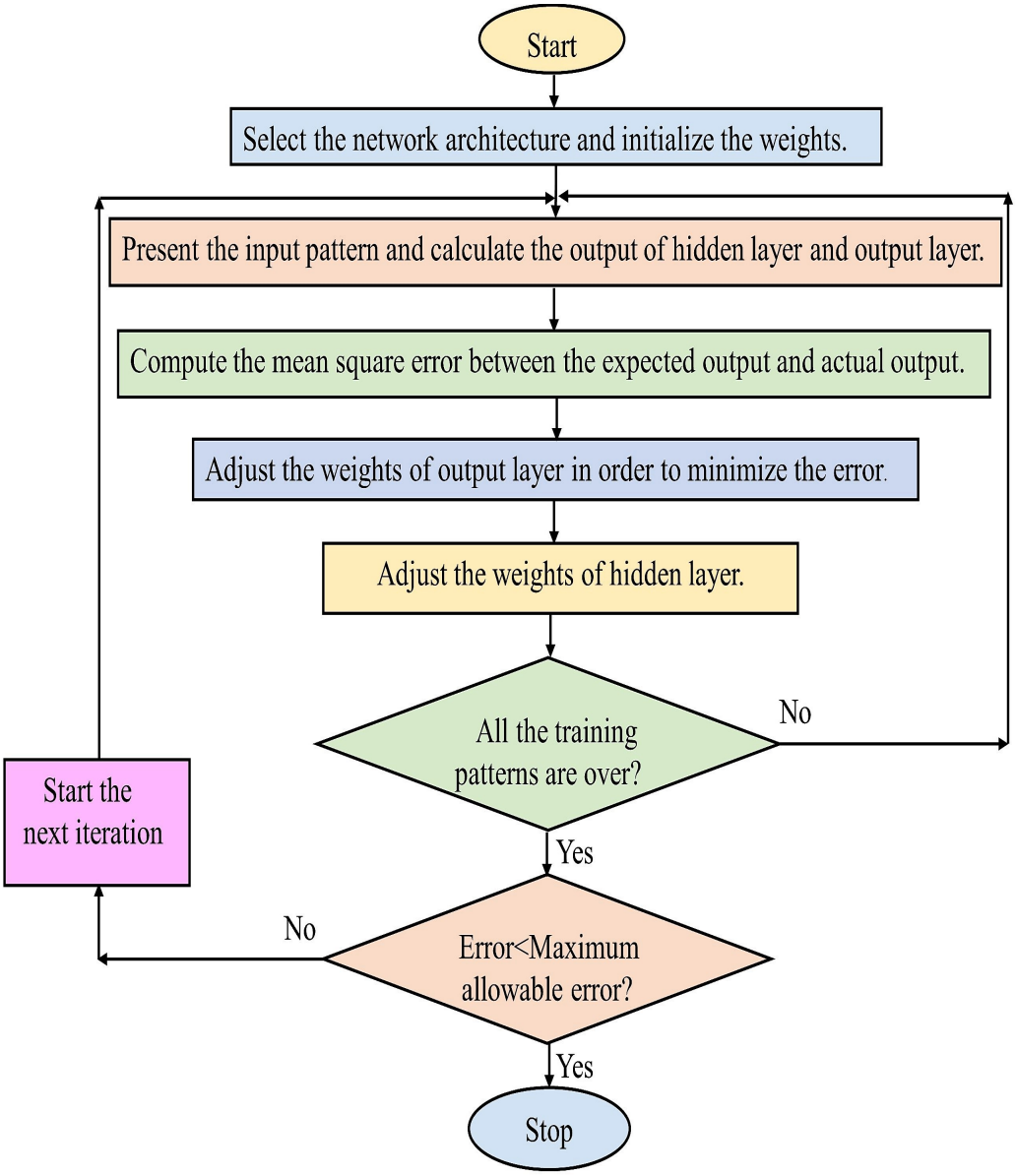


Figure 6.2: Flow chart for EBPA.

The weight update rule for output layer neurons is described using Equation A.3.

$$v_{kj}(t+1) = v_{kj}(t) + c(e_k - z_k){z_k}' y_j \tag{6.1}$$

Where,

$v_{kj}(t+1)$ = weight that connects output of $j^{th}$ neuron in the hidden layer to $k^{th}$

neuron in the output layer at (t+1)

$v_{kj}(t)$ = weight that connects output of $j^{th}$ neuron in the hidden layer to $k^{th}$ neuron in the output layer at t

$c$ = learning constant

$e_k$ = expected output of $k^{th}$ output layer neuron

$z_k$ = actual output of $k^{th}$ neuron in the output layer

$z_k^{'}$ = derivative of actual output of $k^{th}$ neuron in the output layer

$y_j$ = output of $j^{th}$ neuron in hidden layer.

The weight update rule for hidden layer neurons is described using Equation A.7.

$$u_{ji}(t+1) = u_{ji}(t) + cy_j^{'} a_i \sum_{k=0}^{K} (e_k - z_k) z_k^{'} v_{kj} (6.2)$$

where,

$u_{ji}(t+1)$ =weight that connects $i^{th}$ input to $j^{th}$ neuron in the hidden layer at (t+1)

$u_{ji}(t)$ = weight that connects $i^{th}$ input to $j^{th}$ neuron in the hidden layer at t

$y_j^{'}$ = derivative of output of $j^{th}$ neuron in the hidden layer

$a_i = i^{th}$ input to neural network.

EBPA makes the individual weight change proportional to the slope of error curve. The slope of error curve is proportional to the learning constant, difference between input and output and the derivative of the output of corresponding neuron. For larger inputs, as the actual output of neuron increases, derivative of the error drops off. As a consequence of reduction in weight change, the classification accuracy gets affected with increased difference between input and output. Further, the chosen value of learning constant aggravates the effect of derivative. Smaller learning constant diminishes the speed of convergence while, larger learning constant reduces the possibility of reaching the convergence. To overcome this problem, EBPA with momentum is introduced but choice of appropriate momentum parameter is crucial [95], [96], [57].

## 6.2 Resilient Back Propagation algorithm

To eliminate the effect of magnitude of derivative of error on the convergence, RPROP algorithm makes the weight update proportional to sign of the error derivative instead of its magnitude. The RPROP algorithm is explained in Figure 6.3.

Initialize $\Delta_{kj}(t) = 0$ and $\frac{\partial E}{\partial v_{kj}}(t-1) = 0$

Repeat

Calculate $\frac{\partial E}{\partial v_{kj}}(t)$

For all biases and weights {

If $\left(\frac{\partial E}{\partial v_{kj}}(t-1) * \frac{\partial E}{\partial v_{kj}}(t) > 0\right)$ then

{ $\Delta_{kj}(t) = minimum(\Delta_{kj}(t-1) * c^{+}, \Delta_{max})$

$\Delta v_{kj}(t) = -sign\left(\frac{\partial E}{\partial v_{kj}} * \Delta_{kj}(t)\right)$

$v_{kj}(t+1) = v_{kj}(t) + \Delta v_{kj}(t)$

$\frac{\partial E}{\partial v_{kj}}(t-1) = \frac{\partial E}{\partial v_{kj}}(t)$ }

Else if $\left(\frac{\partial E}{\partial v_{kj}}(t-1) * \frac{\partial E}{\partial v_{kj}}(t) < 0\right)$ then

{ $\Delta_{kj}(t) = maximum(\Delta_{kj}(t-1) * c^{-}, \Delta_{min})$

$v_{kj}(t+1) = v_{kj}(t) - \Delta v_{kj}(t-1)$

$\frac{\partial E}{\partial v_{kj}}(t-1) = 0$ }

Else if $\left(\frac{\partial E}{\partial v_{kj}}(t-1) * \frac{\partial E}{\partial v_{kj}}(t) = 0\right)$ then

{ $\Delta v_{kj}(t) = -sign\left(\frac{\partial E}{\partial v_{kj}}(t)\right) * \Delta_{kj}(t)$

$v_{kj}(t+1) = v_{kj}(t) + \Delta v_{kj}(t)$

$\frac{\partial E}{\partial v_{kj}}(t-1) = \frac{\partial E}{\partial v_{kj}}(t)$

}

}

Figure 6.3: RPROP algorithm.

In the Figure 6.3 $\frac{\partial E}{\partial v_{kj}}(t-1)$, $\frac{\partial E}{\partial v_{kj}}(t)$ and $\frac{\partial E}{\partial v_{kj}}(t+1)$ indicates the partial derivative of error with respect to $v_{kj}$ at instance $t-1$, $t$ and $t+1$, respectively. Further, $\Delta v_{kj}(t-1)$, $\Delta v_{kj}(t)$ and $\Delta v_{kj}(t+1)$ indicates change in weight $v_{kj}$ at instance $t-$

1, $t$ and $t+1$, respectively. In two consecutive iterations if the sign of the error derivative remains same, then the amount of weight update,which is controlled by $\Delta$, is increased in order to attain the global minimum of the error curve faster. In two subsequent iterations, the change in sign of the gradient point to larger weight update at the previous instance. Therefore, the $\Delta$ is reduced. Also, if the product of error derivatives in two consecutive iterations becomes zero, the weights are retained to the previous values. Usually, for RPROP algorithm, the default values of the parameters are 0.1, $e^{-6}$, 50, 1.2 and 0.5 for $\Delta_0$, $\Delta_{min}$, $\Delta_{max}$, $c^+$ and $c^-$, respectively. RPROP algorithm does not require any critical parameter setting. Further, it is more complex and requires more memory as compared to EBPA. However, convergence speed of RPROP is faster than EBPA [97], [57].

## 6.3 Levenberg Marquardt algorithm

In practical applications error curve is complex in nature. For such complex error curves, EBPA works efficiently. The EBPA has very slow convergence as its speed depends on various parameters namely, learning constant, number of training samples, complexity of activation function etc. The speed of EBPA is increased to some extent by using adaptable learning constant. Significant increase in the speed of convergence is obtained using Newton's method. Newton's method makes use of Hessian matrix which, is formed using double differentiation of the error. With Newton's method, there is considerable improvement in the speed of convergence but it does not work well for complex error curve. Further, the computation of double differentiation of error, makes the Newton's method computationally expensive. To increase convergence speed with less computational complexity Gauss Newton algorithm is introduced. Gauss Newton algorithm uses Jacobian matrix that is formed by using the first differentiation of error as opposed to Newton's method. Gauss Newton algorithm offers higher speed and less computational complexity but fails while dealing with non-quadratic error curve. LM algorithm is one of the most competent algo-

rithms for training multilayer perceptron network which, is introduced to combine the advantage of EBPA as well as Gauss Newton algorithm. It operates like an EBPA closer to the complex section of error curve while during the quadratic section of the error curve, it operates like Gauss Newton algorithm [57], [98] . The weight update rule for LM algorithm is given by Equation 6.3

$$v_{kj}(t+1) = v_{kj}(t) - ({J_k}^T J_k + \mu B) J_k^T b_k \tag{6.3}$$

Where,

$J_k$ = Jacobian matrix

$B$ = Identity matrix

$\mu$ = fusion coefficient

$b_k$ = error vector.

In the case $\mu$ is very small the algorithm tends to approach Gauss Newtons algorithm while, larger $\mu$ causes the algorithm to behave like EBPA.LM algorithm turns out to be very efficient for small and moderate size multilayer neural network. The increased memory requirement, for large size multilayer neural network makes LM algorithm very sluggish.

## 6.4 Conjugate Gradient algorithms

In order to attain the global minimum of quadratic error curve, EBPA performs a linear search and makes the successive search path orthogonal to the former search direction. In the case of Conjugate Gradient algorithms, the successive search path is A-orthogonal to just preceding search path [99], [57]. It leads to increase in the speed of convergence of Conjugate Gradient algorithms. The new search direction is determined by Equation 6.4.

$$F = g * h + d \tag{6.4}$$

where,

$F$ = new search direction

$g$ = multiplicative factor

$h$ = previous search direction

$d$ = direction of steepest descent.

The multiplicative constant '$g$' is calculated in number of ways for various Conjugate Gradient algorithms. For Conjugate Gradient Back Propagation with Fletcher-Reeves Update (CGFR) algorithm, '$g$' is calculated using Equation 6.5 [100], [57],

$$g = EC/EP \tag{6.5}$$

where,

$EC$ = energy in the current gradient

$EP$ = energy in the previous gradient.

For Conjugate Gradient Back propagation with Polak-Ribire Update (CGPR) algorithm, '$g$' is calculated using Equation 6.6 [101], [57].

$$g = (EC - EP)/EP \tag{6.6}$$

Usually, when the number of iterations becomes same as the number of network parameters, Conjugate Gradient algorithms converge. If the algorithms do not converge within the number of iterations equaling number of neural network parameters, the search direction is usually reset.

In Conjugate Gradient Back Propagation with Powell-Beale Restarts (CGPB) algorithm, the search path resets for very small orthogonality between two succeeding gradients [102], [103], [57].

The flow chart of Conjugate Gradient algorithms shown in the Figure 6.4

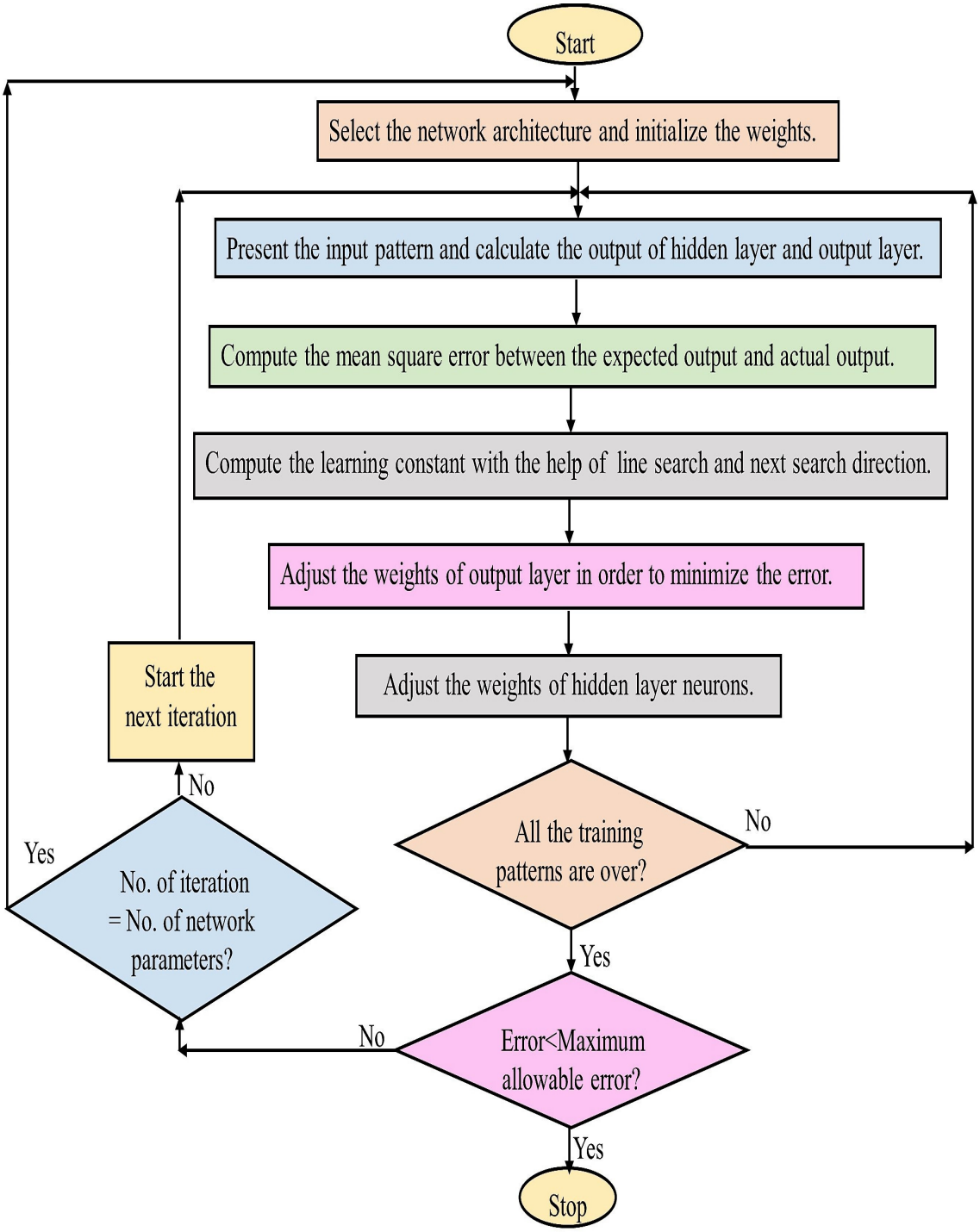


Figure 6.4: Conjugate Gradient Back Propagation algorithm.

## 6.5 Stacked Autoencoder algorithm

An Autoencoder type of neural network makes use of an unsupervised back propagation training algorithm. It comprises of an encoder and a decoder as illustrated in Figure 6.5.

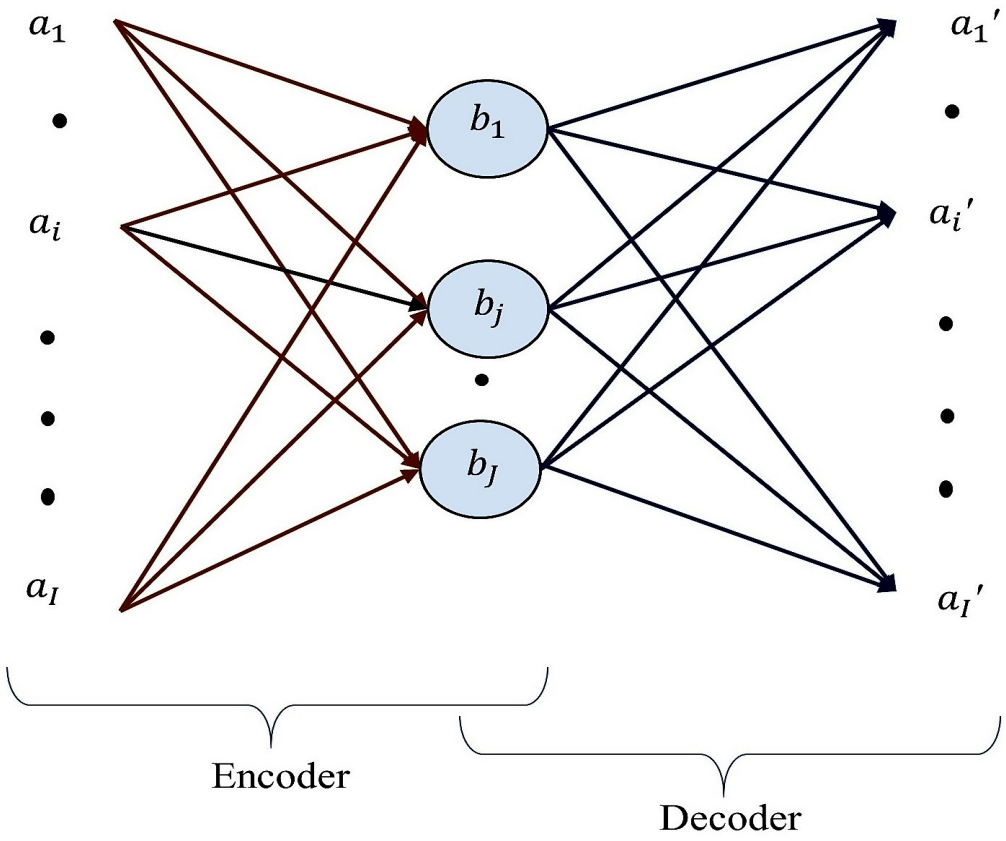


Figure 6.5: Autoencoder.

To extract the features from the input data, an encoder transforms the input data into a hidden representation with the help of Equation 6.7.

$$Y = UD \tag{6.7}$$

where,

$Y$ = transformed input at the output of hidden layer of an autoencoder

$U$ = weight vector of encoder

$D$ = input to an autoencoder.

The decoder transforms the hidden representation back to the input data using Equation 6.8.

$$Y' = VY \tag{6.8}$$

where,

$Y'$ = reconstructed input

$V$ = weight vector of decoder

$Y$ = input to decoder.

To get the best promising representation of the input, the difference between the initial input and reconstructed input is used to update the weights. The SAEN network [104], [104] is made up of one or more Autoencoders followed by the Softmax layer. The SAEN network is shown in Figure 6.6

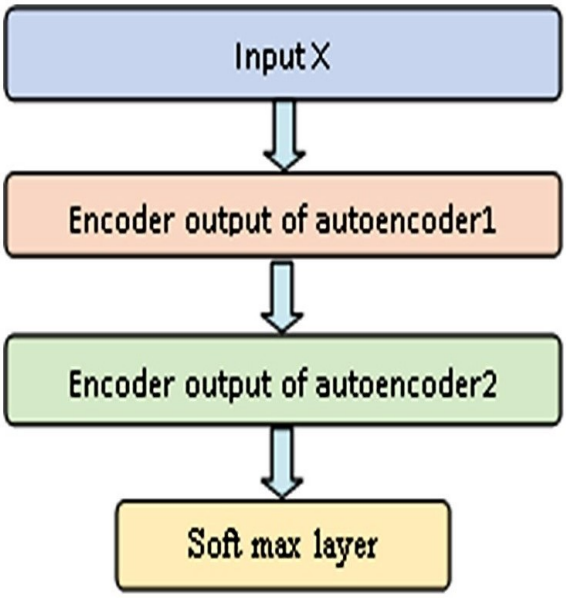


Figure 6.6: Stacked Autoencoder network.

SAEN network training is accomplished by unsupervised prior-training of Autoencoders followed by supervised training of Softmax Layer and finally fine tuning of all

the stages of SAEN. The detailed steps in training the SAEN are explained below:

1. The first Autoencoder is trained to diminish the difference between the initial input and reconstructed input.
2. Neglecting the reconstructed input and considering the output of hidden layer of first Autoencoder as input to second Autoencoder, the training of the second Autoencoder is accomplished using unsupervised learning.
3. This process is repeated for following Autoencoders.
4. Considering output of hidden layer of final Autoencoder as input to Softmax layer, it is trained using the supervised Back Propagation training algorithm.
5. Finally, the entire SAEN network is trained using supervised Back Propagation training algorithm for fine tuning of the weights and biases.

Due to unsupervised pre-training of Autoencoders, supervised training of Softmax layer and fine tuning of entire network, SAEN algorithm delivers high classification accuracy for less number of inputs.

In the proposed system cancer classification is implemented using RPROP, LM, Conjugate Gradient and SAEN algorithm.

# Chapter 7

# Results, Conclusion and Future Scope

## 7.1 Results

In this section, the classification results of Brain tumor dataset GDS1962 with and without use of DCT and DWT (traditional methods) are presented. It includes the comparative results of classification implemented using Thresholding method, Ratio method and Fusion of Thresholding and Ratio methods followed by DWT based feature extraction for Glioma Grade III/ Grade IV datasets viz., GDS1975, GDS1976, GDS1815 and GDS1816. The classification is implemented using RPROP, Conjugate Gradient, LM and SAEN algorithms. Further, the comparative results of proposed system with the methods suggested by Abusamra H et al. [23] and Shen Q et al. [16] for GDS1975 and GDS1976 datasets as well as results of testing of effectiveness of the resultant optimal gene subset for GDS1962 dataset are presented.

### 7.1.1 GDS1962 results

For GDS1962 dataset the classification of sub-types of Brain tumor is implemented for four different levels of malignancy namely,

1. Malignant and Benign Brain Tumor
2. Lymphoma and Glioma Brain Tumor
3. Low Oligodendroglioma, High Oligodendroglioma and Astrocytoma
4. Astrocytoma Grade II, Grade III and Grade IV.

**Malignant and Benign Brain tumor**

The Benign and Malignant Brain tumor gene intensity values are far apart from each other which makes it easy to differentiate between them without using feature extraction method. The classification accuracy of 100 % is achieved using all the DCT coefficients while, with DWT based feature extraction method the 100% classification accuracy is obtained by using less than five wavelet coefficients. The result of classification for Malignant and Benign Brain tumor is presented in Table 7.1

Table 7.1: Result of classification for Malignant and Benign Brain tumor

| Sr. No. | Algorithm | Feature Extraction | Wavelet | Accuracy |
|---|---|---|---|---|
| 1 | RPROP | — | — | 100% |
| 2 | RPROP | DWT | Db2, Db4, Sym2, Sym4, Bior1.3 and Bior2.4 (level 16) | 100% |
| 3 | RPROP | DCT | — | 100% |

**Lymphoma and Glioma Brain tumor.**

For Lymphoma and Glioma Brain tumors, the large difference in the gene intensity values makes it easy to distinguish them without using feature extraction method. The classification accuracy of 97 % is achieved using all the DCT coefficients while,

with DWT based feature extraction method the 100% classification accuracy is obtained by using less than five wavelet coefficients. The result of classification for Lymphoma and Glioma Brain tumor is demonstrated in Table 7.2.

Table 7.2: Result of classification for Lymphoma and Glioma Brain tumor

| Sr. No. | Algorithm | Feature Extraction | Wavelet | Accuracy |
|---|---|---|---|---|
| 1 | RPROP | — | — | 100% |
| 2 | RPROP | DWT | Db2, Db4, Sym2, Sym4, Bior1.3 and Bior2.4 (level 16) | 100% |
| 3 | RPROP | DCT | — | 97% |

**Low Oligodendroglioma, High Oligodendroglioma and Astrocytoma.**

As the level of malignancy goes on increasing, genes intensity values of cancer subtypes appear to be closer to each other making it difficult to differentiate between them. The classification accuracy obtained with and without DCT and DWT for different types of Glioma such as Low Oligodendroglioma, High Oligodendroglioma and Astrocytoma is illustrated in Table 7.3 and Table 7.4.

Table 7.3: Result of classification for sub-types of Glioma using DCT.

| Sr. No. | Algorithm | Accuracy |
|---|---|---|
| 1 | RPROP | 65% |
| 2 | CGPR | 81% |
| 3 | CGPB | 84% |
| 4 | CGFR | 81% |

Table 7.4: Result of classification for types of Glioma using DWT.

| Sr. No. | Algorithm | Wavelet | Level | Accuracy |
|---|---|---|---|---|
| 1 | RPROP | Db4 | 1 | 89% |
| 2 | CGPR | Db4 | 1 | 97% |
| 3 | CGPB | Db4 | 1 | 94% |
| 4 | CGFR | Sym4 | 4 | 91% |

Table 7.5 presents comparison of DCT and DWT results for sub-types of Glioma.

Table 7.5: DCT vs. DWT for classification of sub-types of Glioma.

| Sr. No. | Feature Extraction | Algorithm | Wavelet | Level | Accuracy |
|---|---|---|---|---|---|
| 1 | – | RPROP | — | — | 85.9% |
| 2 | DCT | CGPB | — | — | 84% |
| 3 | DWT | CGPR | Db4 | 1 | 97% |

The custom made filters, higher energy compaction and flexibility of choosing the wavelet function makes DWT outperform DCT for the classification of various types of Glioma as demonstrated in Table 7.5.

**Astrocytoma Grade II, Grade III and Grade IV.**

The malignancy level of Astrocytoma subtypes is higher than that of Glioma subtypes. As a result, the gene intensity values of Astrocytoma subtypes are closer to each other as compared to Glioma subtypes, making the classification a difficult task. The classification accuracy obtained with and without DCT, DWT for different types of Astrocytoma such as Grade II, Grade III and Grade IV is illustrated in Table 7.6 and Table 7.7.

Table 7.6: Result of classification for sub-types of Astrocytoma using DCT.

| Sr. No. | Algorithm | Accuracy |
|---|---|---|
| 1 | RPROP | 49% |
| 2 | CGPR | 68% |
| 3 | CGPB | 81% |
| 4 | CGFR | 68% |

Table 7.7: Result of classification for sub-types of Astrocytoma DWT.

| Sr. No. | Algorithm | Wavelet | Level | Accuracy |
|---|---|---|---|---|
| 1 | RPROP | Bior2.4 | 4 | 87% |
| 2 | CGPR | Bior2.4 | 4 | 92% |
| 3 | CGPB | Db2 | 1 | 87% |
| 4 | CGFR | Sym4 | 4 | 89% |

The comparison of DCT and DWT results for sub-types of Astrocytoma is illustrated in Table 7.8 .

Table 7.8: DCT Vs. DWT for the classification of sub-types of Astrocytoma.

| Sr. No. | Feature Extraction | Algorithm | Wavelet | Level | Accuracy |
|---|---|---|---|---|---|
| 1 | — | RPROP | – | – | 91.89% |
| 2 | DCT | CGPB | – | – | 81% |
| 3 | DWT | CGPR | Bior2.4 | 4 | 92% |

### 7.1.2 Thresholding method

The classification results of Glioma Grade III and Grade IV datasets namely, GDS1975, GDS1976, GDS1815 and GDS1816 for various threshold ranges viz., THD1 (500-

2000), THD2 (2000-10000) and THD3 (10000-100000) are presented in this section. With THD1 (500-2000), the number of genes selected are 574, 330, 493 and 521 for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively. With THD2 (2000-10000), the number of genes selected are 1301, 1413, 147 and 118 for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively. With THD3 (10000-100000), the number of genes selected are 274, 185, 32 and 41 for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively.

Figure 7.1 demonstrates the result of Thresholding method for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively.

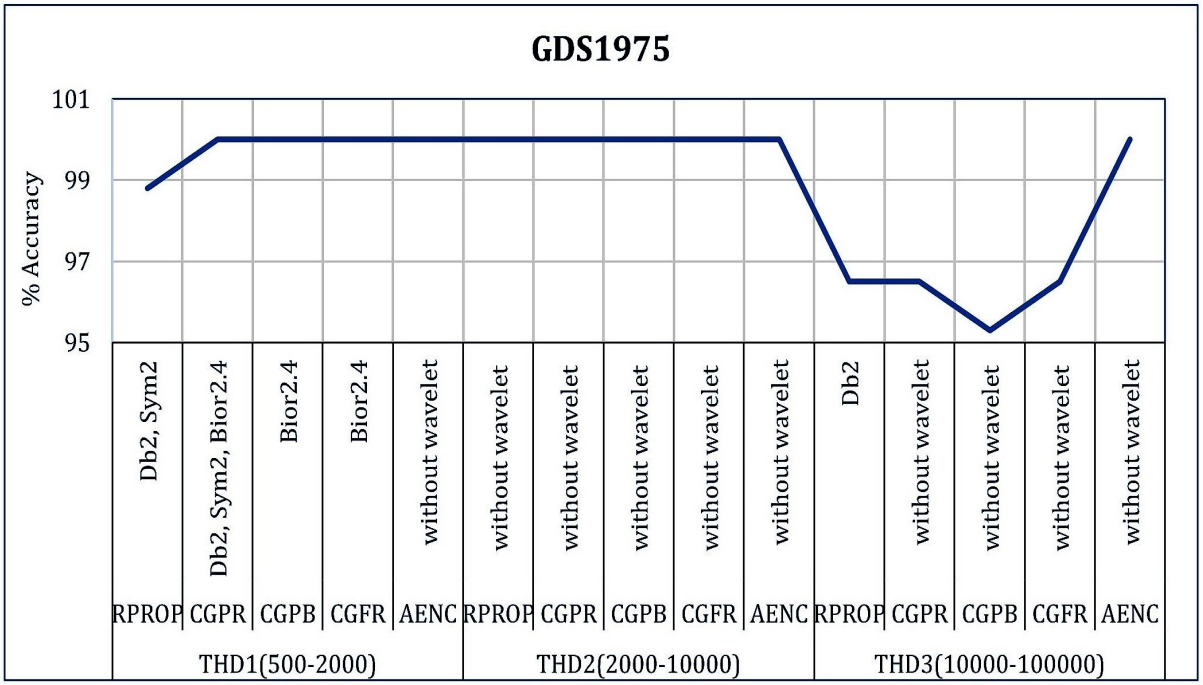


Figure 7.1: Result of Thresholding method for GDS1975 dataset.

Figure 7.2, Figure 7.3 and Figure 7.4 demonstrates the result of Thresholding method for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively.

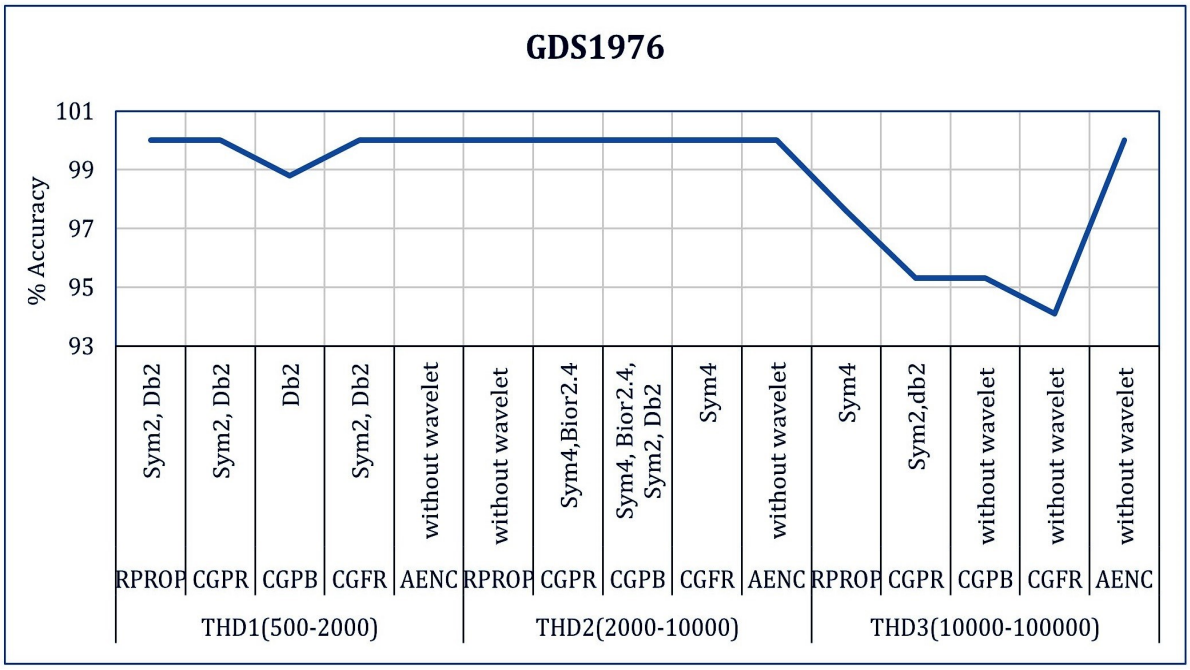


Figure 7.2: Result of Thresholding method for GDS1976 dataset.

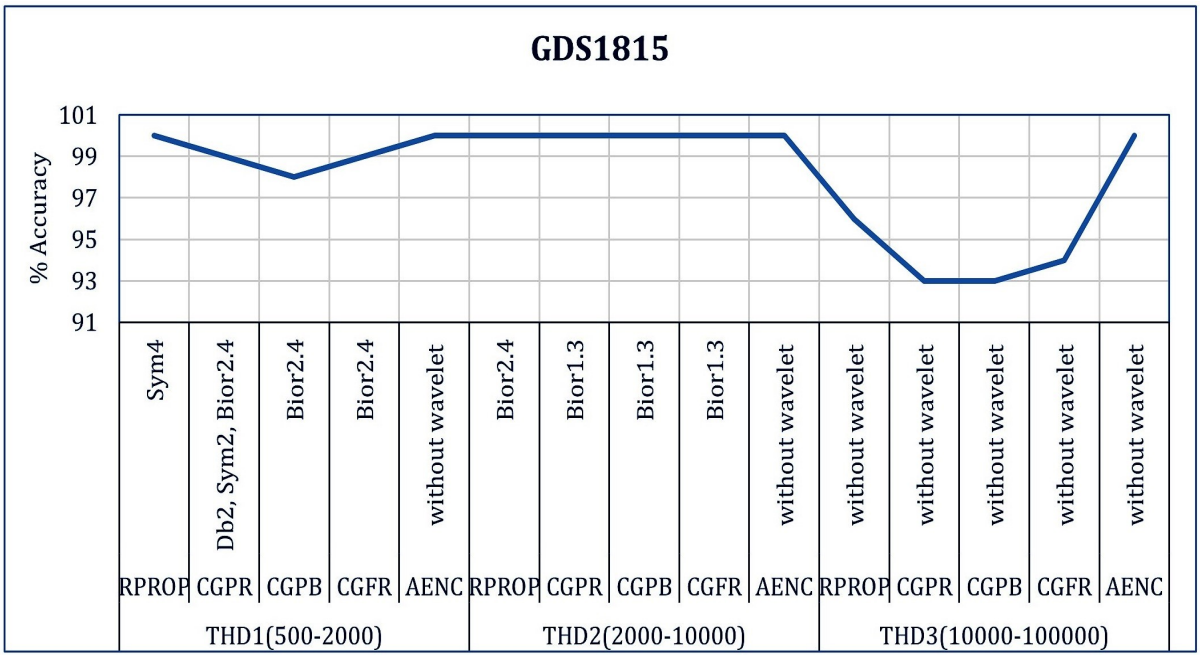


Figure 7.3: Result of Thresholding method for GDS1815 dataset.

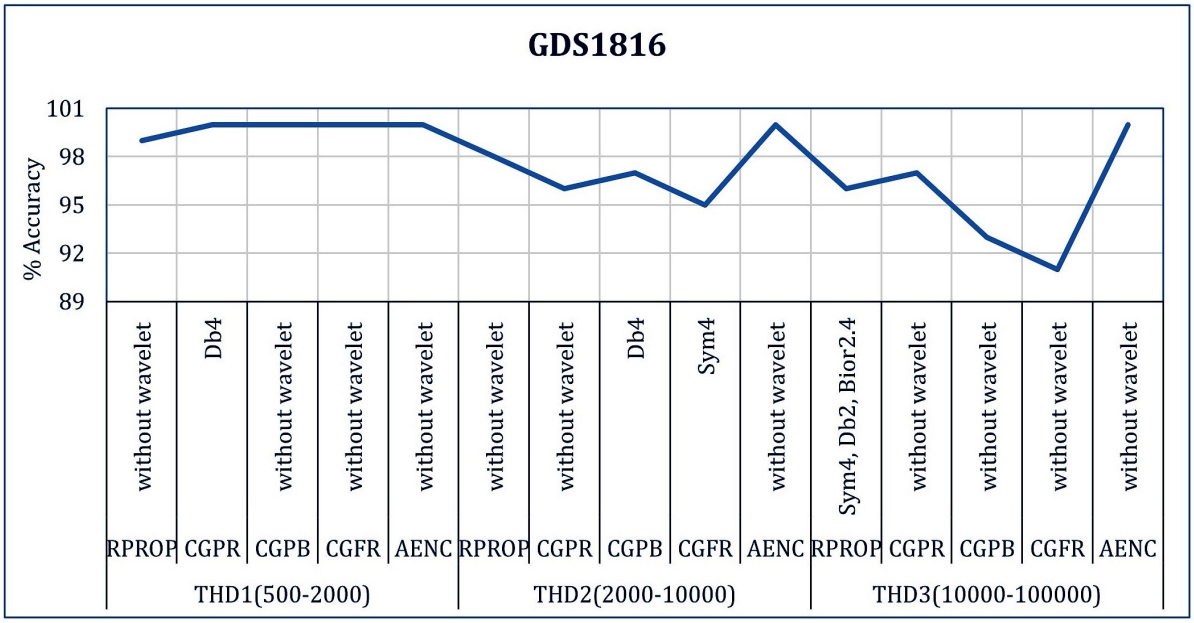


Figure 7.4: Result of Thresholding method for GDS1816 dataset.

The comparison of results of Thresholding method for GDS1975, GDS1976, GDS1815 and GDS1816 datasets is presented in Figure 7.5.

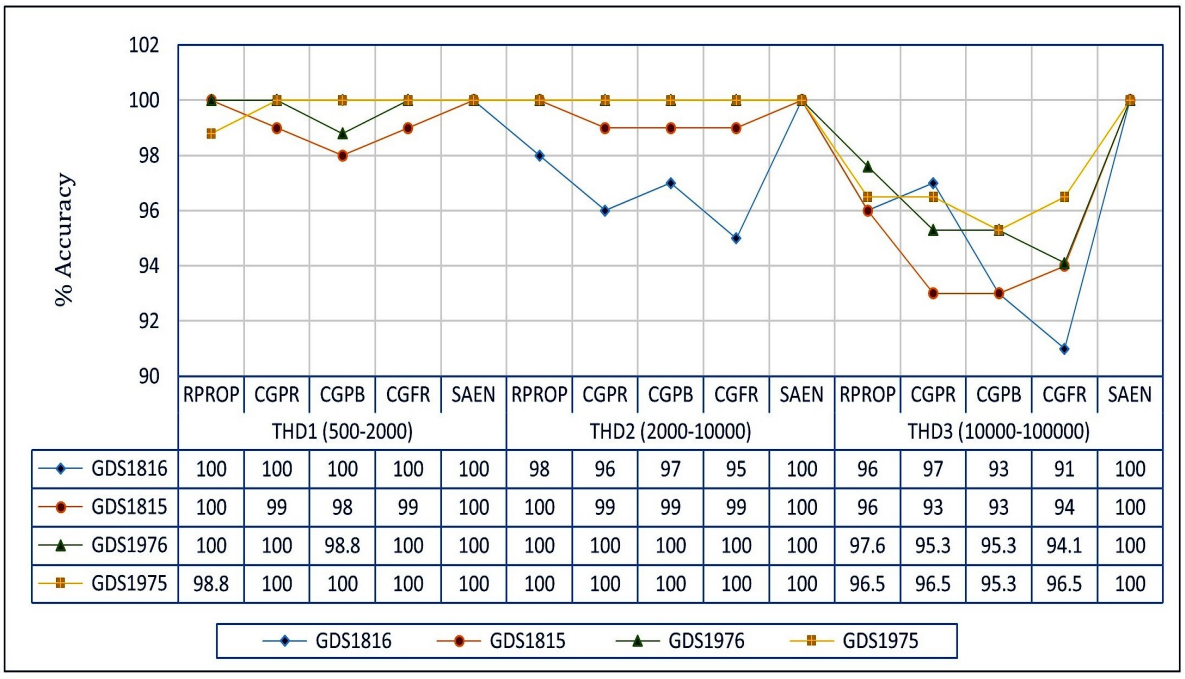


Figure 7.5: Comparison of result of Thresholding method for Glioma datasets.

From the results of Thresholding method, few observation drawn are mentioned below.

1. The threshold range THD2 (2000, 10000) gives 100% classification accuracy with/without wavelet transform for GDS1975, GDS1976 and GDS1815 dataset while, THD1 (500, 2000) performs better for GDS1816 dataset using RPROP, Conjugate Gradient and SAEN algorithm.

2. Threshold range THD1 (500, 2000) at times delivers 100% classification accuracy with wavelet transform for GDS1975, GDS1976 and GDS1815 datasets.

3. For Threshold range THD3 (10000, 100000), SAEN algorithm gives 100% classification accuracy without wavelet transform for all above mentioned datasets.

### 7.1.3 Ratio method

This section demonstrates the classification results of Ratio method for GDS1975, GDS1976, GDS1815 and GDS1816 datasets using various ratios viz., ratio <= 4, ratio <= 3.5, ratio <= 3 and ratio <= 2.5. With ratio <= 4, the number of genes selected are 2791, 1885, 877 and 717 for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively. With ratio <= 3.5, the number of genes selected are 1929, 1187, 436 and 336 for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively. With ratio <= 3, the number of genes selected are 1030, 588, 144 and 104 for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively. With ratio <= 2.5, the number of genes selected are 314, 165, 16 and 48 for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively.

Figure 7.6 demonstrates the result of Ratio method for ratio <= 4 and ratio <= 3.5 while, Figure 7.7 demonstrates the result of the Ratio method for ratio <= 3 and ratio <= 2.5, respectively.

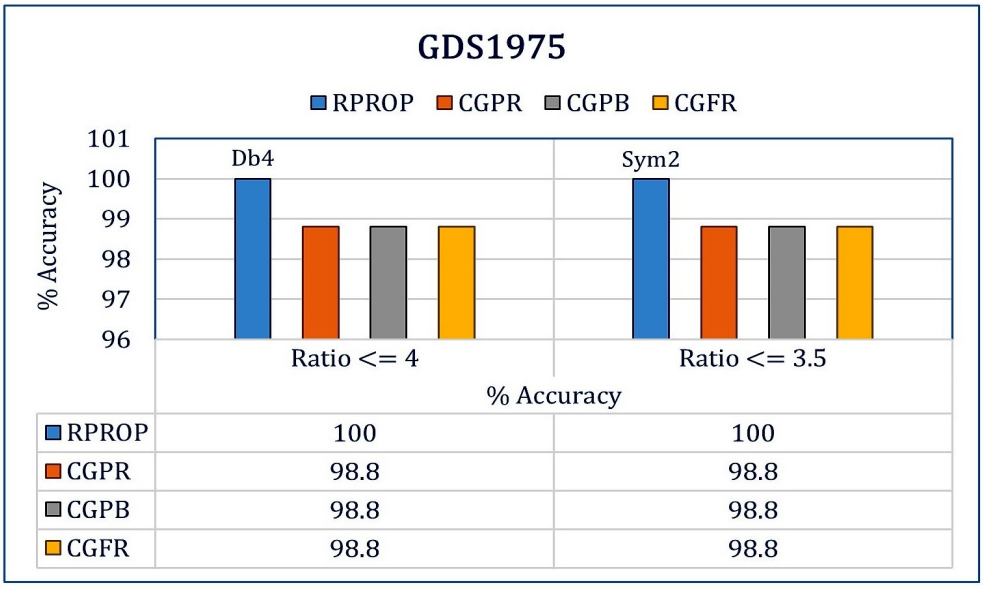


Figure 7.6: Result of Ratio method (ratio <= 4 and ratio <= 3.5 ) for GDS1975 dataset

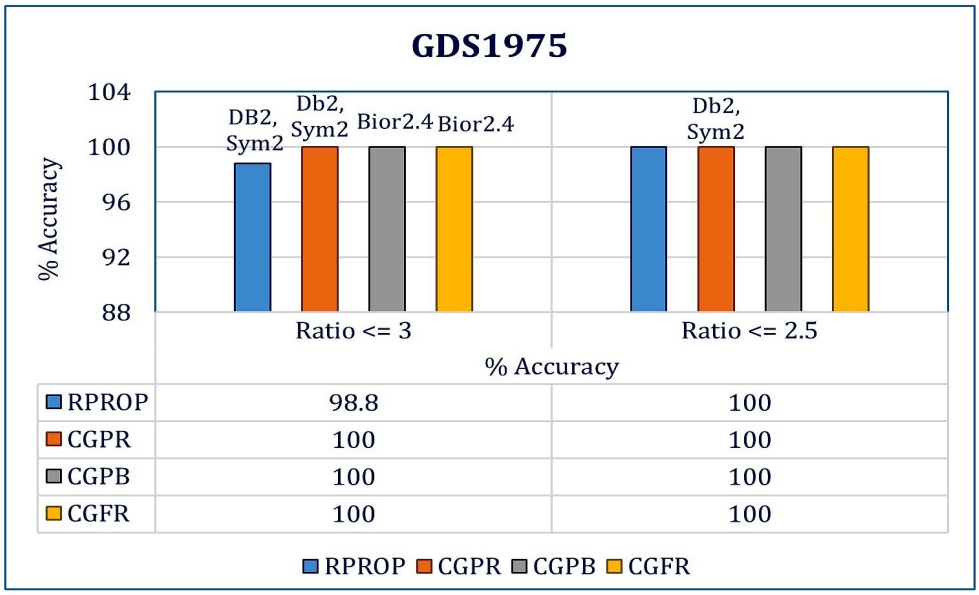


Figure 7.7: Result of Ratio method (ratio <= 3 and ratio <= 2.5) for GDS1975 dataset

For GDS1976 dataset, Figure 7.8 demonstrates the result of Ratio method for ratio <= 4 and ratio <= 3.5 while, Figure 7.9 demonstrates the result of the Ratio method for ratio <= 3 and ratio <= 2.5, respectively.

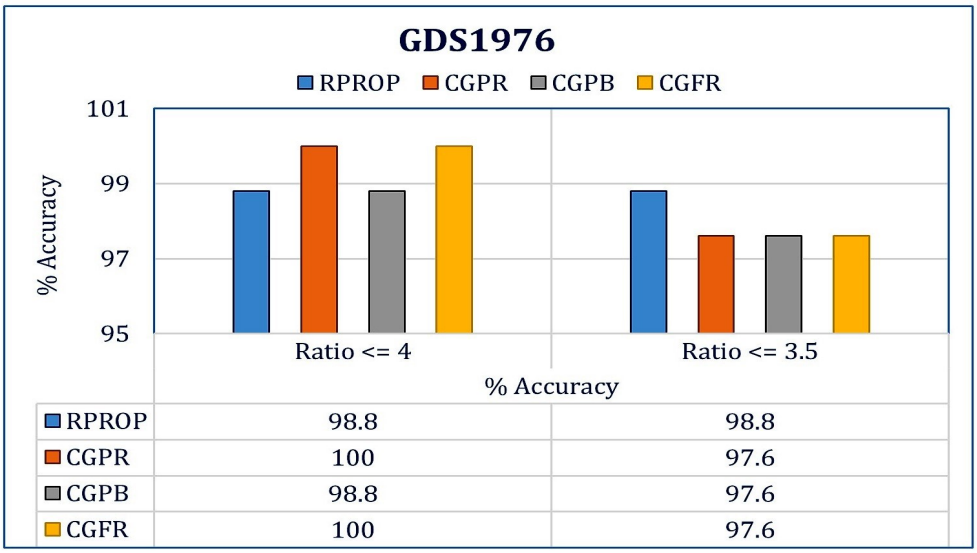


Figure 7.8: Result of Ratio method (ratio <= 4 and ratio <= 3.5) for GDS1976 dataset

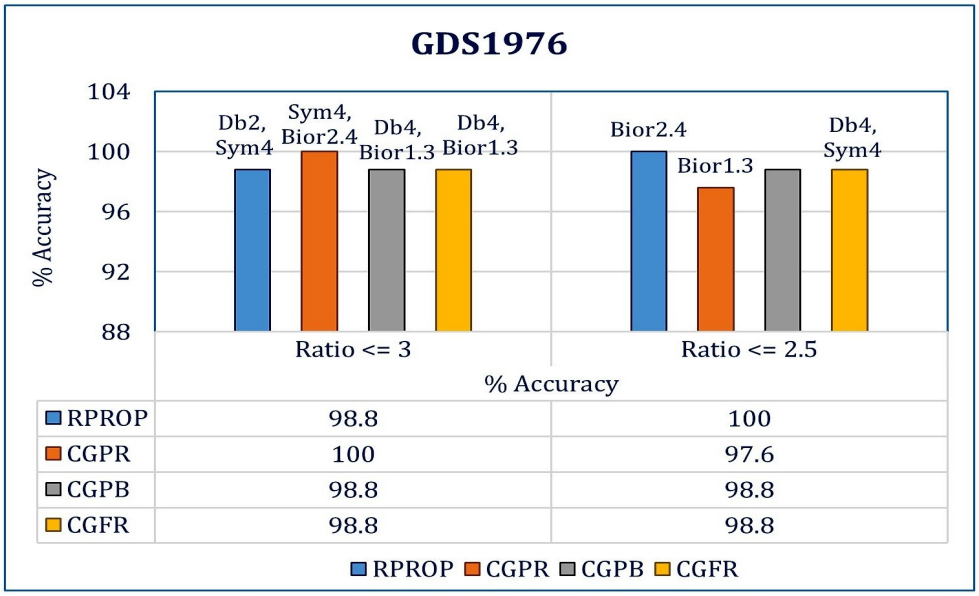


Figure 7.9: Result of Ratio method (ratio <= 3 and ratio <= 2.5) for GDS1976 dataset

For GDS1815 dataset, Figure 7.10 demonstrates the result of Ratio method for ratio <= 4 and ratio <= 3.5 while, Figure 7.11 demonstrates the result of the Ratio method for ratio <= 3 and ratio <= 2.5, respectively.

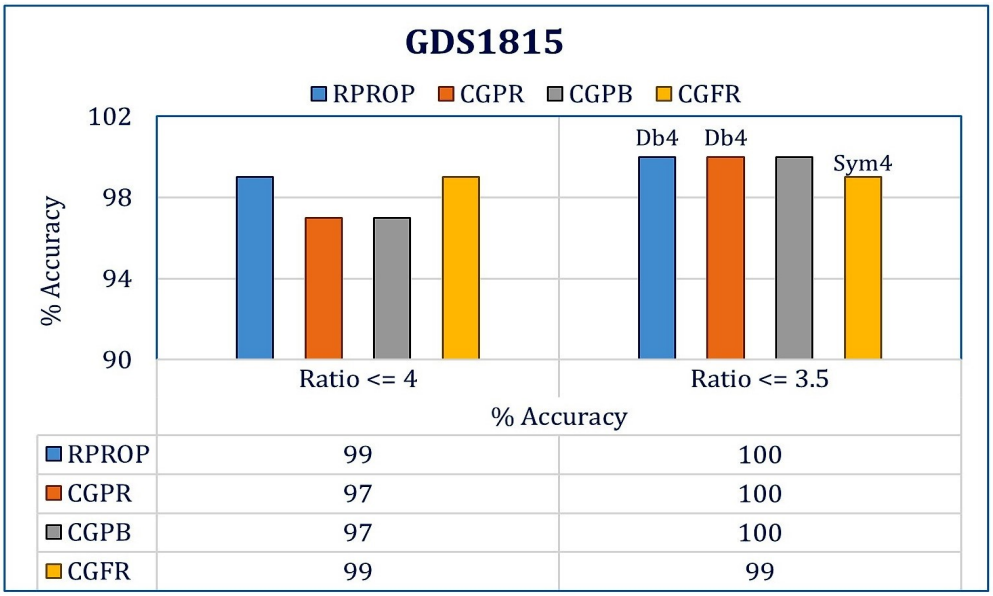


Figure 7.10: Result of Ratio method (ratio <= 4 and ratio <= 3.5) for GDS1815 dataset

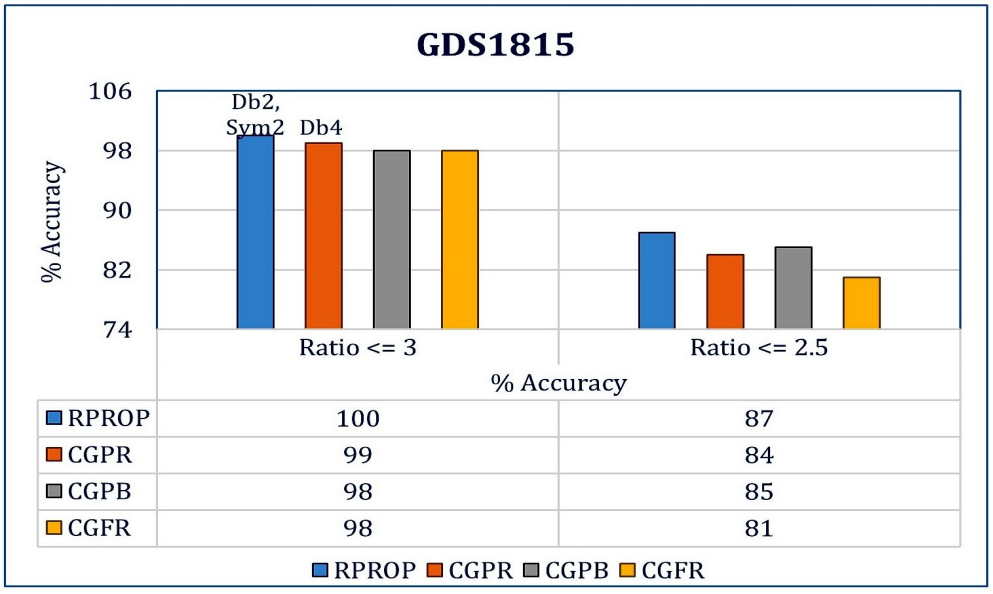


Figure 7.11: Result of Ratio method (ratio <= 3 and ratio <= 2.5) for GDS1815 dataset

For GDS1816 dataset, Figure 7.12 demonstrates the result of Ratio method for ratio <= 4 and ratio <= 3.5 while, Figure 7.13 demonstrates the result of the Ratio method for ratio <= 3 and ratio <= 2.5, respectively.

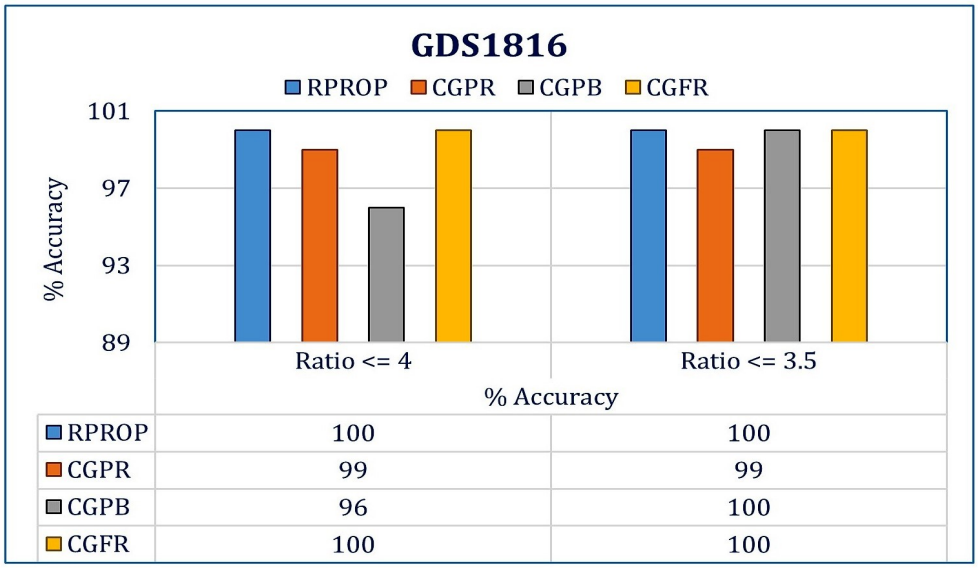


Figure 7.12: Result of Ratio method (ratio <= 4 and ratio <= 3.5) for GDS1816 dataset

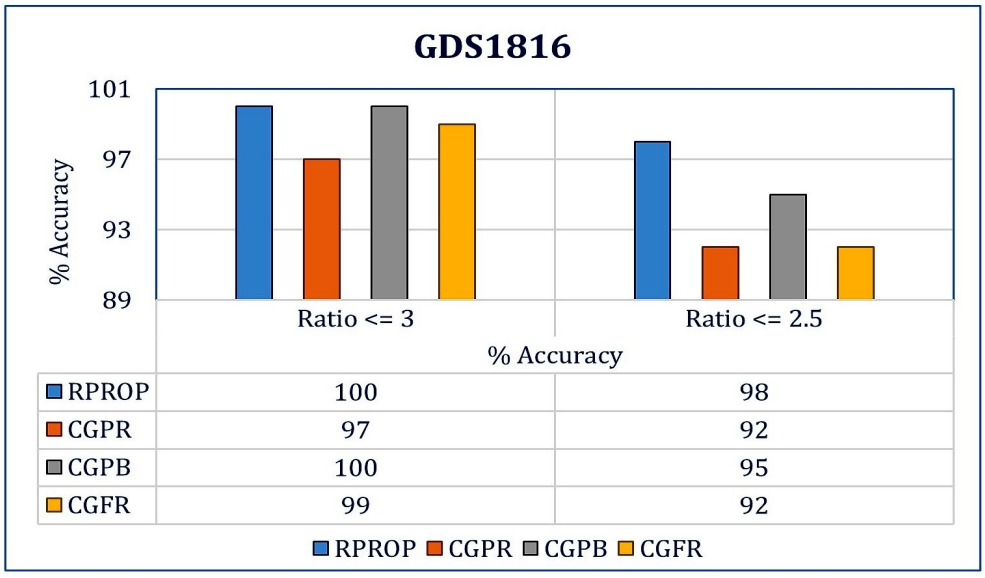


Figure 7.13: Result of Ratio method (ratio <= 3 and ratio <= 2.5) for GDS1816 dataset

The best performance of Ratio method is achieved for ratio <= 2.5, ratio <= 4, ratio <= 3.5 and ratio <= 3.5 for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively. The best of the results of Ratio method for GDS1975, GDS1976, GDS1815 and GDS1816 are shown in the Figure 7.14

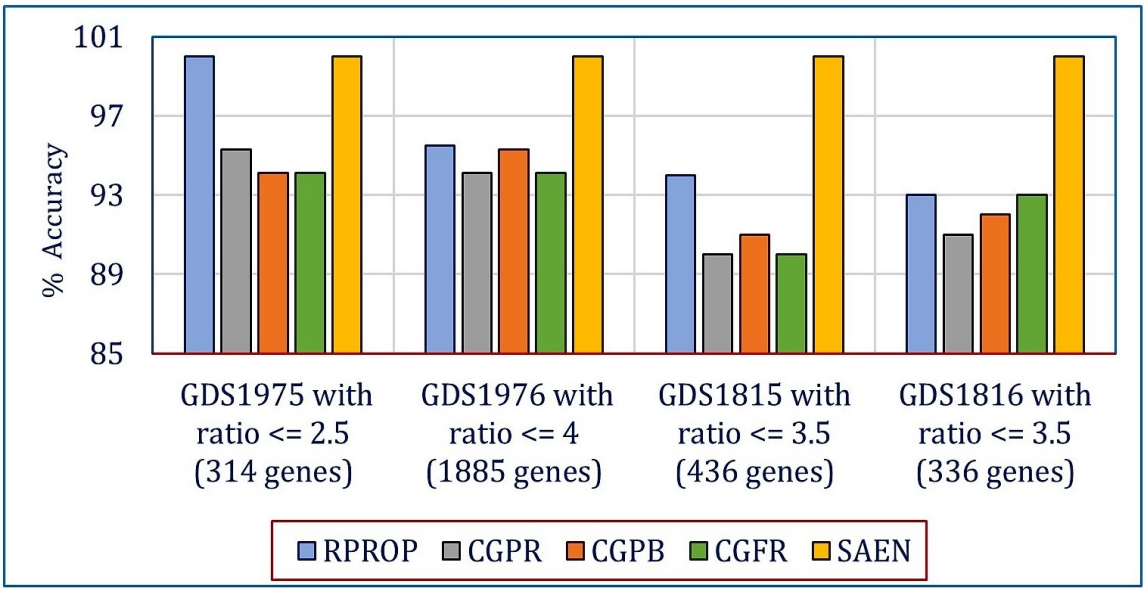


Figure 7.14: Comparison of results of Ratio method for GDS1975, GDS1976, GDS1815 and GDS1816 datasets.

### 7.1.4 Fusion of Thresholding and Ratio method

The classification results of the fusion of Thresholding and Ratio method in combination with the mean intensity difference between the two classes of Glioma datasets, GDS1975, GDS1976, GDS1815 and GDS1816 datasets are presented in this section. For GDS1975 dataset, 10 and 8 genes are selected for difference in the mean intensity values (u1-u2) of the Glioma classes 870 and 1000, respectively. For GDS1976 dataset, 15 and 5 genes are selected for u1-u2 1000 and 1500, respectively. For GDS1815 dataset, 7 and 5 genes are selected for difference in the mean intensity values of the glioma classes 800 and 1000, respectively. For GDS1816 dataset, 12 and 5 genes are selected for difference in the mean intensity values of the glioma classes 170 and 250, respectively.

Figure 7.15 and Figure 7.16 shows two of the best results of fusion of Thresholding and Ratio method for GDS1975 and GDS1976 datasets, respectively.

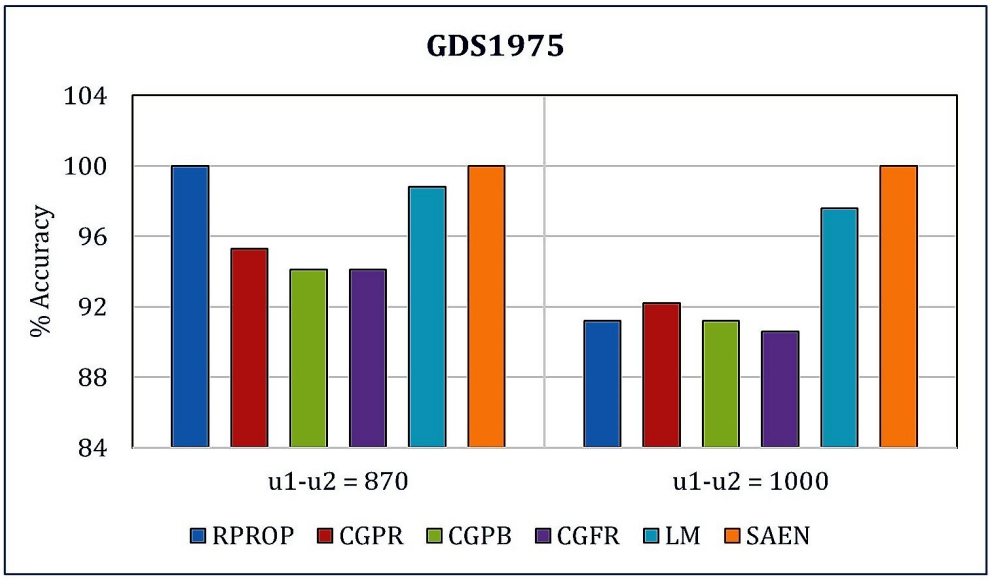


Figure 7.15: Result of Fusion of Thresholding and Ratio method for GDS1975 dataset.

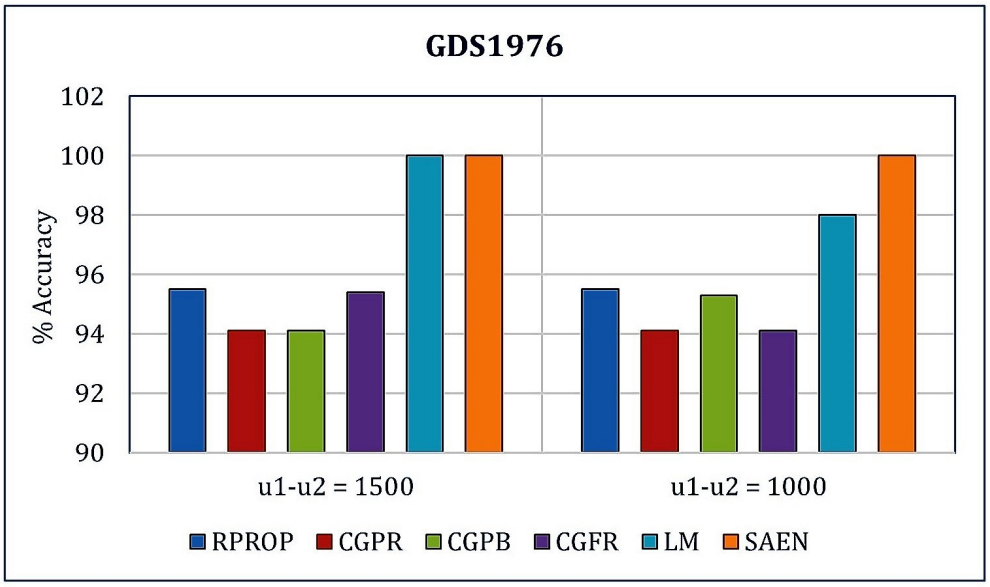


Figure 7.16: Results of Fusion of Thresholding and Ratio method for GDS1976 dataset.

Figure 7.17 and Figure 7.18 shows two of the best results of fusion of Thresholding and Ratio method for GDS1815 and GDS1816 datasets, respectively.

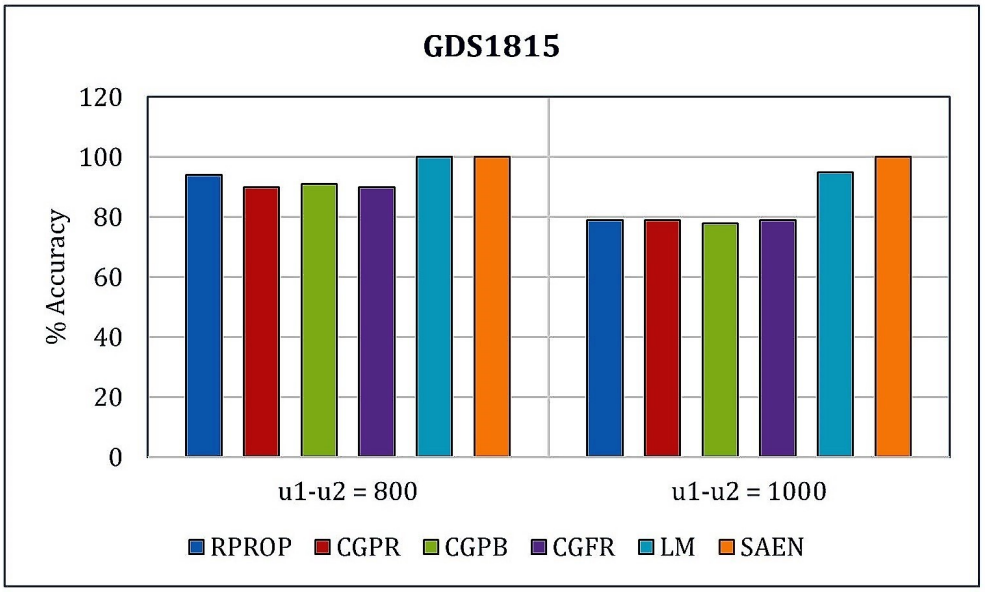


Figure 7.17: Results of Fusion of Thresholding and Ratio method for GDS1815 dataset.

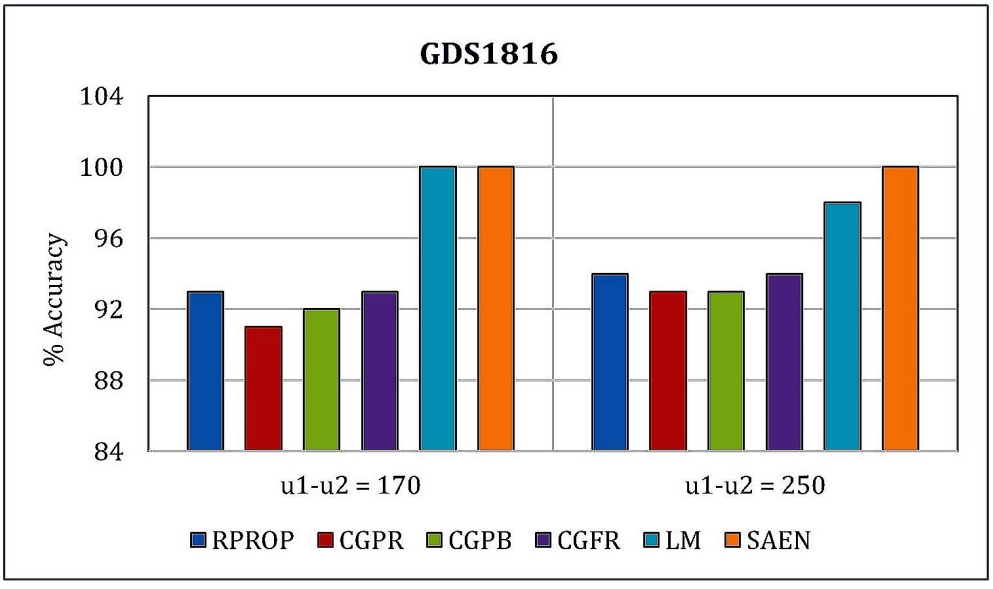


Figure 7.18: Results Fusion of Thresholding and Ratio method for GDS1816 dataset.

Figure 7.19 demonstrates the comparison of best of the results of Fusion of Thresholding and Ratio method for GDS1975, GDS1976, GDS1815 and GDS1816 datasets.

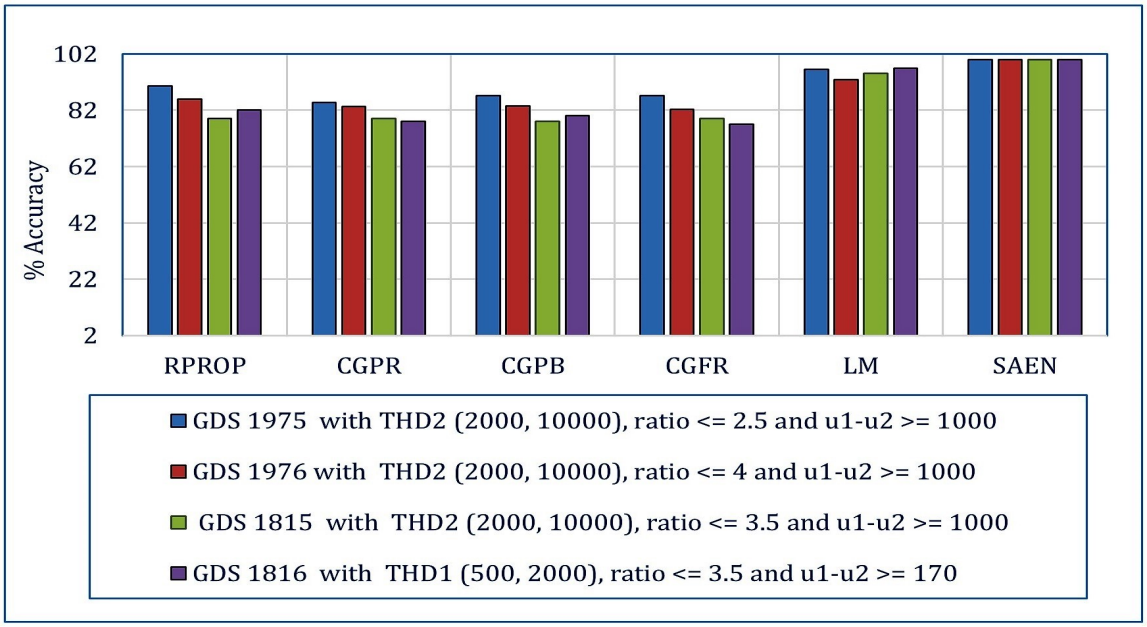


Figure 7.19: Results of Fusion of Thresholding and Ratio method for GDS1975, GDS1976, GDS1815 and GDS1816 datasets.

The comparative analysis of four datasets in proposed study utilizes commonly transcribed genes, such as Protein Kinase B3 (PKB3), Mortality Factor 4 Like 2 (MORF4L2), Ankyrin Repeat Domain 17 (ANKRD17), Signal Recognition Particle 14 (SRP14) and Zinc Finger Protein (ZNF550). PKB3 gene coding for serine/threonine protein kinase is involved in cell proliferation, differentiation and apoptosis. Further, PKB3 gene expression gets down regulated from grade III to grade IV [105]. It may be noted that, MORF4L2 is a vital component of NuA4 HAT and has significant role in transcriptional activation of several genes including oncogenes and proto-oncogenes [106]. An alteration in the gene expression of ANKRD17 observed from glioma grade III to grade IV may be attributed to G1/S transition [107]. SRP14 along with SRP9 and Alu RNA constitute elongation arrest domain signal recognition particle and plays a crucial role in targeting secretary protein to endoplasmic reticulum. Down regulation of SRP14 would alter signal recognition particle mediated vernacular protein transport system leading to cancer progression [108]. The uniport

KB database has reviewed and annotated ZNF550 to be involved in transcriptional regulation. An Alteration in ZNF550 expression may lead to remodeling in expression pattern of cancer related genes promoting oncogenesis. The common transcriptions among four datasets and related functions of these genes leads to direct or indirect correlation of mutations in the above genes with the development of glioma grade III and IV.

The comparison of the best of results of fusion of Thresholding and Ratio method obtained using five common genes across Glioma datasets GDS1975, GDS1976, GDS1815 and GDS1816 is demonstrated in Figure 7.20.

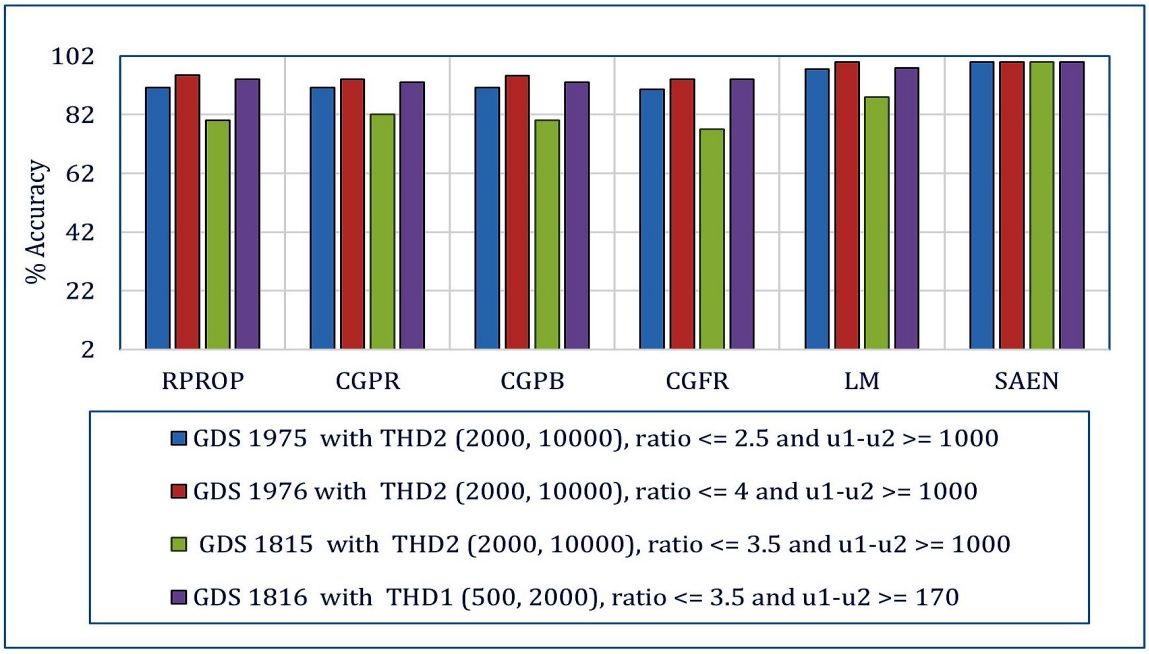


Figure 7.20: Results of Fusion of Thresholding and Ratio method for GDS1975, GDS1976, GDS1815 and GDS1816 datasets.

### 7.1.5 Comparison of proposed system with existing system

Cancer classifications reported in the literature vary widely in respect of microarray datasets as well as methods employed to measure parameters defining and evaluating

types of cancers. Therefore, classification accuracy and optimum number of genes obtained are compared with the results of the authors Abusamra H et al. [23] and Shen Q et al. [16], who employed some of the same dataset as ours. The comparative results of the proposed system with the existing methods are presented in Figure 7.21.

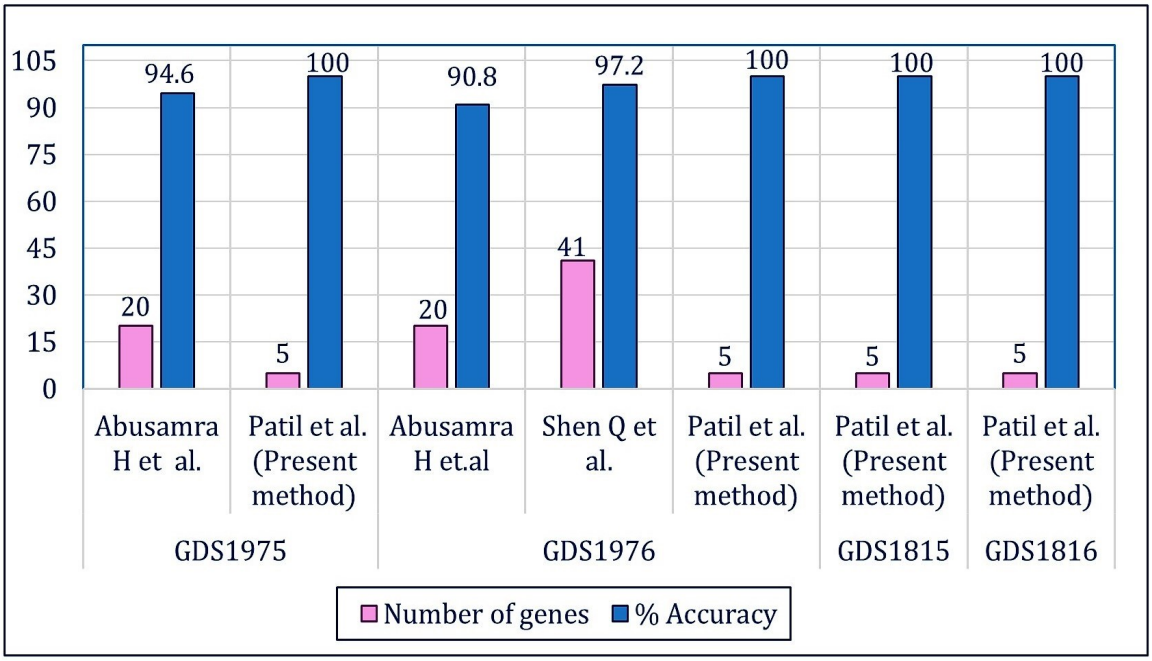


Figure 7.21: Comparative results of the proposed system with existing systems for GDS1975 and GDS1976 datasets.

The proposed system delivers higher classification accuracy with less number of genes as compared to method implemented by Abusamra H et al. [23] and Shen Q et al. [16] for GDS1975 and GDS1976 datasets.

In the proposed system, the classification of GDS1975, GDS1976, GDS1815 and GDS1816 datasets is implemented using Intel(r) Core(TM) i3 CPU M380 @2.53 GHz processor and MATLAB R2017a software.

Table 7.9 presents the comparison of computational time of proposed method with to method suggested by Abusamra H et al. [23] and Shen Q et al. [16] for GDS1975

and GDS1976 datasets.

Table 7.9: Comparison of computational time of proposed system with existing systems

| Sr. No. | Method | Computational Time (Sec.) |
|---|---|---|
| 1 | Abusamra H et al. | 3 |
| 2 | Shen Q et al. | 258 |
| 3 | Proposed System | 17 |

The computational time of the proposed system is moderate as compared to method suggested by Abusamra H et al. [23] and Shen Q et al. [16] for GDS1975 and GDS1976 datasets.

### 7.1.6 Testing of optimal gene subset for GDS1962 dataset.

The effectiveness of optimal gene subset chosen using the fusion of Thresholding and Ratio method is tested for GDS1962 dataset at every level of malignancy using SAEN algorithm. Classification accuracy of 100 % is obtained by using the optimal gene subset at every level of malignancy of GDS1962 dataset as illustrated in Table 7.10.

Table 7.10: Result of classification for GDS1962 dataset using genes from optimal gene Subset.

| Sr. No. | Brain tumor Classes | Accuracy. |
|---|---|---|
| 1 | Benign and Malignant | 100% |
| 2 | Lymphoma and Glioma | 100% |
| 3 | Low Oligodendroglioma, High Oligodendroglioma and Astrocytoma | 100% |
| 4 | Astrocytoma Grade II, Grade III and Grade IV | 100% |

## 7.2 Conclusion and Future scope

### 7.2.1 Conclusion

An alarming increase in cancer deaths every year, existence of more complex methods for cancer classification and occasional failure of biomarkers to identify the cancer type, makes it essential to design an efficient cancer classification system so as to increase the survival rate of cancer patients. In this thesis, a simple and computationally less expensive cancer classification system is designed to obtain 100% classification accuracy for less number of genes using gene expression data obtained by using Microarray technology. Microarray technology is a proven tool for global analysis of gene expression that allows simultaneous investigation of thousands of genes in a sample. The proposed method is implemented for microarray Glioma datasets namely, GDS1975, GDS1976, GDS1815 and GDS1815 and tested for GDS1962 dataset.

Initially, the classification of Brain tumor at different level of malignancy is implemented for GDS1962 dataset with and without using feature extraction (DCT, DWT) method in combination with RPROP and Conjugate Gradient algorithms. The conclusions drawn from these results are discussed below

1. The divergently expressed genes of Benign, Malignant and Glioma, Lymphoma Brain tumors (GDS1962 dataset) makes the classification process effortless and deliver 100% classification accuracy.

2. However, the results of classification of Low Oligodendroglioma, High Oligodendroglioma, Astrocytoma and Astrocytoma Grade II, Grade III, Grade IV (GDS1962 dataset) illustrates the difficulty of attaining 100% classification accuracy (with and without DCT, DWT) on account of less distinctly expressed genes at higher level of malignancy.

3. In absence of feature extraction method, the huge size of the input of the classifier which mostly includes redundant data tends to slow down the classification process.

Therefore, in order to improve the classification accuracy at higher level of malignancy and to increase the speed of classification, cancer classification system is designed using feature selection methods, feature extraction methods and neural network classifiers. The feature selection is implemented using Thresholding method, Ratio method and Fusion of Thresholding and Ratio method. The feature extraction is implemented using DWT while the classification is implemented using RPROP, Conjugate Gradient, LM and SAEN algorithms. The designed cancer classification system is implemented for GDS1975, GDS1976, GDS1815, GDS1816 Glioma datasets and tested for GDS1962 Brain tumor dataset.

The conclusions drawn from the results of implementation are given below

1. Thresholding method

   (a) Thresholding method excludes genes with inconsistent intensity variation across the Glioma samples.

   (b) The number and the range of thresholds needs to be decided based on the intensity variation in a particular cancer dataset. As Glioma datasets in the proposed study are obtained from 16 bit microarray image and majority of the intensities lie below 10000, the thresholds are chosen as THD1 (500-2000), THD2 (2000-10000) and THD3 (10000-100000). The gene intensity values below 500 are not reliable owing likely cross hybridization of genes.

   (c) The threshold range that gives best performance depends on the gene intensity variation for different classes of the dataset. The threshold range THD2 (2000, 10000) delivers 100% classification accuracy with/without wavelet transform for GDS1975, GDS1976 and GDS1815 dataset while, THD1 (500, 2000) performs better for GDS1816 dataset.

   (d) On account of less differentially expressed genes in the Threshold range THD1 (500, 2000) of GDS1975, GDS1976 and GDS1815 datasets, occasionally 100% classification accuracy is obtained with wavelet transform.

(e) THD3 (10000, 100000) contains lesser number of genes with higher values of intensity. Since intensity values are very large, a small change often makes the task of classification difficult. However, SAEN algorithm, delivers 100% classification accuracy without wavelet transform for THD3 (10000, 100000) of all the above mentioned datasets.

(f) Thresholding method appears to have an edge, in the sense, it provides an alternative subset of genes for obtaining 100% classification accuracy.

2. Ratio method

(a) Ratio method eliminates genes with large maximum to minimum intensity ratio across the dataset samples of a particular class.

(b) The results of Ratio method depends on the difference in the mean intensity values of classes of Glioma. The ratios of maximum to minimum gene intensity considered for a classification of Glioma are ratio <= 4, ratio <= 3.5, ratio <= 3 and ratio <= 2.5. The best performance of Ratio method is achieved for ratio <= 2.5, ratio <= 4, ratio <= 3.5 and ratio <= 3.5 for GDS1975, GDS1976, GDS1815 and GDS1816 datasets, respectively.

3. Fusion of Thresholding and Ratio method

(a) Fusion of Thresholding method and Ratio method gives a small subset of genes in comparison with Thresholding method and Ratio method, implemented independently.

(b) In this method, genes common to best performing thresholding and ratio are mined from the Glioma dataset and classification is performed. Further, filtering of genes on the basis of difference in the average gene intensity with or without wavelet transform leads to 100% classification accuracy with less number of genes.

(c) Alternatively, ratio can be chosen first and thresholding can be applied later, yielding the same subset of genes.

4. Feature extraction

   (a) The custom made filters of DWT, ability to detect the discontinuity in the signal, various types of mother wavelet functions makes it most efficient for feature extraction.

   (b) There is no universal function that works well for all microarray datasets. A blend of the type of wavelet and neural network algorithm that gives the best result rely on the nature of variation of classification data and network parameters. Based on the classification accuracy, the most suitable wavelet for the datasets under consideration are found to be Db2 and Sym2.

5. Classification algorithms

   (a) Conjugate Gradient algorithms are designed to restart in case of failure to reach the convergence wherein, the number of neural network parameters become equal to the number of iterations. Hence, Conjugate Gradient algorithms are found to be faster than RPROP, LM algorithms.

   (b) For larger data size, the prerequisite of extensive memory in LM algorithm makes the classification process inefficient and sluggish. On account of less number of genes obtained by the Fusion of Thresholding and Ratio method combined with the mean intensity difference between genes of different classes, LM algorithm performed better in comparison with Conjugate Gradient algorithms.

   (c) Stacked Autoencoder network trained with Back Propagation algorithm delivers the best result as compared to RPROP, Conjugate Gradient algorithms and LM algorithm owing to the prior training of Autoencoder stages, fine tuning of Softmax layer and Stacked Autoencoder network.

6. Optimal gene subset

   (a) Optimal gene subset obtained using the Fusion of Thresholding and Ratio

Method comprises of genes viz., PKB3, MORF4L2, ANKRD17, SRP14 and ZNF550.

(b) The mutations in the genes selected by the Fusion of Thresholding and Ratio method are directly or indirectly associated to the occurrence of Glioma Grade III and Grade IV.

(c) Testing of this optimal gene subset for GDS1962 at different level of malignancies gives 100% classification accuracy.

7. Comparison with existing systems

(a) The implementation of the proposed system with Intel(r) Core(TM) i3 CPU M380 @2.53 GHz processor and MATLAB R2017a software requires moderate computational time of about 17 sec as compared to 3 sec and 258 sec using the methods suggested by Abusamra H et al. [23] and Shen Q et al. [16], respectively. Considering the state of art, the computational time appears to be insignificant.

(b) The proposed system uses simple and computationally less expensive feature selection method.

(c) The SAEN network along with a combination of Thresholding and Ratio method outperforms the methods suggested by Abusamra H et al. [23] and Shen Q et al. [16] giving 100% classification accuracy using only five common genes for GDS1975, GDS1976, GDS1815 and GDS1816 datasets.

(d) Proposed system facilitates the easy selection and precise classification of Cancer at higher malignancy level with highest accuracy in moderate time as compared to the existing methods.

### 7.2.2 Future scope

1. The gene expression data from the different resources, experiments does not follow standard format. Gene expression data is available in various forms such

as intensity values, ratio of red to green intensity of genes, logarithm of the ratio of red to green intensity of genes etc. The proposed system is designed for classification of Glioma using gene intensities obtained from 16 bit microarray image. It can be modified to be applicable to all forms of gene expression data. It can be extended to few of such kind of datasets namely, Colon cancer, ALL/AML, DLBCL, Prostate cancer, Leukemia, Breast cancer etc.

2. The gene selection methods utilized in proposed study do not consider interdependency between the genes. These methods do not interact with the classifier. Therefore, the proposed system may be modified to consider the interdependency between the genes and to automatically interact with the classifier.

3. At times, these gene selection methods may select the genes, whose mutations may not be directly or indirectly related to the occurrence of Grade III and Grade IV Glioma. It can be modified to select the genes whose mutations are directly related with the occurrence of Grade III and Grade IV Glioma.

4. As yet, no method of cancer classification gives the stable subset of genes for the classification of Grade III and Grade IV Glioma. The proposed method may be modified to get more stable gene subset to be used for cancer classification.

# Bibliography


[1] Cancer, url=http://www.who.int/mediacentre/factsheets/fs297/en/.

[2] Cancer, url=http://learn.genetics.utah.edu/content/history/cancer/.

[3] Lixin Sun, Ai-Min Hui, Qin Su, Alexander Vortmeyer, Yuri Kotliarov, Sandra Pastorino, Antonino Passaniti, Jayant Menon, Jennifer Walling, Rolando Bailey, et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer cell*, 9:287–300, 2006.

[4] William A Freije, F Edmundo Castro-Vargas, Zixing Fang, and Horvath et al. Gene expression profiling of gliomas strongly predicts survival. *Cancer research*, 64:6503–6510, 2004.

[5] Genome resource facility, url = http://grf.lshtm.ac.uk/microarrayoverview.html/.

[6] Jaleh Barar, Amir Ata Saei, and Yadollah Omidi. In *Gene Therapy - Developments and Future Perspectives*. InTech, 2011.

[7] Vittal R Srinivas. *Bioinformatics: A Modern Approach*. PHI Learning Pvt. Ltd., 2005.

[8] M Madan Babu. Introduction to microarray data analysis. *Computational genomics: Theory and application*, 17:225–49, 2004.

[9] Thermo fisher scientific, url = http://www.affymetrix.com/.

[10] Y Tu, G Stolovitzky, and U Klein. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences*, 99:14031–14036, 2002.

[11] Ron Dror. Noise models in gene array analysis. *Report in fulfillment of the area exam requirement in the MIT Department of Electrical Engineering and Computer Science*, 2001.

[12] Peter Bajcsy, Lei Liu, and Mark Band. Dna microarray image processing. *DNA Array Image Anal. Nuts Bolts*, pages 1–77, 2007.

[13] Shuanhu Wu and Hong Yan. Microarray image processing based on clustering and morphological analysis. In *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics*, volume 19.

[14] Guifang Shao, Tiejun Li, Wangda Zuo, Shunxiang Wu, and Tundong Liu. A combinational clustering based method for cdna microarray image segmentation. *PloS one*, 10:e0133025, 2015.

[15] Ji-Gang Zhang and Hong-Wen Deng. Gene selection for classification of microarray data based on the bayes error. *BMC bioinformatics*, 8:370, 2007.

[16] Qi Shen, Zhen Mei, and Bao-Xian Ye. Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification. *Computers in biology and medicine*, 39:646–649, 2009.

[17] Debahuti Mishra and Barnali Sahu. Feature selection for cancer classification: a signal-to-noise ratio approach. *International Journal of Scientific & Engineering Research*, 2:1–7, 2011.

[18] B Chandra and Manish Gupta. An efficient statistical feature selection approach for classification of gene expression data. *Journal of biomedical informatics*, 44:529–535, 2011.

[19] Alok Sharma, Seiya Imoto, and Satoru Miyano. A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9:754–764, 2012.

[20] Monalisa Mandal and Anirban Mukhopadhyay. An improved minimum redundancy maximum relevance approach for feature selection in gene expression data. *Procedia Technology*, 10:20–27, 2013.

[21] Mohsen Hajiloo, Babak Damavandi, Metanat HooshSadat, Farzad Sangi, John R Mackey, Carol E Cass, Russell Greiner, and Sambasivarao Damaraju. Breast cancer prediction using genome wide single nucleotide polymorphism data. *BMC bioinformatics*, 14:S3, 2013.

[22] Jagath C Rajapakse and Piyushkumar A Mundra. Multiclass gene selection using pareto-fronts. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10:87–97, 2013.

[23] Heba Abusamra. A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Computer Science*, 23:5–14, 2013.

[24] M Shoman S Tarek, R Elwahab. Gene expression based cancer classification. *Egyptian Informatics J*, 18:151–159, 2016.

[25] J Wu W Zhang X Lu. Feature selection for cancer classification using microarray gene expression data. *Biostat Biometrics Open Acc J*, 1:555–557.

[26] Lei Yu, Yue Han, and Michael E Berens. Stable gene selection from microarray data via sample weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9:262–272, 2012.

[27] Quanjin Liu, Zhimin Zhao, Ying-xin Li, Xiaolei Yu, and Yong Wang. A novel method of feature selection based on svm. *JCP*, 8:2144–2149, 2013.

[28] Satoshi Niijima and Yasushi Okuno. Laplacian linear discriminant analysis approach to unsupervised feature selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:605–614, 2009.

[29] Sebastián Maldonado, Richard Weber, and Jayanta Basak. Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, 181:115–128, 2011.

[30] Ali Anaissi, Paul J Kennedy, Madhu Goyal, and Daniel R Catchpoole. A balanced iterative random forest for gene selection from microarray data. *BMC bioinformatics*, 14:261, 2013.

[31] Yukyee Leung and Yeungsam Hung. A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7:108–117, 2010.

[32] Chien-Pang Lee and Yungho Leu. A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing*, 11:208–213, 2011.

[33] Mohsen Hajiloo, Hamid R Rabiee, and Mahdi Anooshahpour. Fuzzy support vector machine: an efficient rule-based classification technique for microarrays. *BMC bioinformatics*, 14:S4, 2013.

[34] Pengyi Yang, Bing B Zhou, Zili Zhang, and Albert Y Zomaya. A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data. *BMC bioinformatics*, 11:S5, 2010.

[35] Huawen Liu, Lei Liu, and Huijie Zhang. Ensemble gene selection by grouping for microarray data classification. *Journal of biomedical informatics*, 43:81–87, 2010.

[36] Shutao Li, Chen Liao, and James T Kwok. Wavelet-based feature extraction for microarray data classification. In *International Joint Conference on Neural Networks*, pages 5028–5033. IEEE, 2006.

[37] Rosalin Mahapatra, Babita Majhi, and Minakhi Rout. Development and performance evaluation of improved classifiers of microarray data. In *International Conference on Advances in Engineering, Science and Management*, pages 519–523. IEEE, 2012.

[38] M Vimaladevi and B Kalaavathi. A microarray gene expression data classification using hybrid back propagation neural network. *Genetika*, 46:1013–1026, 2014.

[39] Zarita Zainuddin and Ong Pauline. Improved wavelet neural network for early diagnosis of cancer patients using microarray gene expression data. In *International Joint Conference on Neural Networks*, pages 3485–3492. IEEE, 2009.

[40] Sabrina Rashid and Golam Morshed Maruf. An adaptive feature reduction algorithm for cancer classification using wavelet decomposition of serum proteomic and dna microarray data. In *IEEE International Conference Bioinformatics and Biomedicine Workshops*, pages 305–312. IEEE, 2011.

[41] Jaison Bennet, Chilambuchelvan Arul Ganaprakasam, and Kannan Arputharaj. A discrete wavelet based feature extraction and hybrid classification technique for microarray data analysis. *The Scientific World Journal*, 2014.

[42] Tumor markers, url = https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-markers-fact-sheet.

[43] Jun Chin Ang, Andri Mirzal, Habibollah Haron, and Haza Nuzly Abdull Hamed. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, 13:971–989, 2016.

[44] Heidi S Phillips, Samir Kharbanda, Ruihuan Chen, and Forrest et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer cell*, 9:157–173, 2006.

[45] Bruno M Costa and Justin S et al. Smith. Reversing hoxa9 oncogene activation by pi3k inhibition: epigenetic mechanism and prognostic significance in human glioblastoma. *Cancer research*, 70:453–462, 2010.

[46] Xin-Yun Zhang, Fei Chen, Yuan-Ting Zhang, Shannon C Agner, Metin Akay, Zu-Hong Lu, Mary Miu Yee Waye, and SK-W Tsui. Signal processing techniques in genomic engineering. *Proceedings of the IEEE*, 90:1822–1833, 2002.

[47] Mario Mastriani and Alberto E Giraldez. Microarrays denoising via smoothing of coefficients in wavelet domain. *International Journal of Biomedical Sciences*, 1:7–14, 2006.

[48] XH Wang, Robert SH Istepanian, and Yong Hua Song. Microarray image enhancement by denoising using stationary wavelet transform. *IEEE Transactions on Nanobioscience*, 2:184–189, 2003.

[49] Rastislav Lukac, Konstantinos N Plataniotis, Bogdan Smolka, and Anastasios N Venetsanopoulos. A multichannel order-statistic technique for cdna microarray image processing. *IEEE Transactions on Nanobioscience*, 3:272–285, 2004.

[50] Rastislav Lukac and Bogdan Smołka. Application of the adaptive center-weighted vector median framework for the enhancement of cdna microarray images. *International Journal of Applied Mathematics and Computer Science*, 13:369–383, 2003.

[51] Paul O'Neill, George D Magoulas, and Xiaohui Liu. Improved processing of microarray data using image reconstruction techniques. *IEEE transactions on nanobioscience*, 2:176–183, 2003.

[52] Ahmad M Sarhan. Cancer classification based on microarray gene expression data using dct and ann. *Journal of Theoretical & Applied Information Technology*, 6, 2009.

[53] Pardeep Kumar, Vivek Sehgal, Durg Singh Chauhan, et al. Performance evaluation of decision tree versus artificial neural network based classifiers in diversity of datasets. In *World Congress on Information and Communication Technologies*, pages 798–803. IEEE, 2011.

[54] Rafael C Gonzalez and Richard E Woods. Digital image processing, 2005.

[55] Stéphane Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

[56] KP Soman. *Insight into wavelets: From theory to practice*. PHI Learning Pvt. Ltd., 2010.

[57] Matlab neural network toolbox, 2016.

[58] Descriptions of cybertory downloads, url = www.cybertory.org/downloads/.

[59] Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111–135, 2014.

[60] Rabia Aziz, CK Verma, and Namita Srivastava. Dimension reduction methods for microarray data: a review. 2017.

[61] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 1:37, 2014.

[62] Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter, Virginie de Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowe. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9:1106–1119, 2012.

[63] Mark A Hall and Lloyd A Smith. Practical feature subset selection for machine learning. 1998.

[64] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23:2507–2517, 2007.

[65] S. Geeta S. Sasikala, S. Balamurugan. Multi filtration feature selection (mffs) to improve the discriminatory ability in clinical data set. *Applied Computing and Informatics*, 2:117–127, 2016.

[66] S Vanaja and K Ramesh Kumar. Analysis of feature selection algorithms on classification: a survey. *International Journal of Computer Applications*, 96, 2014.

[67] Jasmina Novaković. Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, 21, 2016.

[68] Zheng Zhao, Fred Morstatter, Shashvata Sharma, Salem Alelyani, Aneeth Anand, and Huan Liu. Advancing feature selection research. *ASU feature selection repository*, 1:1–28, 2010.

[69] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286:531–537, 1999.

[70] Pierre Baldi and Anthony D Long. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.

[71] L Ad, M Hj, C By, T Hatfield, and G Baldi. Improved statistical inference from dna microarray data using analysis of variance and a bayesian statistical framework. 2005.

[72] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Em-

pirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96:1151–1160, 2001.

[73] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573:83–92, 2004.

[74] Jeffrey G Thomas, James M Olson, Stephen J Tapscott, and Lue Ping Zhao. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, 11:1227–1236, 2001.

[75] Baris Senliol, Gokhan Gulgezen, Lei Yu, and Zehra Cataltepe. Fast correlation based filter (fcbf) with a different search strategy. In *23rd International Symposium on Computer and Information Sciences*, pages 1–4. IEEE, 2008.

[76] David E Goldberg et al. Genetic algorithms in search, optimization, and machine learning, 1989.

[77] Justin Doak. Cse-92-18-an evaluation of feature selection methodsand their application to computer security. *UC Davis Dept of Computer Science tech reports*, 1992.

[78] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46:389–422, 2002.

[79] Alexander T Basilevsky. *Statistical factor analysis and related methods: theory and applications*. John Wiley & Sons, 2009.

[80] Xuechuan Wang and Kuldip K Paliwal. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern recognition*, 36:2429–2439, 2003.

[81] Principal component analysis, url =ebastianraschka.com/articles/2014 pca step by step.html/.

[82] Syed Ali Khayam. The discrete cosine transform (dct): theory and application. *Michigan State University*, 114, 2003.

[83] Robi Polikar. The wavelet tutorial part i. *IOWA State University, USA*, 1996.

[84] Robi Polikar. The wavelet tutorial part ii. *IOWA State University, USA*, 1996.

[85] Short-time fourier transform, url = https://en.wikipedia.org/wiki/short-time fourier transform.

[86] Robi Polikar. The wavelet tutorial part iii. *IOWA State University, USA*, 1996.

[87] Nitai D Mukhopadhyay and Snigdhansu Chatterjee. Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23.

[88] Jiuzhou Z Song, Kan-Ming Duan, and M Surette. The wavelet transformation for temporal gene expression analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 148–148. IEEE, 2005.

[89] Robi Polikar. The wavelet tutorial part iv. *IOWA State University, USA*, 1996.

[90] Wai Keng Ngui, M Salman Leong, Lim Meng Hee, and Ahmed M Abdelrhman. Wavelet analysis: Mother wavelet selection methods. In *Applied mechanics and materials*, volume 393, pages 953–958. Trans Tech Publ, 2013.

[91] Noor Kamal Al-Qazzaz, Sawal Ali, Siti Anom Ahmad, Md Shabiul Islam, and Mohd Izhar Ariff. Selection of mother wavelets thresholding methods in denoising multi-channel eeg signals during working memory task. In *IEEE Conference on Biomedical Engineering and Sciences*, pages 214–219. IEEE, 2014.

[92] J Rafiee, MA Rafiee, N Prause, and MP Schoen. Wavelet basis functions in biomedical signal processing. *Expert Systems with Applications*, 38:6190–6201, 2011.

[93] AI Megahed, A Monem Moussa, HB Elrefaie, and YM Marghany. Selection of a suitable mother wavelet for analyzing power system fault transients. In *IEEE Conference on Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century*, pages 1–7. IEEE, 2008.

[94] J Saraswathy, M Hariharan, Thiyagar Nadarajaw, Wan Khairunizam, and Sazali Yaacob. Optimal selection of mother wavelet for accurate infant cry classification. *Australasian Physical & Engineering Sciences in Medicine*, 37:439–456, 2014.

[95] Jacek M Zurada. *Introduction to artificial neural systems*. West St. Paul, 1992.

[96] Kishan Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. *Elements of artificial neural networks*. MIT press, 1997.

[97] Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, 1993.

[98] Bogdan M Wilamowski and Hao Yu. Neural network learning without backpropagation. *IEEE Transactions on Neural Networks*, 21:1793–1803, 2010.

[99] Jonathan Richard Shewchuk. An introduction to the conjugate gradient method without the agonizing pain, 2009.

[100] Reeves Fletcher and Colin M Reeves. Function minimization by conjugate gradients. *The computer journal*, 7:149–154, 1964.

[101] LE Scales. *Introduction to non-linear optimization*. Springer-Verlag New York, Inc., 1985.

[102] Michael James David Powell. Restart procedures for the conjugate gradient method. *Mathematical programming*, 12:241–254, 1977.

[103] EML Beale. A derivation of conjugate gradients. *Numerical methods for non-linear optimization*, 1:39–43, 1972.

[104] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

[105] Hideo Mure, Kazuhito Matsuzaki, Keiko T Kitazato, Yoshifumi Mizobuchi, Kazuyuki Kuwayama, Teruyoshi Kageji, and Shinji Nagahiro. Akt2 and akt3 play a pivotal role in malignant gliomas. *Neuro-oncology*, 12:221–232, 2009.

[106] Gregory S Yochum and Donald E Ayer. Role for the mortality factors morf4, mrgx, and mrg15 in transcriptional repression via associations with pf1, msin3a, and transducin-like enhancer of split. *Molecular and cellular biology*, 22:7868–7876, 2002.

[107] Min Deng, Fahui Li, Bryan A Ballif, Shan Li, Xi Chen, Lin Guo, and Xin Ye. Identification and functional analysis of a novel cyclin e/cdk2 substrate ankrd17. *Journal of Biological Chemistry*, 284:7875–7888, 2009.

[108] KATHARINA Strub and PETER Walter. Assembly of the alu domain of the signal recognition particle (srp): dimerization of the two protein components is required for efficient binding to srp rna. *Molecular and cellular biology*, 10:777–784, 1990.

# Appendix A

# Weight update rules for EBPA

## A.1 Weight update calculation of hidden layer and output layer of EBPA

### A.1.1 Weight update calculation for output layer neuron

The weight change for individual weight of hidden layer neuron in the direction of negative gradient is given as,

$$\Delta v_{kj} = -c\frac{\partial E}{\partial v_{kj}} \tag{A.1}$$

$\partial E$ / $\partial v_{kj}$ is given as,

$$\frac{\partial E}{\partial v_{kj}} = \left(\frac{\partial E}{\partial z_k}\right)\left(\frac{\partial z_k}{\partial net_k}\right)\left(\frac{\partial net_k}{\partial v_{kj}}\right) \tag{A.2}$$

The Root Mean Square Error between expected ouput and actual output of output layer neuron is given by equation

$$E = \frac{1}{2}\big(e_k - z_k\big)^2$$

The equation for ouput of ouput layer neuron is given as

$$z_k = f\big(net_k\big)$$

The net value of output layer neuron is given as

$$net_k = \sum_{k=0}^{K} v_{kj} y_j$$

Therefore, Equation A.2 becomes

$$\frac{\partial E}{\partial v_{kj}} = -\big(e_k - z_k\big) z_k{}' y_j$$

Therefore Equation A.1 becomes

$$\Delta v_{kj} = c\big(e_k - z_k\big) z_k{}' y_j$$

Hence, the weight update rule for output layer neuron is given as

$$v_{kj}{}' = v_{kj} + c(e_k - z_k) z_k{}' y_j \tag{A.3}$$

### A.1.2 Weight update calculation for hidden layer neuron

The weight change for individual weight of hidden layer neuron in the direction of negative gradient is given as

$$\Delta u_{ji} = -c \frac{\partial E}{\partial u_{ji}} \tag{A.4}$$

$\partial E$ / $\partial u_{ji}$ is given as,

$$\frac{\partial E}{\partial u_{ji}} = \left(\frac{\partial E}{\partial y_j}\right)\left(\frac{\partial y_j}{\partial net_j}\right)\left(\frac{\partial net_j}{\partial u_{ji}}\right) \tag{A.5}$$

$\partial E$ / $\partial y_j$ is given as,

$$\frac{\partial E}{\partial y_j} = \left(\frac{\partial E}{\partial z_k}\right)\left(\frac{\partial z_k}{\partial net_k}\right)\left(\frac{\partial net_k}{\partial y_j}\right) \tag{A.6}$$

For weight updation of hidden layer neuron the error at the ouput of every ouput layer neuron is utilised and it is given as,

$$E = \frac{1}{2} \sum_{k=0}^{K} \big(e_k - z_k\big)^2$$

Threrfore, Equation A.6 is given as,

$$\frac{\partial E}{\partial y_j} = -\sum_{k=0}^{K} \big(e_k - z_k\big) z_k{}' v_{kj}$$

The ouput of hidden layer neuron and net value of its ouput is given as,

$$y_j = f\big(net_j\big)$$

$$net_j = \sum_{j=0}^{J} u_{ji} a_i$$

Therefore Equation A.5 becomes,

$$\frac{\partial E}{\partial v_{ji}} = -a_i {y_j}' \sum_{k=0}^{K} \big(e_k - z_k\big) {z_k}' v_{kj}$$

Therefore Equation A.4 becomes,

$$\Delta u_{ji} = c a_i {y_j}' \sum_{k=0}^{K} \big(e_k - z_k\big) {z_k}' v_{kj}$$

Hence, the weight update rule for hidden layer neuron is given as,

$${u_{ji}}' = u_{ji} + c {y_j}' a_i \sum_{k=0}^{K} (e_k - z_k) {z_k}' v_{kj} \tag{A.7}$$

# Appendix B

# Publications

**Journals (Peer Reviewed)**

[1] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"Survey of Microarray Data Processing for Cancer Sub-Classification",* International Journal of Emerging Technology and Advanced Engineering, UGC Approved Journal (Serial No:44256), ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 4, April 2013, Impact Factor of 4.027.

[2] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"An Application of Wavelet Transform and Artificial Neural Network for Microarray Gene Expression based Brain Tumor Sub-classification",* International Journal of Emerging Technology and Advanced Engineering, UGC Approved Journal (Serial No:44256), ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 5, May 2015. Impact Factor of 4.027.

[3] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"Microarray Image Denoising",* International Journal of Emerging Technology and Advanced Engineering, UGC Approved Journal (Serial No:44256), ISSN 2250-2459, ISO 9001:2008

Certified Journal, Volume 7, Issue 3, March 2017, Impact Factor of 4.027.

[4] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"Percentage Change Method for Microarray Gene Expression Based Classification of Glioma",* International Journal of Emerging Technology and Advanced Engineering, UGC Approved Journal (Serial No:44256), ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 3, August 2017, Impact Factor of 4.027.

[5] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"Wavelet Transform Based Microarray Image De-noising",* International Journal of Emerging Technology and Advanced Engineering, UGC Approved Journal (Serial No:44256), ISSN 2250-2459, ISO 9001:2008 Certified Journal,Volume 7, Issue 3, August 2017, Impact Factor of 4.027.

[6] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"Stacked Autoencoder Network for Classification of Glioma Grade III and Grade IV",* Biomedical Signal Processing and Control, An Elsevier publication (under review).

**Conferences**

**International Conferences**

[1] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"Microarray Gene Expression Based Classification of Glioma Grade III, Grade IV Using Wavelet Transform and Artificial Neural Network",* at IEEE sponsored 3rd International conference on Electronics and Communication Systems (ICECS), Karpagam College of Engineering, Coimbatore, $25^{th}$ and $26^{th}$ February 2016.

[2] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"An Application of Thresholding Method for Microarray Gene Expression Data based Classification of Glioma Grade III and Grade IV",* at IEEE sponsored International conference on Engineering and Technology (ICET), Karpagam College of Engineering, Coimbatore, $16^{th}$ and $17^{th}$ December 2016.

[3] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"Neural Network Based Classification of Glioma Grade III and Grade IV",* at IEEE sponsored International conference on Engineering and Technology (ICET), Karpagam College of Engineering, Coimbatore, $16^{th}$ and $17^{th}$ December 2016.

**National Conference**

[4] Mrs. Supriya Patil, Prof. G.M. Naik, Dr. K. R. Pai *"An Application of Artificial Neural Network for Microarray Based Gene Expression Level Brain Tumor Sub-Classification",* at $9^{th}$ Annual National Symposium on VLSI and Embedded Systems on $11^{th}$ March 2015 at Carmel College, Goa.