Using Epigenomics Data to Predict Gene Expression in Breast Cancer

by

Nilisha Patel

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science (MSc) in Computational Sciences

The Faculty of Graduate Studies
Laurentian University
Sudbury, Ontario, Canada

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
Laurentian Université/Université Laurentienne
Faculty of Graduate Studies/Faculté des études supérieures

Title of Thesis Titre de la thèse	Using Epigenomics data to predict gene expression in Breast Cancer		
Name of Candidate Nom du candidat	Patel, Nilisha		
Degree Diplôme	Master of Science		
Department/Program Département/Programme	Computational Sciences	Date of Defence Date de la soutenance	August 14, 2019

APPROVED/APPROUVÉ

Thesis Examiners/Examinateurs de thèse:

Dr. Kalpdrum Passi
(Supervisor/Directeur de thèse)

Dr. Ratvinder Grewal
(Committee member/Membre du comité)

Dr. Thomas Merritt
(Committee member/Membre du comité)

Dr. Gulshan Wadhwa
(External Examiner/Examinateur externe)

Approved for the Faculty of Graduate Studies
Approuvé pour la Faculté des études supérieures
Dr. David Lesbarrères
Monsieur David Lesbarrères
Dean, Faculty of Graduate Studies
Doyen, Faculté des études supérieures

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

I, **Nilisha Patel**, hereby grant to Laurentian University and/or its agents the non-exclusive license to archive and make accessible my thesis, dissertation, or project report in whole or in part in all forms of media, now or for the duration of my copyright ownership. I retain all other ownership rights to the copyright of the thesis, dissertation or project report. I also reserve the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that this copy is being made available in this form by the authority of the copyright owner solely for the purpose of private study and research and may not be copied or reproduced except as permitted by the copyright laws without written authority from the copyright owner.

Abstract

Epigenetics is the study that deals with phenotype alterations that do not cause any modification in the DNA sequence of cells. Basically, it adds something to the top of DNA to alter its properties. This subsequently prevents the execution of certain behavior of DNA. Such epigenetic alterations are found in cancerous cells. These alterations are not the only cause of cancer; nevertheless, accurate statistical data that provides adequate shreds of evidence is still missing. In this research, four different types of data are used to bifurcate cancerous cells from non-cancerous cells. The data are Methylation, Histone, Human Genome and RNA-Seq data. The processing of these datasets is done using custom R-script. The tool that is used for feature selection and classification in the presented work is Weka 3.With the help of the machine learning method, the epigenetics data shows the prediction of breast cancer in the given set of cells.

Keywords: Epigenomics, Histone, DNA Methylation, Human Genome, RNA-Sequencing, Feature Selection

Acknowledgments

First and foremost, I would like to express my deepest gratitude and appreciation to my supervisor, Dr. Passi, who gave me an opportunity, support, knowledge and dealt with my queries with prodigious patience throughout the study period. He guided me step by step, which gave me the definition of a supervisor.

I want to thank all my friends and my family members who made my research experience enjoyable. I could never have completed my master's research without the support of my family and my friends. I offer special thanks to my parents for support and patience and believing in me.

Finally, and most significantly, I would like to acknowledge the persistent cooperation of my beloved family.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The source of cancer, which is known as cancer stem cells (CSCs) are a subset of tumor cells which have escaped cell cycle regulatory mechanisms, cell death, and yet have retained the immense self-renewing and proliferative potential of stem cells [4]. In this manner, they have the capacity to regenerate entire tumors from a limited number of cells. Their existence was first documented by Bonnet and Dick [5] in transplantation studies of human acute myeloid leukemia (AML) in mice with severe combined immunodeficiency disease (SCID). Of the transplanted leukemic cells, only an estimated population of between 0.01 and 1% of the total cell population was capable of initiating AML in the immune compromised mice. Termed SCID leukemia–initiating cells, they were found to undergo rapid clonal expansion and appeared to be at the top of a cancer cell hierarchy. Later it was shown that most AML tumor-initiating cells were at the stage of the multi potent progenitors (MPPs), and not the hematopoietic stem cells (HSCs) [6], opening the question as to how does a normally non-self-renewing cells gain both self-renewal and unlimited expansion.

The cells within a tumor are derived from tissues and organs which contain normal stem cells, progenitors, and lineage committed cells, e.g., the lineage hierarchy of the blood system. Identifying the complete roadmap of transitions from HSC through MPPs lacking self-renewal (short term- HSC, MPPs [7,8]) to common myeloid progenitor in mice [9] and humans [10], and Common lymphoid progenitor (CLP) [11], and downstream from them ever more committed progenitors (e.g., granulocyte–macrophage progenitor (GMP) and megakaryocyte/erythroid progenitor) [9,12] allowed Weiss man and colleagues to phenotypically isolate the cell types from which gave rise to Leukemia's. The critical development of a strain of mice where two pathways important in

programmed cell death was blocked in hematopoietic cells [13]. Serial transplantation of leukemia's from mice that developed AML could only be achieved with GMP cells. They further inferred that the progression to leukemia required at least five to seven rare events, either genetic or epigenetic; however, most of these events could not confer self-renewal, and so must have occurred in self-renewing cells to persist sufficiently to form a clone that was leukemic [14]. The pathways from the cell type that acquired the first oncogenic mutation, also known as the cell of origin to CSC, are varied and complex. However, it has been shown that in in vitro models of hematopoietic malignancy, certain oncoproteins may activate genetic programs involved in self-renewal, thereby conferring "stemness" to committed malignant cells and leading to the creation of CSCs. In vivo studies using an MLL1–AF9 mouse model of AML confirmed this finding but with an additional caveat: the amount of translocation product expressed determined the efficiency of CSC generation. Below a certain threshold, oncogenic capacity was limited, demonstrating the importance not only of oncoprotein presence, but also of gene dosage, in tumor formation. It is important to note, however, that the CSC population of a tumor is almost always genetically distinct from the cell of origin. Two models have been proposed to explain why this may be the case.

1.1 Identification of Cancer

CSCs are robust cells which may have acquired characteristics similar to their normal tissue stem cell. The expression of ABC transporters and telomerase and glutathione synthetize are properties of normal tissue stem cells that are extended into their CSC progeny that allow for cell survival and proliferation even after exposure to anticancer therapeutics. For example, certain gastrointestinal cancer cell lines show increased resistance to oxidative stress via interactions between CD44 and cell surface cysteine–glutamate exchange transporters which result in increased synthesis of reduced Glutathione, a key molecule involved in the neutralization of reactive oxygen species. In this manner, they are able to gain a survival advantage in inflammatory environments. Other studies have shown that CSCs are also capable of extensive metabolic reprogramming, rapid DNA damage repair, as well as enhanced drug

excretion through ATP-binding cassette (ABC) transporters of particular concern in the context of chemotherapeutic agents and other anticancer drugs. Given the aggressively proliferative nature of CSCs, it is no wonder that they have been found to contribute significantly to the formation of minimal residual disease (MRD).

1.1.1 Breast Cancer

Breast cancer is the most frequent cancer amongst women and is the second leading cause of cancer death among females. National Cancer Institute has estimated that the diagnosis of 246,660 new cases and 40,450 deaths from this disease in the United States, and the incidence is still rising. Breast cancer is an increasingly serious health problem all over the world, and its incidence and resistance to treatment are increasing significantly. Although improvement in both surgical techniques and neoadjuvant chemotherapy has been achieved, the prognosis of some breast cancer patients is still poor. There is increasing evidence that multiple genes and cellular pathways are involved in the development and progression of breast cancer. Therefore, identify new biomarkers of disease progression and signaling pathways is critical to discover more effective diagnostic and therapeutic strategies.

As of late, propels in microarray and high-throughput sequencing innovations have given a productive instrument to unraveling key hereditary adjustment in tumourigenesis and have discovered promising biomarkers for malignant growth finding and treatment [3, 4]. Numerous quality exes-articulation profiling studies have been performed on bosom malignant growth Carcinogenesis in the most recent decade, and many differentially communicated qualities have been obtained. Near examination of the differentially communicated qualities (DEGs) in free investigations demonstrates a generally constrained level of cover. The coordinated bioinformatics strategies joined with articulation profiling procedures can illuminate this burden.

In this study, raw data for GSE65212 was obtained from the GEO database, from which there were a total of 130 breast cancer cases and 11 normal breast tissue data available. We analyzed DEGs using the limma package with standard data processing. Subsequently, developed Gene ontology (GO) term

enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis for screening of DEGs with DAVID database. The protein-protein interaction (PPI) network was then established by STRING and visualized by Cytoscape software. We picked out hub genes with a high degree of connectivity. Besides, the Kaplan–Meier estimator of these hub genes was performed on the Kaplan–Meier plotter website. An integrated analysis of breast cancer on DEGs will provide further insight into the mechanism of breast cancer.

1.1.2 Bioinformatics

Bioinformatics is an interdisciplinary field that creates computational techniques and programming bundles for examining organic information. As an interdisciplinary field of science, bioinformatics joins innovations from software engineering, insights, and advancement to process natural information. A definitive objective of bioinformatics is to find new natural experiences through the examination of organic information.

Right now, a general pipeline for tending to a science issue in bioinformatics is as per the following 1: Wet labs configuration tests and get ready examples.

2. Enormous measures of organic information are created.

3. Existing (or new) computational and factual strategies are connected (or created).

4. Information examination results are additionally approved by wet lab testing.

5. On the off chance that vital, the methodology of 1–4 is rehashed with refinements.

Be that as it may, the bioinformatics investigate frequently mirrors a two-sided issue [1]: (1) Researchers in software engineering and other related fields simply see bioinformatics as one explicit use of their hypotheses and techniques because of the failure to give exact answers for complex sub-atomic science issues. (2) Biologists center around theory testing of wet labs so that bioinformatics fills in as an instrument for breaking down the natural information created from their trials. It isn't hard to see that the two sides have their own constraints. Computational researchers need a decent

comprehension of science and biomedical sciences, while scholars need to all the more likely comprehend the idea of their information examination issue from an algorithmic viewpoint. In this way, the absence of coordination of these different sides not just constrains the advancement of life science inquire about, yet in addition restricts the improvement of computational strategies in bioinformatics.

To understand bioinformatics, it is necessary to have a rudimentary grasp of biology. This section gives a brief introduction to some basic concepts of molecular biology that are relevant to the bioinformatics problems discussed in later chapters. The cell is the basic unit of life. Despite of the diversity of cells, they all have a life cycle: they are born, eat, replicate, and die. During the life cycle, a cell makes different decisions through the manifestation in pathways. Three types of basic molecules are present in a cell: deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. Intuitively, DNA, RNA, and proteins can be viewed as strings. DNA is a very long molecule that is composed of four types of bases: adenine (A), thymine (T), guanine (G), and cytosine (C). Similar to DNA, there are four bases in RNA as well. The major difference is that the T base is replaced by the base uracil (U) in RNA. Each protein is a string sequence consisting of 20 types of amino acids. DNA carries the genetic information of a cell and is composed of thousands of genes. Every cell contains the genetic information so that the DNA is duplicated before a cell divides (replication).

The respective genes are translated into RNA (transcription) when proteins are required. Therefore, RNA's primary responsibility is to synthesize the particular protein according to the protein-encoding information within the DNA (translation). Proteins are responsible for performing biochemical reactions, sending signals to other cells, and forming the body's major components.

1.2 Gene Expression

Advancement in DNA microarray technologies has made simultaneous monitoring of the expression levels of thousands of genes under different experimental conditions possible. The quality articulation information acquired through such advancements can be helpful for some applications in

bioinformatics if appropriately examined. For example, they can be utilized to encourage quality capacity forecast. The quality capacity expectation issue can be defined as a bunching issue so that, given a database of quality articulation information, a grouping calculation can be utilized to gather qualities that have comparable articulation profiles into groups. Since genes that perform the same biological functions are expected to exhibit similar expression patterns across different experimental conditions, the expression profiles of the genes belonging to the same cluster are expected to perform the same functions. Clustering gene expression data can, therefore, be useful in identifying different gene functional groups and a gene that is grouped with another gene in the same group can be expected to perform the same functions. For the purpose of clustering gene expression data, clustering techniques such as the hierarchical agglomerative clustering algorithm, the k-means algorithm, the self-organizing map (SOM), and the support vector machine (SVM) algorithm have been commonly used.

Other than formulating the gene function prediction problem as a clustering problem and tackling it using clustering algorithms, it should be noted that the problem can also be formulated as a Classification problem and tackled using classification techniques. Given the expression profiles of several classes of genes that perform known biological functions, classification techniques can be used to discover the characteristics of each gene class so that the gene whose class membership is not known earlier can be classified based on the characteristics discovered. The function that the gene performs can then be taken to be the same as that of the genes that belong to the class it is classified into. Classification techniques that have been used for gene function prediction include those that are based on the k-nearest-neighbor (k-NN) and the support vector machine (SVM).

Since biological processes are naturally complex, irregular expression patterns can always exist among genes that belong even to the same functional classes. Additionally, since quality articulation information are boisterous and have high dimensionality, quality capacity expectation, regardless of whether it is figured as a grouping or order issue, is troublesome and customary bunching and

characterization procedures, which are not initially created to manage quality articulation information, may not generally be the most appropriate For example, it may not be easy for these algorithms to discover that genes in a particular functional class are usually “highly expressed” under some experimental conditions, whereas they are likely to be “lowly expressed” under some others. The use of distance or correlation measures by these algorithms may not be able to uncover such patterns and this is especially difficult when the data being dealt with are very noisy. To discover such patterns, the quantitative gene expression data should best be discretized into intervals representing “highly expressed,” “lowly expressed,” etc. To do so, it should be noted that the discretization process divides quantitative data into non-overlapping “crisp” intervals, and such an approach has the disadvantage that it does not handle values at interval boundaries very well. A slight change in interval "Transcription of most protein coding genes in mammals is linked guanine nucleotide" limits may lead to very distinct interpretations of gene expression values and may introduce more noise in the information making it difficult to readily discover significant patterns.

In light of the ability of grouping in managing the vulnerabilities emerging from uproarious and vague information, which are very ordinary in articulation information and furthermore to make the examples found effectively interpretable by human clients, as of late, a few information digging approaches for quality articulation information examination have been proposed. In and, approaches were applied to search gene expression data for regulatory triplets consisting of the activator, repressor, and target genes. Gene expression levels are first converted into three different states (low, medium, and high) to varying degrees based on a set of predefined membership functions. Genes are then paired as an activator–repressor pairs to determine the expression value of the target gene based on a set of rules. These regulatory triplets are then ranked based on a residual score between predicted and actual expression values and a variance score of the activator– repressor gene pair. The triplets with low residual score and low variance score are then the most likely to exhibit the regulatory relationships. These approaches are not applicable to gene function prediction, as they are specifically developed for

solving gene regulatory networks reconstruction problems. In addition, the similarity or distance measures that existing data mining approaches – use do not tell us what expression levels under what experimental conditions are important in characterizing the genes in a functional class.

1.2.1 Epigenetic

Epigenetics is the investigation of cell and physiological attributes that are not brought about by changes in the DNA grouping. Epigenetics portrays the investigation of steady, long haul adjustments in the transcriptional capability of a cell, yet in addition can prompt transient changes. A portion of those modifications are heritable. For instance, during embryogenesis, totipotent undifferentiated organisms become the different pluripotent cell lines of the developing life, which
thusly become completely separated cells. This procedure is controlled by epigenetics.

1.2.2 Methylation

The second epigenetic system is the expansion of methyl gatherings to the DNA, generally at CpG locales, to change over cytosine to 5-methylcytosine (5-mC). 5-mC performs much like a customary cytosine, blending with a guanine in twofold stranded DNA. The chromatin structure adjacent to CpG island promoters allows transcription, while methylated CpG islands give chromatin a powerful compaction that prevents transcription and hence the gene expression of early CpG islands gives chromatin a tight compaction that stops transcription and hence gene expression. In the human genome, 60-80% of 28 million CpG dinucleotide are methylated [11]. The chromatin structure adjacent to the CpG island encourages development, while the methylated CpG islands give chromatin a tight compaction that prevents development and hence the gene expression of late CpG islands imparts a small compaction to chromatin that stops development and hence gene expression. A few regions of the genome are methylated more intensely than others, and exceedingly methylated territories will in general be less transcriptionally dynamic. Methylation of cytosine's can likewise persevere from the germ line of one of the guardians into the zygote, denoting the chromosome as being acquired from one parent or the other;

this is called hereditary engraving [12]. DNA methylation as often as possible happens in rehashed arrangements, and stifles the articulation and versatility of transposable components, for example, LINE-1 [13]. DNA methylation is related with histone alterations, especially the nonappearance of histone H3 lysine 4 methylation (H3K4me3) and the nearness of H3 lysine 9 methylation (H3K9me2).

DNA methylation examples are known to be built up and changed in light of ecological factors by a mind boggling interchange of in any event three autonomous DNA methyltransferases (DNMTs): DNMT1, DNMT3A, and DNMT3B [15]. These catalyze the methyl group transfer from S- adenosyl-methionine to cytosine bases on the DNA [16]. By specially altering hemi methylated.

DNA, DNA methyltransferase 1 (DNMT1) moves examples of methylation to a recently incorporated strand after DNA replication; it is hence regularly alluded to as the "support" methyltransferase [17]. DNMT1 is fundamental for legitimate embryonic improvement, engraving, and X-inactivation [18]. DNMT3 is a group of DNA methyltransferases that can methylate hemi methylated and unmethylated CpG at a similar rate. The design of DNMT3 compounds is like that of DNMT1, with an administrative locale connected to a synergist area [19]. There are three known individuals from the DNMT3 family: DNMT3A, 3B, and 3L. DNMT3A and 3B can intercede methylation-autonomous quality suppression, while DNMT3A can co-confine with heterochromatin protein (HP1) [20] and methyl-CpG-space restricting proteins (MBDs). They can likewise cooperate with DNMT1, which may be a co-usable occasion during DNA methylation. DNMT3L contains DNA methyltransferase themes and is required for setting up maternal genomic engraves, notwithstanding being chemically dormant. DNMT3L is communicated during gametogenesis when genomic engraving happens, yet additionally assumes a job in undifferentiated cell science [21].

1.2.3 Methyl binding domain portion

DNA methylation may influence the interpretation of qualities in two different ways. To start with, the methylation of DNA itself may physically hinder the authoritative of transcriptional proteins to the

quality; second and likely increasingly significant, methylated DNA might be bound by proteins known as MBDs [22]. MBDs at that point enroll extra proteins to the locus, for example, histone deacetylases and other chromatin redesigning proteins that can adjust histones, in this manner shaping minimized, latent chromatin, named heterochromatin. This connection between DNA methylation and chromatin structure is significant. Specifically, loss of methyl-CpG- restricting protein 2 (MeCP2) has been embroiled in Rett disorder, and methyl-CpG-restricting Space protein 2 (MBD2) intercedes the transcriptional hushing of hyper methylated qualities in malignant growth.

1.2.4 DNA Methylation

DNA methylation may influence the interpretation of qualities in two different ways. In the first place, the methylation of DNA itself may physically hinder the authoritative of transcriptional proteins to the quality; second and likely progressively significant, methylated DNA might be bound by proteins known as MBDs [22]. MBDs at that point select extra proteins to the locus, for example, histone deacetylases and other chromatin renovating proteins that can adjust histones, along these lines shaping conservative, inert chromatin, named heterochromatin. This connection between DNA methylation and chromatin structure is significant. Specifically, loss of methyl-CpG- restricting protein 2 (MeCP2) has been embroiled in Rett disorder, and methyl-CpG-restricting area protein 2 (MBD2) intercedes the transcriptional hushing of hyper methylated qualities in disease.

1.2.5 Histone Modification

Histones are the center protein segments of chromatin buildings and they give the auxiliary spine around which DNA wraps at customary interims, creating chromatin. The nucleosome speaks to the first dimension of chromatin association and is made out of two of every one of histones H2A, H2B, H3, and H4, amassed in an octamericcenter with DNA firmly folded over the octamer [2]. The first epigenetic component is the post-translational adjustment of the amino acids that make up histone proteins. On the off chance that the amino acids in the chain are changed, the state of the histone may be changed. DNA isn't totally loosened up during replication, and in this way it is conceivable that the altered histones

might be conveyed into each new duplicate of the DNA. Once there, these histones may act as layouts, starting forming of the encompassing new histones in the new way. By changing the state of the histones around them, these changed histones guarantee that a genealogy explicit interpretation program is kept up after cell division. In spite of the fact that histone alterations happen all through the whole grouping, the histone tails are especially exceptionally altered. These changes incorporate acetylation, methylation, ubiquitylation, phosphorylation, sumoylation, ribosylation, and citrullination, of which acetylation and methylation are the most very considered.

Histone adjustments are connected to basically every phone procedure requiring DNA get to, including translation, replication, and fix. Histone acetylation is completed by compounds called histone acetyltransferases (HATs) that are in charge of adding acetyl gatherings to lysine deposits on histone tails, while histone deacetylases (HDACs) are those that expel acetyl bunches from acetylated lysine's [3,4]. For instance, acetylation of the K14 and K9 lysine's of the tail of histone H3 by HATs is commonly identified with transcriptional capability. The nearness of acetylated lysine on histone tails prompts a casual chromatin express that advances transcriptional initiation of chose qualities; interestingly, deacetylation of lysine deposits prompts chromatin compaction and transcriptional inactivation [5].

1.3 Feature Selection

One major issue when applying an enormous epigenetic dataset onto the distinctive classifier is Redundancy. A few types of research demonstrated that the utilization of a feature selection strategy can some way or another improve the accuracy and furthermore decrease the redundancy. In our research, we are using feature selection methods including the following:

1. PCA (Principle Component Analysis)
2. CFS (Correlation-based feature selection)
3. Gain Ratio
4. ReliefF

1.4 Classification

In this research, we utilize the classification technique to measure and compare at the distinction between various feature selection techniques. Any classification strategy utilizes a lot of parameters to describe each object. These features are significant to the data being examined. Here we are discussing strategies for supervised learning. In supervised learning, there are labels on the data and the algorithms figure out how to predict the output from the input data.

In our research, we have used 8 classifiers: Gaussian SVM, Linear SVM, SVM, Logistic Regression, Random Forest, Neural Network, Naive Bayes and KNN.

1.5 Objectives of the study and outline of the thesis

In this research, we proposed estimation of fluctuate classification techniques. This incorporated the feature selection using Principal Component Analysis (PCA) , Correlation based feature selection (CFS), Gain ratio and ReliefF and implemented these strategies on raw data and in percentage-wise 95%,90%,85%,80%,75%,70%,65%,60%,55% and 50% as well.

The refined dataset will go through classification methods including, Gaussian SVM, Linear SVM, Logistic Regression, Naive Bayes, Random Forest, K-Nearest-Neighbor (KNN), and Neural Network. Cross validation being used is 10 fold and further in each classification, the cross validation percentage varies 95%,90%,85%,80%,75%,70%,65%,60%,55% and 50% as well.

The thesis is organized as below:

We will present the literature review and some of the previous works done on the epigenetic data. Plus, some results and works about feature selection methods will also be mentioned in this chapter.

In Chapter 3, we will present Datasets and Data processing. Chapter 4 includes Feature extraction and Feature Selection Methods. Chapter 5 describes Classification methods used for prediction of breast cancer. Chapter 6 describes about results and discussion. Chapter 7 presents Conclusions and Future work.

Chapter 2

Literature Survey

A non-linear method of multivariate analysis, weighted digital analysis (WDA), and estimated they have ability to predict lung cancer utilizing volatile biomarkers in the breath has been implemented by **Phillips** *et al.* [33]. By determining weight, a cutoff value, and a sign for each predictor variable used in the model, WDA produces a discriminating feature to predict disease membership vs. no disease groups. The weight of each predictor variable was the area under the curve (AUC) of the receiver controlling the characteristic (ROC) curve minus a fixed offset of 0.55, where the AUC was found by utilizing that predictor variable alone, as the sole marker of disease. The sign (±) was applied to invert the predictor variable if a lower value depicted a higher probability of disease. When utilized to evaluate the presence of a disease in a particular patient, the discriminant function was evaluated as the sum of the weights of all predictor variables that exceeded their cutoff values. The algorithm that generates the discriminant function was deterministic since the parameters were computed from each individual predictor variable without any optimization or adjustment. They used WDA to re-evaluate information from a latest research of lung cancer breath biomarkers, which included volatile organic compounds (VOCs) in the alveolar breath of 193 individuals with original lung cancer and 211 tests with adverse chest CT. Their method WDA discriminant function accurately distinguished the patients with lung cancer in a model utilizing 30 breath VOCs (ROC curve AUC = 0.90; sensitivity = 84.5%, specificity = 81.0%).

A multi-dimensional linear adaptive filters and support vector regression to evaluate the motion of lung tumors tracked at 30 Hz has been analyzed by Nadeem Riaz *et al.* [34]. To determine the motion of a tumor, they expand on the prior work of other groups who have looked at adaptive filters by applying a general framework of a multiple-input single-output (MISO) adaptive system that uses

multiple correlated signals. They equate these two fresh methods ' efficiency with standard techniques such as linear regression and single-input, single-output adaptive filters. The average root-mean-square-errors (RMSEs) are 2, 58, 1, 60, 1, 58, 1, 71 and 1, 26 mm at 400 ms latency for the 14 therapy sessions studied without prediction, linear regression, single-output adaptive filter, MISO and vector support regression. At 1 s, the RMSEs are 4.40, 2.61, 3.34, 2.66 and 1.93 mm, respectively. They discover that supporting vector regression assesses the future tumor position of the analyzed techniques most appropriately and can provide an RMSE of less than 2 mm at 1 s latency. Also, a multi-dimensional adaptive filter framework provides developed performance over single-dimension adaptive filters.

A support vector machine (SVM) classification for lung cancer was inquired; presenting a systematic quantitative evaluation against Boosting, Decision trees, k-nearest neighbor, LASSO regressions, neural networks and random forests has been invented by **Tao Sun** et al. [35]. For their outcomes, a large database of 5984 fields of concern (ROIs) and 488 features of input (including textural features, patient characteristics, and morphological features) was used to train and evaluate the classifiers. The evaluation for classifiers' performance was established on a tenfold cross validation framework, receiver operating characteristic curve (ROC), and Matthews's correlation coefficient. Area under curve (AUC) of SVM, Boosting, Decision trees, k-nearest neighbor, LASSO, neural networks, random forests were 0.94, 0.86, 0.73, 0.72, 0.91, 0.92, and 0.85, respectively. To the reference methods, it was proved that SVM classification offered crucially enhanced classification performance equated.

Temesguen Messay et al. [36] have presented novel pulmonary nodule segmentation algorithms for computed tomography (CT). Those admit a fully-automated (FA) system, a semi-automated (SA) system, and a hybrid system. Like most traditional systems, the novel FA system needs only a single user-supplied cue point. On the other hand, the SA system refers a novel algorithm class requiring 8

User-supplied control points. Those enhance the burden on the user, but they establish that their

resulting system was highly robust and can handle a variety of challenging cases. Their presented hybrid system starts with the FA system. If developed segmentation results were required, the SA system was then deployed. There are 2 free parameters for the FA segmentation engine and 3 for the SA scheme. These parameters were adaptively evaluated for each nodule in a search process guided by a regression neural network (RNN). The RNN applies a number of features computed for each candidate segmentation. They train and test their systems using the new Lung Image Database Consortium and Image Database Resource Initiative (LIDC–IDRI) data. To the best of their knowledge, that was one of the first nodule-specific performance benchmarks using the new LIDC–IDRI dataset. They also equate the performance of the introduced methods with several previously reported results on the same data applied by those other methods. Their results proposed that their suggested FA system develops upon the state-of-the-art, and the SA system offers a considerable boost over the FA system.

Lung cancer was induced by abnormal and uncontrolled growth of cells in the lungs and the mortality rate of lung cancer was the highest between all types of cancer. They can be inquired and treated with the help of computed tomography (CT) images. From an image was a key concern, for an automated classifier, identifying good features. Deep feature extraction applying pre-trained convolutional neural networks has been successful for some image domains recently. A pre-trained convolutional neural network (CNN) to extract deep features from lung cancer CT images and then train classifiers to predict short and long term survivors has been evaluated by Rahul Paul *et al.* [37]. The best accuracy of 77.5% was with a cropping approach applying a decision tree classifier in a leave one out cross validation with ten features selected and applying symmetric uncertainty feature ranking. They mixed extracted deep neural network features along with quantitative (traditional image) features and found the best accuracy of 82.5% with a nearest neighbor classifier in a leave one out cross validation applying the symmetric uncertainty feature ranking algorithm.

Lung cancer was one of the leading induces of death worldwide. There were three main types of lung cancers, non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC) and carcinoid. NSCLC was further separated into lung adenocarcinoma (LADC), squamous cell lung cancer (SQCLC) as well as large cell lung cancer. Many previous studies established that DNA methylation has emerged as potential lung cancer-specific biomarkers. However, whether there exists in a set of DNA methylation markers simultaneously distinguishing such three types of lung cancers remains elusive. A ROC (Receiving Operating Curve), RFs (Random Forests) and mRMR (Maximum Relevancy and Minimum Redundancy) to capture the unbiased, informative as well as compact molecular signatures followed by machine learning methods to separate the LADC, SQCLC and SCLC has been analyzed by Cai *et al*. [38]. As a consequence, a 16-DNA methylation marker panel achieves an optimal classification authority with 86.54 percent precision, 84.6 percent precision and 84.37 percent recall, 85.5 percent in cross-validation leave-one-out (LOOCV) and autonomous test set tests, respectively. Besides, comparison results depicted that ensemble-based feature selection methods outperform individual ones when combined with the incremental feature selection (IFS) strategy in terms of the informative and compact property of features. Taken together, the findings collected suggest the efficacy of the ensemble-based strategy to selecting features and the possible presence of a particular panel of DNA methylation markers among these three kinds of lung cancer tissue that would enable clinical diagnosis and therapy.

Lung cancer was one of the diseases responsible for a large number of cancer related death cases worldwide. The recommended standard for screening and early detection of lung cancer was the low dose computed tomography. However, lots of patients discovered die within one year, which makes they require to find alternative approaches for screening and early detection of lung cancer. A computational technique for the classification, screening and early detection of victims of lung cancer that can be introduced in a functional multi-genomic scheme has been suggested by Adetiba et al.[39]. In order to validate computational methods, samples of the top ten biomarker genes previously reported to have the highest frequency of lung cancer mutations and sequences of normal biomarker genes were

collected from the COSMIC and NCBI databases. Experiments have been performed on the mixing of Z-curve and tetrahedron affine transform neural networks, Histogram of Oriented Gradient (HOG), Multilayer Perceptron and Gaussian Radial Base Function (RBF) to obtain a suitable combination of computational techniques for acquiring sophisticated classification of biomarker genes for pulmonary cancer. Results show that a mixture of Voss representation affine transformations, HOG genomic characteristics, and Gaussian RBF neural network perceptibly develops precision, specificity, and sensitivity classification of lung cancer biomarker genes as well as tiny mean square error [39].

Due to current progress in Convolutional Neural Networks (CNN), developing image-based CNN models for predictive diagnosis was gaining enormous interest. However, to date, insufficient imaging samples with truly pathological-proven labels impede the estimation of CNN models at scale. A domain-adaptation framework that learns transferable deep features for patient-level lung cancer malignancy prediction has been proposed by Shen , *et al.* [40]. The major aim of their presented article was learns CNN-based features from a large discovery set (2272 lung nodules) with malignancy likelihood labels admitting multiple radiologists' assessments, and then tests the transferable predictability of those CNN-based features on a diagnosis-definite set (115 cases) with true pathologically-proven lung cancer labels. They estimate their approach on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset, where both human expert labeling information on cancer malignancy likelihood and a set of pathologically- proven malignancy labels were rendered. Their model submitted can significantly reduce demand for pathologically proven information, promising to empower the diagnosis of cancer by leveraging multi-source CT imaging datasets. The scope of the applications of breath sensors was abundant in disease diagnosis. The diagnosis of lung cancer was a well-fitting health-related implementation of this technology, which was of utmost importance in the health industry, as lung cancer has the largest death rate among all kinds of cancer and carries a heavy annual worldwide burden. Yekbun Adiguzel et al.[41] tested a rational basis for the creation of breath sensors for the diagnosis of lung cancer from a historical view

that will enable the transfer of the concept into the rapidly changing sector of sensors. Colorimetric, composite, carbon nanotube, gold nanoparticle-based, and surface acoustic wave sensor arrays are admitted following examples with diagnostic apps. These select sensor applications were widened by the state-of-the-art developments in the sensors field. Coping with sampling sourced artifacts and cancer staging are among the debated topics, along with the other concerns like proteomics strategies and biomimetic media utilization, feature selection for data classification, and commercialization. A Multi-crop Convolutional Neural Network (MC-CNN) to automatically extract nodule salient information by utilizing a novel multi-crop pooling strategy which crops various regions from convolutional feature maps and then applies max-pooling various times has been estimated by **Shen** *et al.* [32]. Extensive experimental findings show that their introduced technique not only achieves state-of - the-art nodule suspiciousness classification efficiency, but also characterizes nodule semiconductor characteristics (subtlety and margin) and nodule diameter that may have been useful in modeling nodule-malignancy.

Chapter 3

Dataset and Data Processing

3.1 Dataset Selection

Data processing is done using R-Script whereas; classification, as well as feature selection, is done using data mining software – Weka. In the context of breast cancer, an approach is introduced here that can manifest the cancerous cells and bifurcate them from non-cancerous cells. Several data sets are used for feature selection and classification purposes. The data sets include methylation, human genome, histone, and RNA- Sequencing. These data of breast cancer are processed further using smart machine learning tools and algorithms.

3.2 Data Processing

3.2.1 DNA Methylation Data

The Cancer Genome Atlas (TCGA) Methylation information from Illumina's Infinium HumanMethylation450 Bead chip (Illumina 450k) were utilized to separate CpG methylation related highlights, as indicated by their comment document. The data is selected from TCGA which is produced by clinicians. The genomic directions of CpG, their exons and coding areas were acquired from the Illumina comment document. Since the comment document just given data of transcripts, exons, and coding DNA arrangements (CDS), we re-annoted the protein coding qualities utilizing the Illumina iGenomes hg19 Refseq explanation so as to remove progressively exhaustive data from different areas of the transcripts: all introns (with uncommon classes for the first and last intron), just as first and last exons, untranslated locales in the 5' and 3' heading (5' UTR and 3' UTR, individually), and a "single exon" or "single intron" assignment for transcripts that just had a single exon or single intron [42].

3.2.2 Histone Data

Two cell lines were regarded for three schemes of histone marker CHIP-Seq data, H3k4me3, H3k27me3, and H3k36me3: MCF-7 cell line (0.2% EtOH therapy) from breast carcinoma tissue and SAEC normal breast epithelial line (no therapy). The data is selected from UCSC genome browser which is produced by clinicians. In a joint attempt with the ENCODE project through the UCSC genome browser at http:/genome.ucsc.edu. Raw CHIP-Seq data was downloaded from the Broad Institute / Bernstein Lab at the Massachusetts General Hospital / Harvard Medical School and the University of Washington [10, 11]. The raw data were handled in-house to guarantee consistency of all standardization systems. Raw data were first aligned to hg19 utilizing bowtie2 [12], followed by the removal of duplicate reads by the Sam tools toolbox (explicitly, the "rmdup" device) [13]. The aligned reads were intersected with the relevant segments of the transcript by utilizing the Bed tools toolbox (explicitly, the "multicov" instrument) [14]. A custom R script was utilized to standardize the data over total number of reads after removing PCR duplicates [42].

3.2.3 Human Genome Data

We have extracted nucleotide composition data from hg19 genome FASTA files downloaded from the UCSC genome browser. The data is selected from UCSC genome browser which is produced by clinicians. Conservation scores across three classes of species: vertebrates, primates, and placental animals, were additionally considered. PhastCons46Way scores were downloaded from the UCSC genome browser [11, 15]. Conservation scores were then intersected with the relevant segments of the transcripts using a custom Perl script, in order to extract conservation features [42].

3.2.4 RNA-Seq Data

RNA-Seq gene expression data from lung cancer samples with coupled CpG methylation data were downloaded from TCGA Research Network: http://cancergenome.nih.gov. Lung adenocarcinoma and

lung squamous cell carcinoma data were combined for this project, as they are two subtypes of non-small cell lung cancer. The data is selected from TCGA which is produced by clinicians. Differential expression analysis was done with the DESeq2 package in R [16]. In cases where multiple transcripts are mapped to the same Refseq ID, the geometric mean of the differential expression results was used to represent the gene level expression. In the case that any of these read counts was zero, the counts from all transcripts were artificially increased by one in order to calculate the geometric mean, followed by final subtraction of one. The expression of a gene was then classified as binary outcomes: either up-regulated or down-regulated, once it passed two thresholds: 1) having an adjusted p value < .05 after Holm's multiple hypothesis test [17] and 2) having an absolute value of log2 fold change greater than 1. As a result, 2874 genes were selected as "differentially expressed" genes [42].

Chapter 4

Feature Extraction and Feature Selection Methods

Using these diverse data sets, various features are obtained. All these features are then extracted using exclusive methods. Histone modification data, DNA sequence data, and DNA methylation 450K data are combined as predictors. Later, RNA-Seq expression data (up versus down-regulated genes put into binary) is used as response variables. Illumina 450K annotation file is used to get the specific features for breast cancer cells [42].

4.1 Feature Extraction

The extracted features are categorized into four major sub-groups. All feature were considered on a segment-wise basis.

4.1.1 CpG Methylation feature

Differential expression of the methylated CpG sites was processed using the limma library in R. Specifically, the function toptable was used to determine the log fold change (logFC) between the cancer and normal tissues as well as the average methylation (avgMval) of each CpG site across the two types of tissue [18]. A positive logFC indicates hypermethylation whereas a negative logFC indicates hypomethylation. Additional segment-based features were also considered. These include the number of hypermethylated (numHyper) and hypomethylated probes (numHypo) on a segment of a given transcript. For example, first_exon_numHyper refers to the number of hypermethylated probes on the first exon. Two other types of features are the average of logFC and avgMval of all CpG probes on a segment of the transcript, e.g. the average logFC of all probes on the first exon of a given transcript (first_exon_avglogFC).

Special effort was paid to compute distances of CpG probes to exon-exon junctions. Given that one or more CpG sites may exist on the individual exon segments of a transcript (including the first and last exons), transcript-level maximum, minimum and average distances of any hyper/hypo-methylated probe to the nearest 5' or 3' exon-exon junction were computed (maxHypoTo5, minHypoTo5, avgHypoTo5, maxHypoTo3, minHypoTo3, avgHypoTo3, maxHyperTo5, minHyperTo5, avgHyperTo5, maxHyperTo3, minHyperTo3, and avgHyperTo3) [42].

4.1.2 Histone marker change feature

After the alignment of raw histone marker data, the aligned histone marker reads were intersected with the segments of each transcript using the multicov function from the BEDTools package [19]. The histone reads were then normalized per 1000 bp length of each segment per 1 million aligned read library. Similar to the CpG methylation features, the histone marker modification features were extracted on a segment-by-segment basis. Initials are used to represent the individual cell lines where the features come from: *A* for the MCF-7 cell line and *S* for the SAEC cell line. Following the initial is a number representing the specific histone H3 methylation marker: 4 for H3k4me3, 27 for H3k27me3, and *36* for H3k36me3. As a result, features are named as segment_cell type and histone modification type (e.g. first_exon_A4). In order to compare histone modification between the cancer and non-cancer cell types, the differences of the reads between them were divided by the average of the two (e.g. a feature named first_exon_A4_minus_S4_divavg) [42].

4.1.3 Nucleotide feature

In each segment of the transcript, four different types of nucleotide features were extracted: single nucleotide composition, dinucleotide composition, trinucleotide composition, and the length of each segment. Nucleotide sequences of Hg19 reference genome were processed using the Biostrings library in R [20, 42].

4.1.4 Conservative feature

Conservation score per segment was calculated as the arithmetic mean of the conservation score per nucleotide in that segment. Three separate sets of conservation scores with different comparative species were extracted from UCSC genome browser - vertebrate, primate, or placental. Thus, features such as first_exon_vertebrate emerge from this set [42].

4.2 Feature Selection Techniques

Selection of features, also known as machine learning variable selection, selection of attributes or selection of variable subsets. It is the process towards choosing a subset of relevant features (factors, indicators) for use in model development. There are three different approaches of feature selection. They are as follows:

4.2.1 Wrapper strategy

This strategy gives high computing complexity. It utilizes a learning algorithm to assess the precision of classification using the chosen characteristics. Wrapper techniques can offer specific classifiers elevated classification precision. Examples of these methods are recursive feature elimination, greedy algorithms.

4.2.2 Filter method

This method selects a subset of characteristics without any learning algorithm being used. This technique is used by higher-dimensional data sets and is comparatively quicker than methods based on the wrapper. Examples of filter methods are PCA, CFS, Gain Ratio and RelifF methods.

4.2.3 Embedded approach

The applied learning algorithms determine this approach's specificity and select characteristics during the information set training phase. In this method, we have to add a penalty against unpredictability to decrease the level of over fitting or fluctuation of a model by including more bias. The examples of this method are L1 (LASSO), decision tree.

4.3 Feature selection Methods

In this research, we have used four different types of feature selection methods. They are Principle Component Analysis (PCA), Correlation based feature selection (CFS), Gain ration and ReliefF.

4.3.1 Principal Component Analysis (PCA)

Industrial process data are becoming huge and increasingly valuable assets for decision making in process operations, process control, and monitoring. Since process measurements are often highly correlated, latent variable methods, such as principal component analysis (PCA) and partial least squares (PLS), are useful analytic tools for data modeling and process monitoring. In PCA, the objective is to extract latent variables from the data such that the variance of the extracted latent variables is maximized. By applying PCA, the measurement space can be decomposed into a principal subspace with maximized variability and a residual subspace. Fault detection statistics are developed for each subspace for process monitoring.

The PCA is a mathematical process, which is used to reduce the dimensions of dataset by taking the first several PCs to represent the original high dimensional dataset. For a specific initial dataset, after PCA, the average and PC matrices would be unique and each sample in the initial dataset can be approximated and controlled by a specific PCS group.

Principal Component Analysis, or PCA, is a dimensionality-decrease technique that is frequently used to lessen the dimensionality of enormous informational collections, by changing a huge arrangement of factors into a littler one that still contains the majority of the data in the huge set.

Diminishing the quantity of factors of an informational collection normally comes to the detriment of exactness; however the trap in dimensionality decrease is to exchange a little precision for effortlessness. Since littler informational collections are simpler to investigate and picture and make dissecting information a lot simpler and quicker for AI calculations without unessential factors to process.

Step 1: Standardization

The point of this progression is to institutionalize the scope of the persistent beginning factors with the goal that every single one of them contributes similarly to the examination.

All the more explicitly, the motivation behind why it is basic to perform institutionalization preceding PCA is that the last is very touchy with respect to the differences of the underlying factors. That is, if there are enormous contrasts between the scopes of beginning factors, those

factors with bigger reaches will rule over those with little ranges (For instance, a variable that ranges somewhere in the range of 0 and 100 will command over a variable that ranges somewhere in the range of 0 and 1), which will prompt one-sided results. Along these lines, changing the information to practically identical scales can anticipate this issue.

Scientifically, this should be possible by subtracting the mean and partitioning by the standard deviation for each estimation of every factor.

z=(value-mean)/standard deviation

When the institutionalization is done, every one of the factors will be changed to a similar scale.

Step 2: Covariance Matrix computation

The point of this progression is to see how the factors of the information informational collection are changing from the mean regarding one another, or as it were, to check whether there is any connection between them. Since now and again, factors are profoundly corresponded so that they contain excess data. In this way, so as to distinguish these relationships, we figure the covariance framework.

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Covariance matrix for three dimensional data. Since the covariance of a variable with itself is its

difference (Cov (a, a) =Var (a)), in the primary corner to corner (Top left to base right) we really have the fluctuations of each underlying variable. Furthermore, since the covariance is commutative (Cov (a, b) =Cov (b, a)), the sections of the covariance lattice are symmetric as for the primary inclining, which implies that the upper and the lower triangular segments are equivalent.

Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components

Eigenvectors and eigenvalues are the straight polynomial math ideas that we have to process from the covariance network to decide the main parts of the information. Before getting to the clarification of these ideas, allows first comprehend what we mean by primary segments.

Essential segments are new factors that are developed as straight mixes or blends of the underlying factors. These blends are done so that the new factors (i.e., essential segments) are uncorrelated, and the vast majority of the data inside the underlying factors is crushed or packed into the original parts. In this way, the thought is 10-dimensional information gives you 10 essential parts, yet PCA

Attempts to put most extreme conceivable data in the primary segment, at that point highest residual data in the second, etc., until having something like appeared in the plot beneath.

Organizing information in principal components this way, will allow you to reduce dimensionality without losing much information, and this by discarding the components with low information and considering the remaining components as your new variables.

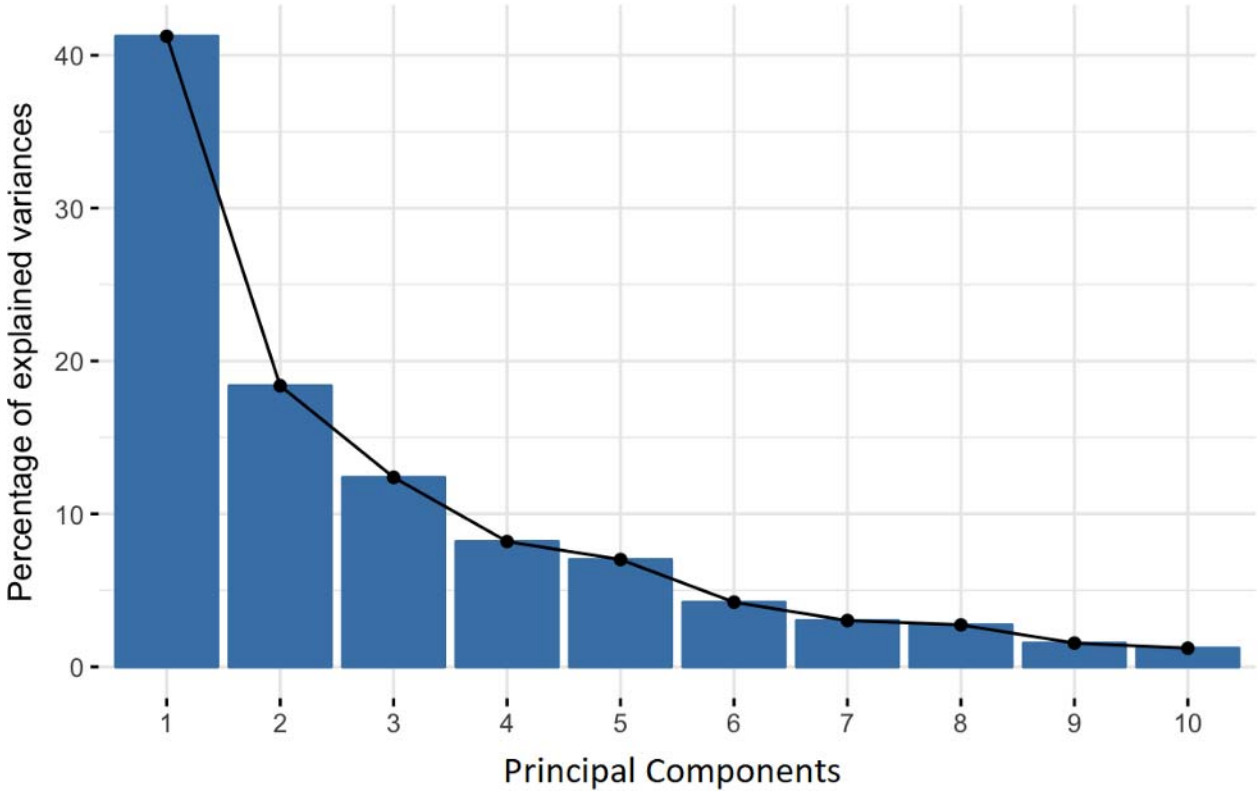


Figure 1: Graph of PCA

An important thing to realize here is that, the principal components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables.

Geometrically speaking, principal components represent the directions of the data that explain a maximal amount of variance, that is to say, the lines that capture most information of the data. The relationship between variance and information here is that, the larger the variance carried by a line, the larger the dispersion of the data points along it and the larger the dispersion along a line, the more the information it has. To put all this simply, just think of principal components as new axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible.

Construction of the Principal Components

As there is the same number of vital parts as there are factors in the information, essential segments are built in such a way, that the main important segment represents the most significant conceivable fluctuation in the informational index. For instance, how about we expect that the disperse plot of our informational collection is as demonstrated as follows, would we be able to figure the central, vital part is a question. Honestly, it is roughly the line that coordinates the purple imprints since it experiences the inception, and it is the line where the projection of the focuses (red specks) is the most spread out. Alternatively, on the other hand numerically, the line augments the change (the normal of the squared separations from the anticipated focuses (red spots) to the starting point).

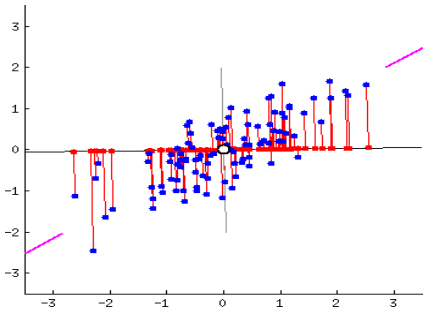


Figure 2: PCA Demonstration

The second principal component is determined similarly, with the condition that it is uncorrelated with (i.e., opposite to) the original central part and that it represents the following most noteworthy change. This proceeds until a sum of p key segments have been determined, equivalent to the first number of factors. Since we comprehended what we mean by crucial parts, we should return to eigenvectors and eigenvalues. What you right off the bat need to think about them is that they generally come two by two, so every eigenvector has an eigenvalue. What is more, their number is equivalent to the number of measurements of the information. For instance, for a 3-dimensional informational index, there are 3 factors; in this manner, there are 3 eigenvectors with 3 relating eigenvalues.

Right away, it is eigenvectors and eigenvalues which are behind all the enchantment clarified above,

because the eigenvectors of the Covariance framework are the bearings of the tomahawks where there is the most fluctuation (most data) and that we call Principal Components. Furthermore, eigenvalues are just the coefficients appended to eigenvectors, which give the measure of difference conveyed in every Principal Component.

The covariance framework is a p × p symmetric network (where p is the quantity of measurements) that has as passages the covariance related with every single imaginable pair of the underlying factors. For instance, for a 3-dimensional informational index with 3 factors x, y, and z, the covariance framework is a 3×3 network of this from:

Use of PCA in breast cancer

AI is a part of man-made reasoning that utilizes an assortment of factual, probabilistic and streamlining strategies that enables Principal Component to gain from past models and to identify designs from enormous informational indexes. This ability is especially appropriate to medicinal applications.

Utilizing the covariance lattice and its eigenvalues and eigenvectors, PCA finds the "head segments" in the information which are uncorrelated eigenvectors each speaking to some extent of difference in the information. PCA and numerous varieties of it have been connected as a method for decreasing the dimensionality of the information in malignant growth microarray information [58–64]. It has been contended [65, 66] that when figuring the vital parts (PCs) of a dataset there is no assurance that the PCs will be identified with the class variable. Accordingly, managed head segment investigation (SPCA) was proposed, which chooses the PCs dependent on the class factors. They named this additional progression the quality screening step. Despite the fact that the managed variant of PCA performs superior to the unsupervised, PCA has a significant confinement: it can't catch nonlinear connections that regularly exist in information, particularly in complex natural frameworks. SPCA fills in as pursues:

(1) Compute the connection measure between every quality with result utilizing direct, calculated, or corresponding perils models.

(2) Select qualities most connected with the result utilizing cross-approval of the models in step.

(3) Estimate head segment scores utilizing just the chose qualities.

(4) Fit relapse with result utilizing model in step.

The strategy was exceptionally viable in distinguishing significant qualities and in cross- approval tests was just outflanked by quality shaving, a factual technique for grouping, like progressive bunching. The principle distinction is that the qualities can be a piece of more than one bunch. The expression "shaving" originates from the expulsion or shaving of a level of the qualities (ordinarily 10%) that have the littlest supreme inward item with the main head segment [67].

A comparative straight methodology is old style multidimensional scaling (old style MDS) for Principal Coordinates Analysis [68] which ascertains the lattice of dissimilarities for some random grid input. It was utilized for enormous genomic datasets on the grounds that it is effective in mix with Vector Quantization or - Means [69] which doles out every perception to a class, out of a sum of classes [70].

4.3.2 Correlation-based feature selection (CFS)

Correlation-based feature selection (CFS) as stated by Hall [17] mainly pursues that "a large subset of elements is one that contains excessively corresponding to the class yet uncorrelated to each other". CFS evaluates a subset by considering only the prescient capacity of each of its last characteristics and, moreover, its surplus (or connection) level. The difference between CFS and various methods is that it provides free "heuristic validity" to a subset of components rather than to each element[18].This implies given a capacity (heuristic), the calculation can settle on its best courses of action by choosing the choice that expands the yield of this capacity. Heuristic capacities can likewise be intended to limit the expense to the objective.

4.3.3. ReliefF

ReliefF is likewise generally utilized with malignant growth microarray information. It is a multivariate technique that picks the feature that is the most recognizable among the various classes. It over and again draws a case (test) and, in view of its neighbors, it gives most weight to the feature that help segregate it from the neighbors of an alternate class. A strategy utilizing free calculated relapse with two stages was additionally proposed. The initial step is a univariate strategy in which the qualities are positioned by their Pearson connection coefficients. The top qualities are considered in the second stage, which is stepwise factor choice. This is a restrictively univariate technique dependent on the incorporation (or prohibition) of a solitary quality at any given moment, molded on the factors effectively included.

ReliefF in breast cancer

ReliefF [19] is also commonly used for microarray information on cancer. It is a multivariate technique that selects the most distinguishable characteristics among the various categories. It constantly brings an example (sample) and, depending on its neighbors, provides the most weight to the characteristics that assist to distinguish it from other category neighbors [20, 21]. A technique was also suggested using an autonomous two-step logistical regression [22].The primary stage is a univariate strategy in which the qualities are grouped by their coefficients of Pearson connection. In the subsequent stage, which is step by step factor choice, the top qualities are respected. This is a univariate technique restrictively dependent on the incorporation (or prohibition) of a solitary quality at any given moment, adapted on the factors effectively included.

4.3.4 Gain ratio

Gain Ratio (GR) is a data gain alteration that decreases its prejudice. When selecting an attribute, the gain percentage takes into consideration the amount and size of branches. It corrects the benefit of data by taking into consideration a split's inherent data. Intrinsic information is entropy of instance

allocation into branches (i.e. how much information we need to say which branch an instance belongs to). Attribute value reduces as data becomes bigger intrinsically.

Gain Ratio = Gain (Attributes)/ () intrinsic_info (Attributes)

Use of gain ratio in breast cancer

Investigating the concealed examples in the datasets of the restorative field is the monotonous assignment in therapeutic information mining. These examples can be used for clinical analysis. Information preprocessing incorporates information cleaning, information combination, information change and information decrease. These information preprocessing strategies can considerably improve the general nature of the fascinating examples mined as well as the time required for the genuine mining. Information preprocessing is significant for learning disclosure process as quality choices depends on quality information. The point of information decrease is to locate a base arrangement of properties with the end goal that the subsequent likelihood dissemination of the information classes is as close as conceivable to the first dispersion got utilizing all qualities. Mining on the decreased arrangement of traits has extra advantages. It decreases the quantity of traits showing up in the found examples, making the examples more obvious. Further, it improves the precision of characterization and learning runtime (Han and Kamber 2001).

Split strategy is the most significant segment of choice tree student. To achieve high predictive exactness in various circumstances, the split strategy (data gain proportion) is the best one.

The data increase measure is one-sided towards tests with numerous results. The significant downside of utilizing data increase is that it will in general pick properties with enormous quantities of unmistakable qualities over traits with less quality despite the fact that the last is increasingly enlightening (Asha et al 2012). For instance, think about a quality IE, name of the malady in the patient database. A split on a malady name would result in an enormous number of allotments; as each record in the database has an alternate name for various patients. So the data required to arrange

database with this apportioning would be of little significance and such a parcel is pointless for order.

C4.5, a successor of ID3 (http://www.cs.waikato.ac.nz/□ml/weka/), utilizes an expansion to pick up data known as an addition proportion (GR), which endeavors to conquer the predisposition. The WEKA (Quinlan 1986) classifier bundle has its own rendition of C4.5 known as J4.8. We have utilized J4.8 to distinguish the critical properties. Give D a chance to be a set comprising of'd' information tests with n unmistakable classes. At that point the normal data expected to group the example is given by Equation 1:

$$I\,(D) = -\sum_{i=1}^{n} p_i \log_2 p_i,$$

Where pi is the probability of an arbitrary sample which belongs to class Ci. Let attribute A have 'V' distinct values. Let ' dij ' be amount of Class Ci specimens in a Dj sub-set. Dj includes the specimens in D with the significance ' aj ' of A. The entropy based on partitioning into subsets by A, is given by

Equation 2:

$$E\,(A) = -\sum_{i=1}^{n} I\,(D)\,\frac{d_{1i} + d_{2i} + \ldots + d_{mi}}{d}.$$

The encoding data that branching A would gain is

Equation 3:

Gain (A) = I (D) − E (A)

C4.5 applies a sort of standardization to data addition utilizing a 'split data' esteem which is characterized comparably with Info (D) as

Equation 4:

$$E\,(A) = -\sum_{i=1}^{n} I\,(D)\,\frac{d_{1i} + d_{2i} + \ldots + d_{mi}}{d}.$$

This worth speaks to the data registered by part the dataset D, into v allotments, comparing to the v results of a test on quality A (Han and Kamber 2001). For every conceivable result, it considers the result

of the quantity of tuples concerning the all-out number of tuples in D. The increase proportion is characterized as

Equation 5:

$$Gain\ ratio\,(A) = \frac{Gain\,(A)}{Split\ info\,(A)}.$$

The parameters are categorized on the basis of three variables. All three are mentioned below:

1. Correlation: Mostly based on their correlation factor, the characteristics are categorized as linked or similar. We have many linked characteristics in the information collection. Now the issue with getting correlated characteristics is that if f1 and f2 are two correlated characteristics of a data set, then the classification or regression model including both f1 and f2 will yield the same as the predictive model compared to the situation where either f1 or f2 was included in the information set. This is because both f1 and f2 are correlated, thus contributing the same data set model information. There are different techniques for calculating the correlation factor, but the correlation coefficient of Pearson is mostly widely used.

2. Entropy: Entropy is the data's mean valuation. The larger the entropy, the higher that function's input to information. The entropy of a function f1 is calculated in Data Science by excluding function f1 and then calculating the entropy of the other characteristics. The reduced the temperature price (excludingf1) the greater the f1 info material. This calculates the entropy of all the characteristics. Ultimately, either a limit valuation or further relevantity test will determine the optimality of the characteristics based on which characteristics are chosen. Entropy is mostly used for Unsupervised Learning as we have a category domain in the dataset and therefore feature entropy can provide significant data.

3. Mutual Information: In information theory, the quantity of ambiguity in X owing to Y awareness is mutual information I(X;Y). Mutual info in scientific data is mostly calculated to understand how much data a function shares about the category. Therefore, Supervised Learning is

mostly used for the decrease of dimensionality. The features in a supervised learning that have a high mutual information value corresponding to the class are considered optimal as they can influence the predictive model towards the correct prediction and thus increase the model's accuracy.

4.4 Analysis of Selected Features

There are total 245 features selected in the feature selection process. There are 74 features of methylation are selected, 75 features of histone, 90 features of nucleotide composition, 4 of the conservation features and rest 2 of the element length. We first studied the connection between the characteristics chosen. Using hierarchical clustering on absolute correlation values between characteristics, we discovered that the chosen characteristics tend to cluster by type of information as anticipated. The conservation characteristics in the coding regions (CDS) are grouped together, for instance, and so are most methylation characteristics. As expected, the promoter's CpG islands are very essential for gene expression prediction, as evidenced by the three chosen and extremely correlated characteristics of CG structure, TSS200 GC, TSS200 CG and TSS200 CGG.

Table 1: Feature Selection

Categorization by Data Type	Number of selected features
Histone	75
Methylation	74
Nucleotide Composition	90
Conservation	4
Element Length	2
Categorization By Gene Location	Number of selected features
TSS200	2
CDS	2
First Exon	87
Full Transcript	87
TSS1500	2
UTR5	2
First Intron	57
Last Exon	2
Last Intron	2
UTR3	2

4.5 Model assessment

The dataset was divided into training and testing sets. The training dataset experienced 10-fold cross validation with different combination of feature selection ratio and classification techniques, so as to get the best model.

The data that has to go through the training and testing phase, we fragmented it into the following parts:

1. 10:90 where 10 percent of data acts as testing data while 90 percent data works as training data
2. 20:80 where 20 percent of data acts as testing data while 80 percent data works as training data
3. 30:70 where 30 percent of data acts as testing data while 70 percent data works as training data
4. 40:60 where 40 percent of data acts as testing data while 60 percent data works as training data.

Chapter 5

Classification Methods used for Prediction of Breast Cancer

5.1 Classification Methods

Classification is regarded as an example of supervised learning, i.e. learning where there is a training set of properly defined observations. The associated unsupervised operation is called clustering and includes grouping information into classifications based on some intrinsic resemblance or distance measure.

The individual observations are often evaluated into a collection of quantifiable characteristics, known as explanatory variables or characteristics in different ways. These may be categorical as ordinal (e.g. "big," "medium" or "small"), whole-evaluated (e.g. amount of telephone occurrences) or real-evaluated (e.g. measurement of blood stress). Other classifiers operate by using a similarity or distance function to compare observations with prior observations.

The efficiency of classification is depicted by scalar attributes such as precision, sensitivity, and specificity in distinct metrics. Comparing different classifiers using these measures is easy, but it has many problems such as the sensitivity to imbalanced data and ignoring the performance of some classes. Different classification efficiency definitions are given by graphical evaluation methods such as receiver working features (ROC) and precision-recall curves.

In our research, we have used eight classification methods. They are SVM (Support Vector Machine), Gaussian SVM, Linear SVM, Logistic Regression, Random Forest, Neural Network, Naive Bayes, and KNN (k-nearest-neighbor).

5.1.1 Gaussian SVM

Kernel methods, such as Support Vector Machines (SVMs), are the most popular classification tools in machine learning and data mining areas, which are based on statistical learning theory and deliver state-of the-art results for non-linear learning problems [34, 37]. They map the input data into the Reproducing Kernel Hilbert Space (RKHS), where one can use linear algorithms to abstract the non-linear relations in the input data. Lifted information depiction may contribute to stronger generalization results, but the need to edit kernel matrices of kernel methods imposes an important computing bottleneck with computational time in O(N)3 and storage costs in O(N)2, where N is the size of the training set. Hence it is difficult for kernel methods to scale up to large-scale problems.

For relatively small-scale problems, for instance, the fault diagnosis of discrete14 event systems, the most important thing is to pursue high accuracy rather than efficiency. While for huge input datasets, it is necessary to improve the computational efficiency, and there are typically two possible ways to proceed. One strategy is to decrease the sample size in a manner that does not drastically alter the outcomes [15, 32, 41].This is valid because in many applications approximate solutions are acceptable, many data may be over generated. Thus, we could subsample examples from the whole data to do classification, regression, clustering, etc.

We have used this classifier because the kernel trick is SVM's true strength. We can fix any complicated issue with a suitable kernel function. Gaussian SVM is not achieved for local optima. It scales comparatively well to high-dimensional statistics. In practice, Gaussian SVM models are generalized, with less risk of over fitting in Gaussian SVM.

The name of the package and all parameters which we have used in our experiment is shown Table 3.

5.1.2 SVM

A Support Vector Machine (SVM) is officially described by a separate hyper plane as a discriminatory classifier. In other words, given the marked training data (supervised learning), an ideal hyper plane is

produced by the algorithm that categorizes fresh instances. This hyper plane is a line dividing a plane into two sections in two dimensional spaces where it lies on either side in each class. The fundamentals of Support Vector Machines and how it functions are best comprehended with a straightforward precedent. We should envision we have two labels: red and blue, and our information has two highlights: x and y. We need a classifier that, given a couple of (x, y) organizes yields if it's either red or blue. We plot our effectively named preparing information on a plane:

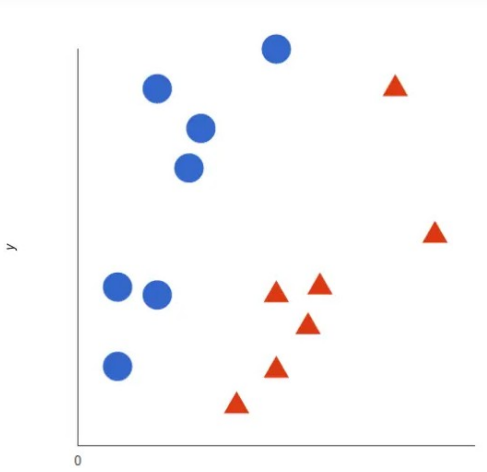


Figure 3: Labeled data for SVM

It help vector machine takes these information focuses and yields the hyper plane (which in two measurements it's basically a line) that best isolates the labels. This line is the choice limit: whatever tumbles to the other side of it we will characterize as blue and anything that tumbles to the next as red.

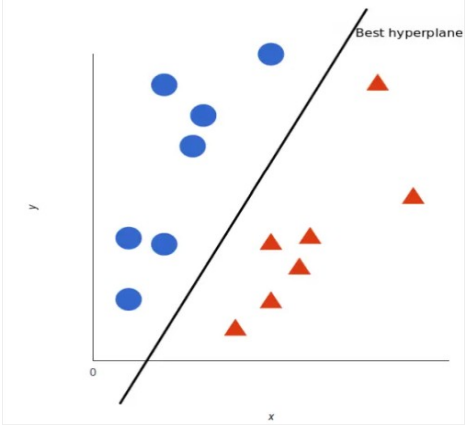


Figure 4: Hyper plane

Be that as it may, what precisely is the best hyper plane? For SVM, the one expands the edges from the two labels. At the end of the day: the hyper plane (recall it's a line for this situation) whose separation to the closest component of each tag is the biggest.

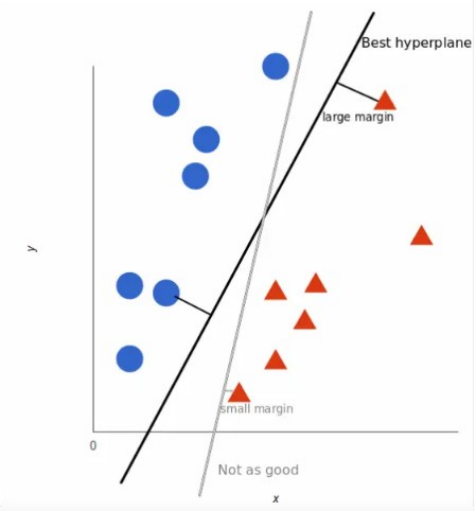


Figure 5: Best hyper plane

Nonlinear data

Presently this model was simple, since unmistakably the information was directly detachable — we could attract a straight line to isolate red and blue. Unfortunately, more often than not things aren't that straightforward. Investigate this case:

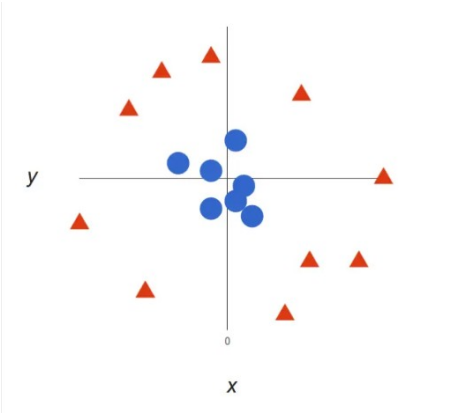


Figure 6: Nonlinear data

It's truly certain that there's not a direct choice limit (a solitary straight line that isolates the two labels).

Be that as it may, the vectors are all around plainly isolated and it looks just as it ought to be anything but difficult to isolate them.

So this is what we'll do: we will include a third measurement. As of not long ago we had two measurements: x and y. We make another z measurement, and we decide that it be determined a specific way that is advantageous for us: $z = x^2 + y^2$ (you'll see that is the condition for a circle).

This will give us a three-dimensional space. Taking a cut of that space, it would appear that this:

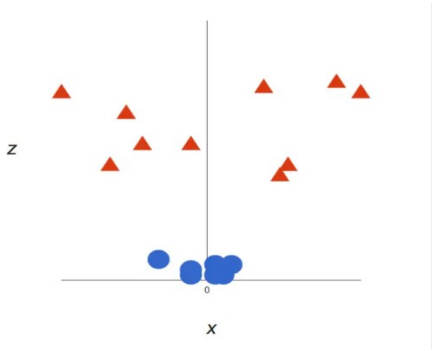


Figure 7: Three dimensional data presentation

The SVM feature we use in the experiment can be found in Weka as a plugin. Yasser EL-Manzalawy created the initial wrapper, called WLSVM. The present version we are using is the wrapper's full rewrite to prevent compilation mistakes. The name of the package and all parameters in Table 3.

We have used this classifier because SVM classifiers offer good accuracy and faster prediction. They also use less memory because in the decision stage they use a subset of learning points. SVM functions well with clear separation margins and high dimensional space.

5.1.3 Linear SVM

To predict one variable from another requires the use of an extension of linear correlation called linear regression. Linear SVM analysis is a “workhorse” in applied statistics. Math is not too complicated, and regression analysis is supported by most software packages. Linear SVM extends the idea of the scatterplot used in correlation and adds a line that best “fits” the data. Because it is an extension of linear correlation, linear SVM models the linear component of the relationship between variables. If the

relationship has no linear component, then the correlation will be close to 0 and the linear regression will have little to no predictive accuracy.

Although there are many ways to draw lines through the data, least-squares analysis is a mathematical approach that minimizes the squared distance between the line and each dot in the scatterplot. This analysis can be done by hand or using software such as Minitab, SPSS, SAS, R, or Excel. We have used liner SVM because Linear SVM's are very useful when we have no concept about the information. It operates well with even unstructured and semi-structured data such as text, images and trees. They also use less memory because in the decision stage they use a subset of learning points. It operates well with clear segregation margins and high dimensional space.

The name of the package and all parameters which we have used in our experiment is shown in Table 3.

5.1.4 Logistic Regression

Logistic regression is an arrangement calculation used to allocate perceptions to a discrete arrangement of classes. Not at all like linear regression which yields consistent number have qualities, calculated relapsed changes its yield utilizing the strategic sigmoid capacity to restore likelihood esteem which would then be able to be mapped to at least two discrete classes.

Different types of logistic regressions:

- Binary (Pass/Fail)
- Multi (Cats, Dogs, Sheep)
- Ordinal (Low, Medium, High)

Binary Logistic Regression

Let's assume we're given information on understudy test results and our objective is to anticipate whether an understudy will pass or bomb dependent on number of hours dozed and hours spent examining. We have two feature (hours dozed, hours considered) and two classes: passed (1) and failed (0). Graphically in a scatter plot, we can represent the data as below:

Table 2: Data of student's result

Studies	Slept	Passed
4.85	9.3	1
8.62	4.5	0
5.43	6.7	1
9.11	6.3	0

Activating Sigmoid

So as to guide anticipated qualities to probabilities, we utilize the sigmoid capacity. The capacity maps any genuine incentive into another incentive somewhere in the range of 0 and 1. In AI, we utilize sigmoid to delineate to probabilities.

$$S(z)=1/(1+e^{-z})$$

Graphical representation:

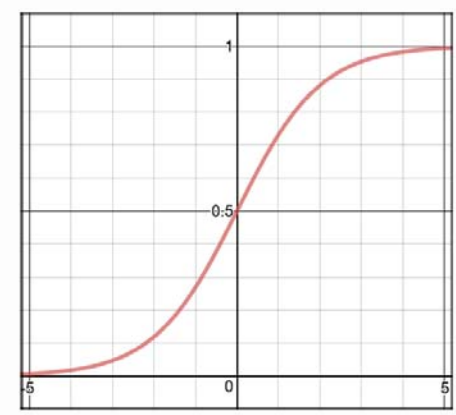


Figure 8: Activation sigmoid

Decision Boundary

Our present forecast capacity restores a likelihood score somewhere in the range of 0 and 1. So as to outline to a discrete class (genuine/false, feline/hound), we select an edge worth or tipping point above which we will order esteems into class 1 and underneath which we arrange values into class 2.

$$p \geq 0.5, class=1$$

$$p < 0.5, class=0$$

For instance, if our edge was .5 and our forecast capacity returned .7, we would arrange this perception as positive. On the off chance that our expectation was .2 we would group the perception as negative. For calculated relapse with different classes we could choose the class with the most

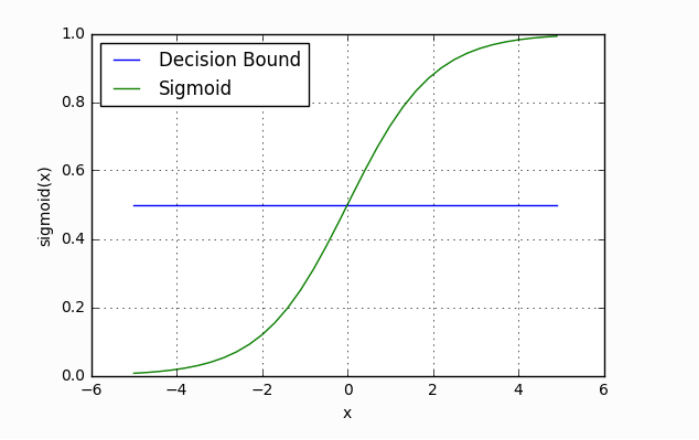


Figure 9: Decision Boundary

Predicting Results

Utilizing our insight into sigmoid capacities and choice limits, we would now be able to compose a forecast capacity. A forecast capacity in calculated relapse restores the likelihood of our perception being sure, true, or "Yes". We call this class 1 and its documentation is P (class=1). As the likelihood draws nearer to 1, our model is increasingly certain that the perception is in class 1.

Equation for the same will be

Equation 6:

$$z = W_0 + W_1 \text{Studied} + W_2 \text{Slept}$$

This time anyway we will change the yield utilizing the sigmoid capacity to restore likelihood esteem somewhere in the range of 0 and 1.

Equation 7:

$$P(class=1)=1/(1+e^{-z})$$

In the event that the model returns .4 it accepts there is just a 40% shot of passing. In the event that our choice limit was .5, we would arrange this perception as “fail”.

We have used this classifier because it is very effective, it does not involve too many computing resources, it is extremely interpretable, it does not require input characteristics to be scaled, it does not involve any tuning, it is simple to regularize, and it produces well-calibrated expected probabilities.

The name of the package and all parameters which we have used in our experiment is shown in Table 3.

5.1.5 Random Forest (RF)

Random forest is a popular machine learning procedure which can be used to develop prediction models. First introduced by Breiman in 2001 (Breiman, 2001), random forests are a collection of classification and regression trees (Breiman, Friedman, Olshen, & Stone, 1984), which are simple models using binary splits on predictor variables to determine outcome predictions. Decision trees are easy to use in practice, offering an intuitive method for predicting outcome which splits “high” versus “low” values of a predictor related to outcome. Though it offers many benefits, decision tree methodology often provides poor accuracy for complex datasets (e.g. large datasets and datasets with complex variable interactions). In the random forest setting, many classification and regression trees are constructed using randomly selected training datasets and random subsets of predictor variables for modeling outcomes. Results from each tree are aggregated to give a prediction for each observation. Therefore, random forest often provides higher accuracy compared to a single decision tree model while maintaining some of the beneficial qualities of tree models (e.g. ability to interpret relationships

between predictors and outcome) (Speiser, Durkalski, & Lee, 2015). Random

Forests consistently offer among the highest prediction accuracy compared to other models in the setting of classification (Fernandez Delgado, Cernadas, Barro, & Amorim, 2014).

A major benefit of using random forest for prediction modeling is the ability to handle datasets with a large number of predictor variables; however, often in practice, the number of predictors required for obtaining outcome predictions should be minimized to improve efficiency. For example, rather than using all variables available in the electronic medical record, one may prefer to use only a subset of the most important variables when developing a medical prediction model. In prediction modeling, an interest is often to determine the most important predictors that should be included in a reduced, parsimonious model. This can be achieved by performing variable selection, in which optimal predictors are identified based on statistical characteristics such as importance or accuracy. Developing prediction models using variable selection may reduce the burden of data collection and may improve efficiency of prediction in practice. Since many modern datasets have hundreds or thousands of possible predictors, variable selection is often a necessary part of prediction model development.

Variable selection in the random forest framework is a relevant consideration for many applications in expert systems and applications. In general, the overall goal of many expert systems is to aid in decision making for a complex problem. This fits the goal of prediction modeling, in which we use a dataset to develop a model (random forest in this study) which will provide predictions of an outcome of interest. To increase efficiency of obtaining model predictions, variable selection may be used in order to identify a subset of predictor variables to be included in a final, simpler model. There are many applications for which this occurs in expert system development, for instance, developing a medical decision support tool, a projection model for stock market prices and a business analytics model to maximize profits. There are several methods available for performing variable selection in the setting of random forest classification. Many R packages provide random forest variable selection procedures, including boruta (Kursa & Rudnicki, 2010), varSelRF (Díaz- Uriarte & De Andres, 2006),

VSURF (Genuer, Poggi, & Tuleau-Malot, 2015), caret (Kuhn, 2008), party (Hothorn, Hornik, Strobl, & Zeileis, 2010), random forest SRC (Ishwaran & Kogalur, 2014), RRF (Deng & Runger, 2013), vita (Janitza, Celik, & Boulesteix, 2015), AUCRF (V. Urrea & M. L. Calle, 2012) and fuzzy Forest (Conn, Ngun, Li, & Ramirez, 2015). Several other methods have been proposed in the literature (e.g. Hapfelmeier (2013), Svetnik (2004), Jiang (2004) and Altmann (2010)). While there are many methods for random forest variable selection for classification problems available, there is a paucity of guidance in the literature about which methods are preferable in terms of prediction error rate (out-of-bag), parsimony (number of variables), computation time and area under the receiver operating curve (AUC) for different types of datasets. Sanchez-Pinto (2018), Degenhardt (2017), Cadenas (2013) and Hapfelmeier (2013) assess variable selection methods for random forest classification, but most of these papers compare only a handful of methods. Additionally, these papers are limited in scope due to the use of synthetic simulated data which are not always representative of real-world datasets or a small number of application datasets. A final limitation of the current random forest variable selection literature is that computation times for different procedures are rarely reported. Given these limitations, there is a need to compare variable selection procedures for a large number of random forest classification problems in order to provide recommendations about which procedures are appropriate for different types of datasets.

We have used this classifier because Random forest is regarded as an extremely precise and stable technique due to the amount of decision trees involved in the process. Random forest prevents over-fitting by creating trees on random subsets. The primary reason is that it requires the average of all projections, which cancels the biases. Random forests can manage missing values as well. The parameters we use in the experiment is available in Weka and are given in Table 3.

5.1.6 Neural Network

In information technology (IT), a neural system is an arrangement of equipment as well as

programming designed after the activity of neurons in the human cerebrum. Neural systems - additionally called fake neural systems - are an assortment of profound learning innovation, which likewise falls under the umbrella of man-made brainpower, or AI. Business uses of these advances by and large spotlight on illuminating complex sign preparing for example acknowledgment issues. Instances of huge business applications since 2000 incorporate penmanship acknowledgment for check preparing, discourse to-content interpretation, oil- investigation information examination, climate expectation and facial acknowledgment.

A neural system more often than not includes an enormous number of processors working in parallel and masterminded in levels. The principal level gets the crude info data - undifferentiated from optic nerves in human visual preparing. Each progressive rate receives the output from the stage before it, rather than from the crude details – likewise, cells beyond the optic cell receive signals from those nearer to it. The last level creates the yield of the framework.

Each handling hub has its own little circle of learning, including what it has seen and any standards it was initially modified with or created for itself. The levels are much interconnected, which means every hub in level n will be associated with numerous hubs in level n-1 - its sources of info - and in level n+1, which gives contribution to those hubs. There might be one or various hubs in the yield layer, from which the appropriate response it produces can be perused.

Neural systems are prominent for being versatile, which means they change themselves as they gain from starting preparing and consequent runs give more data about the world. The most essential learning model is fixated on weighting the info streams, which is the way every hub loads the significance of contribution from every one of its ancestors. Sources of info that add to finding right solutions are weighted higher.

Neural systems are in some cases portrayed regarding their profundity, including what number of layers they have among information and yield, or the model's supposed shrouded layers. This is the reason the term neural system is utilized synonymously with profound learning. They can likewise

be portrayed by the quantity of concealed hubs the model has or as far as what number of sources of info and yields every hub has. Minor departure from the great neural system configuration permit different types of forward and in reverse proliferation of data among levels.

The least complex variation is the feed-forward neural system. This sort of counterfeit neural system calculation goes data straight through from contribution to handling hubs to yields. It could conceivably have concealed hub layers, making their working increasingly interpretable.

Increasingly unpredictable are repetitive neural systems. These profound learning calculations spare the yield of handling hubs and feed the outcome over into the model. This is the means by which the model is said to learn.

Convolutional neural systems are main stream today, especially in the domain of picture acknowledgment. This particular kind of neural system calculation has been utilized in a large number of the most progressive uses of AI including facial acknowledgment, content digitization and normal language preparing.

We have used this classifier because neural networks are more flexible. Neural networks are good for the nonlinear dataset with a large number of inputs such as images. Neural networks can work with any number of inputs and layers. Neural networks have the numerical strength that can perform jobs in parallel.

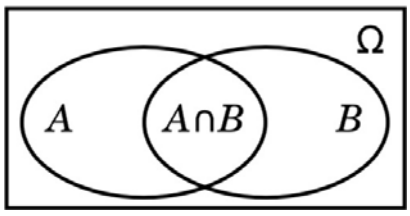
We use the ' MultilayerPerceptron ' feature in Weka for the experiment. Detailed parameters are specified in Table 3.

5.1.7 Naive Bayes

For inferential statistics and many sophisticated machine learning designs, Bayes ' theorem is of basic significance. Bayesian reasoning is a logical approach to updating the likelihood of hypotheses in the light of fresh proof, and thus plays a crucial position in science (Berry, 1996). The Bayesian analysis enables us to answer questions about frequency statistical methods.

A statistical test can be widely described as a method leading to one outcome and only one outcome of several. The sample space, marked by O, is called the set of all feasible results. We can portray occurrences at the introductory stage using notation from set theory. For instance, one roll of a reasonable die may be our test. The sample space is then O={1, 2, 3, 4, 5, 6}, which in the terminology of set theory is also referred to as universal. For example, a straightforward occurrence is result 2, which we call E1= {2}. P (E) denotes the likelihood of a case E. According to the classic concept of possibility, the likelihood of an event E is the number of outcomes that are favorable to this event, divided by the total number of possible results for the experiment.

Let A and B be two events from a sample space O, either finite or countlessly infinite with N elements. Let P: O —[0, 1] be a range of likelihood on O, so0 < P(A) < 1 and 0 < P(B) < 1 and clearly P(O)= 1. In a Venn diagram, we can depict these occurrences (Fig. 9(a)). The association of A and B occasions, intended by A, B, is an opportunity that occurs either A or B or both. The crossing point for A and B occasions are the opportunity for both A and B to occur. Lastly, two occasions, A and B, are called essentially unrelated if one of these occasions ' case precludes the probability of the other occasion's case. Two A and B occurrences, P(A) >0 and P(B) > 0, are called independent if the occurrence of one event does not affect the probability of occurrence of the other event, i.e., P(A | B)= P(A) or P(B | A) = P(B), and P(A - B) = P(A). P (B).



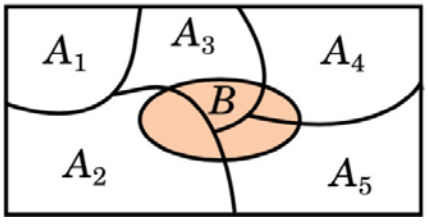


Figure:10 (a) Venn diagram for sets A and B **(b) Illustration of the total probability**

Consider a population of people in which 1% really has a disease, D. A medical screening test is

applied to 1000 randomly selected persons from that population. It is known that the sensitivity of the test is 0.90, and the specificity of the test is 0.91.

(a) If a tested person is really sick, then what is the probability of a positive test result (i.e., the result of the test indicates that the person is sick)?

(b) If the test is positive, then what is the probability that the person is really sick?

The probability that a randomly selected person has the disease is given as P (D) = 0.01 and therefore P (D^c) = 0.99. These are the marginal probabilities that are known a priori, that is, without any knowledge of the person's test result. The sensitivity of a test is defined as TP/TP+FN where TP denotes the number of true positive predictions and FN denotes the number of false negative predictions. Sensitivity is therefore also known as true positive rate; in information retrieval and data mining, it is also called recall. The specificity of a test is defined as TN/TN+FP' where TN denotes the number of true negative predictions and FP denotes the number of false positive predictions. Therefore, the response to (a) is straightforward – indeed, it is already provided: the conditional likelihood P (') is the same as the awareness, since the amount of people who are really ill is the same as the amount of real favorable projections (people are ill and are properly recognized by the experiment) plus the amount of fake adverse projections (people are ill but they are not recognized as such).The details of parameters which we have used are given in Table 3.

We have used this method because it is not only a simple approach but also a fast and accurate prediction method. Naive Bayes has very low computation cost. It can operate effectively on a big dataset. It works well in case of discrete response variable relative to the continuous variable. It can be used with various category prediction problems. It also works well in case of text analytics issues.

5.1.8 K-Nearest-Neighbor Algorithm

KNN classification aims at classifying imbalanced data and was selected as top 10 data mining algorithms (Wu et al. 2008; Zhang et al. 2017; Zhang et al. 2018b; Zheng et al. 2017).There are two main research directions. One is to set a proper K value. Another is the distance function for identifying K nearest neighbors. For setting K value, a usually-used method is the cross-validation in probability theory. It is useful for identifying a proper K value when a training dataset is given. However, training samples are distributed with different densities in the training sample space. This raises a new challenging issue that different samples need different K values for class prediction. Recently, Cheng, et al (2014) studied the computation of parameter K for KNN classification, which is an optimal value for each new data. Zhang, et al (2018b) designed a KNN algorithm with data-driven K parameter computation. Zhang, et al (2017) designed an algorithm to efficiently learn K for KNN Classification. Although there are many distance functions, most KNN classification algorithms use Euclidean distance which is defined as follows.

Equation 8:

$$d(X,Y)=\sqrt{\sum_{i=1}^{N}(Xi-Yi)^2}$$

Where Xi and Yi (i = 1 … N), respectively, are an attribute of two samples/instances X and Y.

Some lately research reports, for example, Deng, et al (2016) pioneered KNN method to classify big data. It first conducts a k-means clustering to separate the whole dataset into several parts. And then, each subset is classified with KNN method. Liu, et al (2016) proposed a neighbor selection for multilabel classification. Liu and Zhang (2012) studied a noisy data elimination using the mutual KNN for data classification. Zhang (2010) proposed to replace the majority rule with CF measure for KNN classification. This leads to a minor class can be become a winner. Zhang (2011) studied a shell-neighbor method for KNN classification. It assists in learning from datasets with missing values. Zhang, et al (2016) proposed a self-representation nearest neighbor search for data classification. The authors proposed representing each sample by other samples with a new self- reconstruction method.

The obtained coefficient is used to compute the value of K for every sample, rather than all samples used in the traditional methods (Zheng et al., 2017; Lei and Zhu 2017; Zhu et al. 2018a). Finally, this literature proposed builds a decision tree with the obtained value of K in the leaf to output the labels of training samples. KNN classifiers are lazy learners, which is time consuming since the distance between every test sample and other samples should be calculated. To deal with this issue, Zhang et al (2018) pioneered a K-tree and a k*Tree to use different numbers of nearest neighbors for KNN classification. The K-tree method needs less running cost but achieves similar classification accuracy, compared with those KNN methods that assign different K values to different test samples. The technique k*Tree is a K-tree expansion. It speeds up its experiment phase by additional storing data from the coaching samples in K-tree's leaf nodes, such as the coaching samples in the leaf clusters, their KNNs, and those KNN's closest neighbor. It makes KNN only using a subset of training samples in the leaf nodes. This is different from previous methods, e.g., (Zhu et al. 2014; Zheng et al. 2018), which use KNN method to visit all samples. Therefore, our proposed method may decrease the computation cost of the test process. We have used this classifier because the K-nearest neighbor classification learning stage is much quicker relative to other classification algorithms. There is no need to train a generalization model, which is why KNN is recognized as the easy, instance-based learning algorithm. KNN can be helpful for nonlinear data. The output value for the object is calculated by the average value of the nearest k neighbors.

We have use K=1 and the Euclidean Distance in our experiment. Details of parameters are given in Table 3. This Table 3 shows the list of parameters and Weka function used in our research.

Table 3: Parameters

Classifier	Weka Function	Parameters
SVM	Lib SVM	-S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -B -model /Users/Stanley -seed 1
RF	Random Forest	-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
NN	Logistic	-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
KNN	IBK	-K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Naïve Bayes	Bayes	-P 0 –M 3.0 –norm 1.0 –Inorm2.0 –stopwords-handler weka.core.stopwords
Gaussian SVM	Lib SVM	-S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -B -model /Users/Stanley -seed 1
Linear SVM	SMO	-C 1.0 –L 0.001 –P 1.0E-12 –N 0 –V -1 –W 1 -K
Logistic Regression	Logistic	-R 1.0E-8 –M -1 –num-decimal-places 4

5.2 Tools

Data processing is done using R-Script whereas; classification, as well as feature selection, is done using data mining software - Weka.

5.2.1 R tool

R is a language and condition for measurable registering and designs. It is a GNU venture which is like the S language and condition which was created at Bell Laboratories (once in the past AT&T, presently Lucent Technologies) by John Chambers and associates. R can be regarded as an alternative use of S. There are some significant contrasts, yet much code composed for S runs unaltered under R.

R gives a wide assortment of factual (straight and nonlinear demonstrating, old style measurable tests, time-arrangement investigation, order, bunching) and graphical strategies, and is profoundly extensible. The S language is often the medium of decision-making in measurable strategy for studies, and R provides assistance for an open source course in this intervention.

One of R's qualities is the straightforwardness with which well-planned distribution quality plots can be delivered, including numerical images and formulae where required. Extraordinary consideration has been assumed control over the defaults for the minor plan decisions in illustrations; however the client holds full control.

R is accessible as Free Software under the details of the Free Software Foundation's GNU General Public License in source code structure. It accumulates and keeps running on a wide assortment of UNIX stages and comparative frameworks (counting FreeBSD and Linux), Windows and MacOS.

5.2.2 Weka

Weka is a set of machine-learning algorithms for information mining functions. It includes instruments for preparing, classifying, regressing, clustering, mining association rules and visualizing information. Found only on New Zealand's islands, the Weka is a curious-looking flightless bird. Such is the name pronounced, and the bird smells like this. Weka is GNU General Public License open source software. Weka encourages deep learning as well.

5.3 Feature Selection Ratio

Selection of features is split into two components:

a. Evaluator Search

b. Method attributes.

Each chapter has several methods to choose from.

The evaluator function is the method used to evaluate each object in your dataset (also known as a row or function) in the event of the input vector (e.g. category). The query method is the technique by which distinct mixes of characteristics can be tried or navigated in the dataset to reach a brief range of selected characteristics.

Some methods of Attribute Evaluator involve particular search methods to be used.

More the ratio of feature selection, more irrelevant data is eradicated.

In our research we have used percentiles as follows:

a. Raw where whole dataset has been used for feature selection

b. 95%

c. 90%

d. 85%

e. 80%

f. 75%

g. 70%

h. 65%

i. 60%

j. 55%

k. 50%

In each and every part, the results differed.

5.4 10 fold cross validation

Cross-validation is a method for frequent holdout to improve. Cross-validation is a comprehensive method to do continuous holdout that effectively enhances it by decreasing the estimation variance. We take a training set and a classifier is created. Then we're looking to assess that classifier's efficiency, and there's a certain level of variance in that assessment because it's all underneath the statistics. We want to maintain the difference as small as feasible in the assessment. And cross validation also prevents the over fitting.

Cross-validation is a method to reduce the variance, and it is further reduced by a cross-validation version called "stratified cross-validation".

We split it only once with cross-validation, but we split it into, say, 10 parts. Then we bring 9 of the parts and use them to train, and we use the last item to test. Then we bring another 9 parts with the same separation and use them for practice and experimentation with the hold-out item. We do the whole process 10 occasions, each moment we use a distinct section to test. In other cases, we split the dataset into 10 parts, and then we keep each piece in turn for monitoring, training on the remainder, monitoring, and averaging the 10 outcomes. That would be a "cross-validation of 10 times."

Divide the dataset into 10 components, keep each portion in turn, and evaluate the outcomes. Therefore, each data point in the dataset is used for testing once and for training nine times. That's a cross-validation of 10 times.

5.5 Ratio Comparison

Comparison of ratio is the ratio taken using cross validation. In

our research, we have used ratios as follows:

1. **90:10** where 90% is training set and 10% is testing
2. **80:20** where 80% is training set and 20% is testing
3. **70:30.** where 70% is training set and 30% is testing
4. **60:40.** where 60% is training set and 40% is testing

5.6 Methodology

Figure 11 illustrates the work flow of data analysis of breast cancer. Four different types of datasets, namely CPG Methylation Data, Histone Marker Modification Data, Human Genome Data, RNA-Seq Data is used for data analysis. The next step after downloading the data is feature extraction, useful features are extracted from each dataset. These features are useful for better prediction of breast cancer as well as to reduce the complexity of dataset. By using transcript id of each dataset, we have combined all four datasets and created model using R tool, which includes all extracted features with most promising number of genes. On this model we have applied four different feature selection techniques like PCA, CFS, Gain ratio, ReliefF. We have used different feature selection ratios (Raw, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, 50%) with different training-testing ratios (90:10, 80:20, 70:30, 60:40) with the combination of 8 different classifiers (Gaussian SVM, Linear SVM, KNN, Naïve Bayes, Random forest, SVM, Logistic regression and Multi-Layer Perceptron). We have also used 10-fold cross validation in the research.

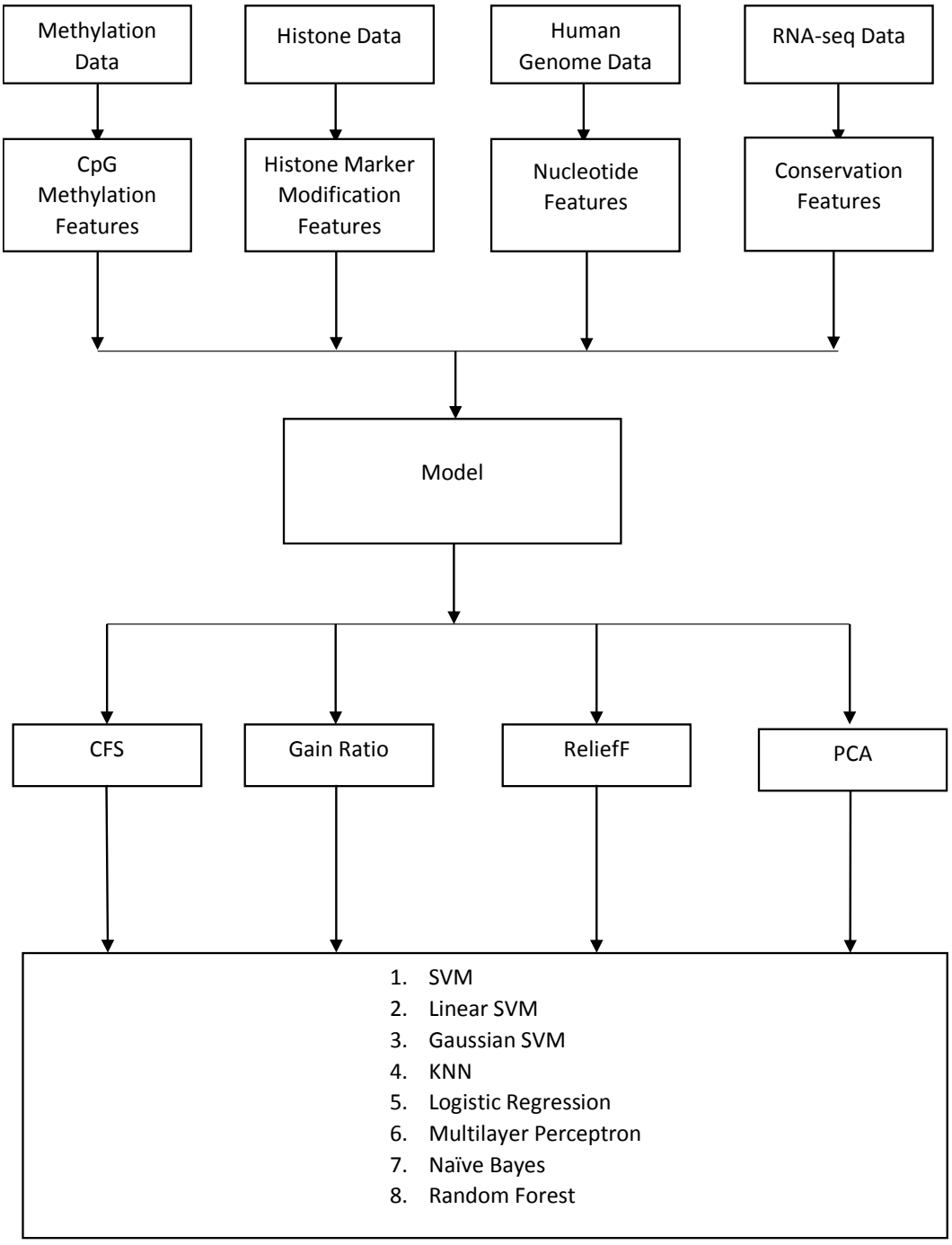


Figure 11: Work Flow of Data Analysis

Chapter 6

Results and Discussion

6.1 Results

Four distinct types of data were used to extract the features of cancerous cells and four different feature selection methods were used for distinguishing the features. The percentage ratio for feature selection is raw, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, and 50%. Further, the results are concluded using eight different classifiers. 10 cross-fold validations are done for different training and testing ratios. (90:10, 80:20, 70:30, 60:40)

The results are highlighted with different colour codes. The light pink colour indicates the highest accuracy in each row of the classifier, middle pink colour is for the highest accuracy in each feature selection % from different ratios and the dark pink colour cell is the highest accuracy of the sheet.

Similarly, light blue colour cells have the highest AUC in each row of the classifier, the middle blue colour is for the highest accuracy in each feature selection % from various ratios and the dark blue colour cell is the highest AUC of the sheet.

The Table 4 is the data analysis of breast cancer dataset by CFS feature selection method. In the table, light pink colour indicates the row-wise highest accuracy for each of the eight classifiers used, middle pink colour is for the highest accuracy of CFS data in each feature selection % taken for different ratios and the dark pink colour cell is the highest accuracy out of the entire breast cancer data analysis.

Then, the shades of blue colour indicate the AUC for CFS data. For each classifier, light blue colour is the highest accuracy; the middle blue colour is for the highest accuracy in each feature selection % done for different ratios and the dark blue colour is the highest AUC of the complete result analysis of breast cancer data.

Table 4: CFS results for individual ratio of dataset

CFS									
Ratio		**09:01**		**08:02**		**07:03**		**06:04**	
Method		**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
GAUSSIAN SVM'	Raw	0.87	8.9	0.93	9.26	0.62	8.8	0.92	8.18
LINEAR SVM'		0.99	8.22	0.85	7.8	0.93	9.84	0.82	7.89
LOGISTIC REGRESSION'		0.71	6.8	0.87	7.22	0.67	6.78	0.63	9.06
NAVIE BAYES'		0.81	7.47	0.93	9.02	0.61	8.88	0.96	7.51
RANDOM FOREST'		0.98	8.36	0.71	9.2	0.91	7.91	0.7	9.32
KNN'		0.66	6.49	0.87	8.91	0.82	6.84	0.79	9.7
SVM'		0.92	7.45	0.61	8.65	0.73	8.2	0.73	9.17
NEURAL NETWORK'		0.64	8.54	0.82	9.16	0.83	8.09	0.91	7.63
GAUSSIAN SVM'	95%	0.96	7.06	0.81	8.27	0.95	9.57	0.92	9.01
LINEAR SVM'		0.81	7.05	0.95	7.62	0.86	8.8	0.92	6.56
LOGISTIC REGRESSION'		0.99	8.85	0.67	6.45	0.77	7.43	0.95	8.64
NAVIE BAYES'		0.91	8.47	0.98	8.51	0.82	8.78	0.63	6.84
RANDOM FOREST'		0.89	8.41	1	7.06	0.67	8.74	0.87	6.63
KNN'		0.6	9.89	0.73	7.14	0.69	9.64	0.8	6.64
SVM'		0.68	8.68	0.85	6.98	0.68	9.59	0.73	6.54
NEURAL NETWORK'		0.7	7.6	0.81	8.24	0.7	7.64	0.84	9.85
GAUSSIAN SVM'	90%	0.63	8.3	0.78	6.51	0.98	7.97	0.74	8.31
LINEAR SVM'		0.9	9.52	0.74	8.12	0.86	9.31	0.68	9.14
LOGISTIC REGRESSION'		0.93	9.98	0.73	6.61	0.93	9.73	0.7	8.54
NAVIE BAYES'		0.74	9.43	0.63	8.02	0.75	9.53	0.64	9.08
RANDOM FOREST'		0.65	7.68	0.76	7.03	0.66	8.22	0.66	7.58
KNN'		0.87	9.11	0.69	8.48	0.9	8.53	0.73	9.46
SVM'		0.94	9.79	0.79	9.36	0.95	9.74	0.84	8.2

NEURAL NETWORK'		0.87	8.82	0.88	6.75	0.74	7.04	0.66	6.85
GAUSSIAN SVM'	85%	0.65	8.16	0.65	7.27	0.67	7.78	0.62	9.34
LINEAR SVM'		0.75	7.46	0.93	7.32	1	7.37	0.6	8.82

LOGISTIC REGRESSION'		0.87	8.56	0.98	7.7	0.99	9.4	0.91	7.8
NAVIE BAYES'		0.81	7.64	0.93	7.9	0.66	6.69	0.87	9
RANDOM FOREST'		1	9.86	0.79	6.85	0.98	7.45	0.64	7.94
KNN'		0.7	8.02	0.87	7.89	0.77	8.42	0.89	6.64
SVM'		0.95	9.87	0.78	8.85	0.69	8.65	0.66	8.53
NEURAL NETWORK'		0.75	8.85	0.77	8.38	0.65	7.9	0.64	7.51
GAUSSIAN SVM'	80%	0.65	8.96	0.61	7.72	0.7	7.37	0.87	7.9
LINEAR SVM'		0.63	9.85	0.98	8.8	0.82	7.93	0.65	8.53
LOGISTIC REGRESSION'		0.9	9.79	0.61	8.86	0.63	7.81	0.62	7.57
NAVIE BAYES'		0.75	9.41	0.79	6.91	0.84	6.79	0.65	6.97
RANDOM FOREST'		0.75	7	0.75	8.01	0.61	6.91	0.6	9.97
KNN'		0.72	9.4	0.81	9.78	0.78	8.62	0.67	6.83
SVM'		0.72	8.1	1	8.47	0.73	8.78	0.62	7.63
NEURAL NETWORK'		0.7	7.51	0.97	8.5	0.98	9.44	0.61	6.58
GAUSSIAN SVM'	75%	0.79	8.31	0.6	9.17	0.72	8.22	0.72	9.73
LINEAR SVM'		1	9.41	0.92	7.32	0.77	7.54	0.81	7.77
LOGISTIC REGRESSION'		0.75	8.36	0.79	8.92	0.87	7.05	0.76	9.97
NAVIE BAYES'		0.98	7.37	0.75	7.49	0.75	6.62	1	8.38
RANDOM FOREST'		0.66	7.71	0.93	8.48	0.65	7.25	0.78	8.45
KNN'		0.97	7.11	0.95	8.7	0.63	7.75	0.66	7.11
SVM'		0.9	7.69	1	7.87	0.72	8.47	0.91	7.21
NEURAL NETWORK'		0.95	7.12	0.68	7.33	0.93	8.47	0.9	8.17
GAUSSIAN SVM'	70%	0.84	9.42	0.95	9.77	0.85	9.22	0.63	7.23

LINEAR SVM'		0.64	7.81	0.72	8.04	0.68	8.55	0.79	9.78
LOGISTIC REGRESSION'		0.78	6.48	0.76	8.34	1	6.84	0.72	7.43
NAVIE BAYES'		0.86	7.55	0.93	8.71	0.77	8.73	0.76	9.84
RANDOM FOREST'		0.62	9.42	0.75	8.82	0.94	6.7	0.77	9.35
KNN'		0.6	7.45	0.77	8.01	0.82	7.75	0.95	7.69
SVM'		0.71	9.23	0.63	7.06	0.61	7.47	0.88	6.52
NEURAL NETWORK'		0.98	8.34	0.87	8.92	0.6	9.08	0.6	7.66
GAUSSIAN SVM'	65%	0.81	6.99	0.89	9.74	0.92	6.68	0.61	7.81
LINEAR SVM'		0.95	8.31	0.72	7.76	0.78	9.14	0.69	6.78
LOGISTIC REGRESSION'		0.75	7.45	0.73	8.72	0.71	7.3	0.84	9.9
NAVIE BAYES'		0.96	8.99	0.9	9.61	0.95	7.27	0.97	7.12
RANDOM FOREST'		0.81	9.06	1	8.16	0.86	6.49	0.73	7.03
KNN'		0.96	8.05	0.88	8.69	0.71	6.92	0.78	8.25
SVM'		0.64	7.09	0.63	8.1	0.86	7.19	0.64	8.25
NEURAL NETWORK'		0.68	6.93	0.79	6.83	0.7	7.21	0.89	6.63
GAUSSIAN SVM'	60%	0.88	7.42	0.86	6.85	0.67	9.67	0.89	7.75
LINEAR SVM'		0.74	8.71	0.81	6.83	0.75	7.43	0.92	8.53
LOGISTIC REGRESSION'		0.6	9.32	0.65	8.03	0.82	6.79	0.98	9.99
NAVIE BAYES'		0.75	8.82	0.71	9.79	0.97	9.23	0.61	7.99
RANDOM FOREST'		0.73	7.47	1	8.7	0.9	8.82	0.92	8.92
KNN'		0.76	6.95	0.92	9.75	0.66	8.63	0.79	6.5
SVM'		0.67	7.32	0.98	6.67	0.6	7.67	0.71	6.99
NEURAL NETWORK'		0.68	8.91	0.74	7.54	0.65	8.9	0.61	9.65
GAUSSIAN SVM'	55%	0.84	6.7	0.68	9.83	0.97	7.28	0.92	6.61
LINEAR SVM'		0.84	8.22	0.87	6.48	0.87	7.54	0.96	6.55
LOGISTIC REGRESSION'		0.67	9.96	0.74	7.04	0.64	7.72	0.91	8.49
NAVIE BAYES'		0.98	9.67	0.77	9.13	0.99	7.22	0.94	7.66

RANDOM FOREST'		0.62	8.49	0.97	9.41	0.64	8.86	0.74	7.23
KNN'		0.68	9.17	0.94	8.6	0.8	9.07	0.67	9.66
SVM'		0.8	9.1	0.76	8.29	0.74	6.48	0.75	8.88
NEURAL NETWORK'		0.91	8.92	0.9	9.8	0.79	7.71	0.7	7.77
GAUSSIAN SVM'	50%	0.65	8.86	0.71	9.08	0.62	7.3	0.6	7.87
LINEAR SVM'		0.98	7.45	0.71	8.94	0.73	9.75	0.72	7.17
LOGISTIC REGRESSION'		0.77	8.52	0.72	9.1	0.7	9.16	0.69	7.29
NAVIE BAYES'		0.62	8.65	0.61	9.56	0.7	7.8	0.68	9.59
RANDOM FOREST'		0.87	9.34	0.97	9.88	0.98	8.89	0.74	9.72
KNN'		0.71	6.5	0.88	9.94	0.61	8.93	0.82	9.23
SVM'		0.91	7.21	0.69	9.72	0.97	8.55	0.65	7.31
NEURAL NETWORK'		0.99	9.96	0.8	7.45	0.75	9.87	0.67	8.52

The Table 5 is the data analysis of breast cancer dataset by Gain Ratio feature selection method. In the table, light pink colour indicates the row-wise highest accuracy for each of the eight classifiers used, middle pink colour is for the highest accuracy of Gain Ratio data in each feature selection % taken for different ratios and the dark pink colour cell is the highest accuracy out of the entire breast cancer data analysis.

Then, the shades of blue colour indicate the AUC for Gain Ratio data. For each classifier, light blue colour is the highest accuracy; the middle blue colour is for the highest accuracy in each feature selection % done for different ratios and the dark blue colour is the highest AUC of the complete result analysis of breast cancer data.

Table 5: Gain Ratio results for individual ratio of dataset

Gain Ratio									
Ratio		09:01		08:02		07:03		06:04	
Method		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
GAUSSIAN SVM'	RAW	0.94	7.84	1	7.63	0.61	8.39	0.99	7.7
LINEAR SVM'		0.69	7.58	0.63	8.56	0.97	7.67	0.97	6.8
LOGISTIC REGRESSION'		0.9	7.87	0.79	9.65	0.9	8.74	0.78	7.8
NAVIE BAYES'		0.71	9.57	0.88	9.45	0.62	9.76	0.68	9.93
RANDOM FOREST'		0.73	9.84	0.98	9.54	0.89	8.54	0.75	8.63
KNN'		0.73	6.46	0.85	6.81	0.71	7.37	0.9	8.5
SVM'		0.62	8.59	0.85	9.25	0.99	9.65	0.96	8.05
NEURAL NETWORK'		0.63	9.24	0.86	7.89	0.66	6.51	0.95	8.03
GAUSSIAN SVM'	95%	0.8	8.72	0.66	7.54	0.84	7.26	0.99	7.9
LINEAR SVM'		0.75	6.86	0.7	6.54	0.79	9	0.66	9.64
LOGISTIC REGRESSION'		0.77	8.97	0.82	8.97	0.76	6.8	0.65	7.93
NAVIE BAYES'		0.85	6.79	0.69	9.45	0.96	9	0.67	8.87
RANDOM FOREST'		0.87	8.53	0.8	6.62	0.69	7.3	0.7	6.57
KNN'		0.86	9.46	0.92	7.32	0.82	7.91	0.89	7.08
SVM'		0.73	7.27	0.87	9.83	0.98	7.48	0.69	6.78
NEURAL NETWORK'		0.9	7.14	0.61	6.84	0.86	6.98	0.84	6.89
GAUSSIAN SVM'	90%	0.88	6.58	0.68	9.86	0.61	7.55	0.67	7.12
LINEAR SVM'		0.84	9.39	0.76	7.67	0.75	8.7	0.83	8.1
LOGISTIC REGRESSION'		0.96	6.55	0.97	9.58	0.95	6.53	0.78	7.06
NAVIE BAYES'		0.99	7.75	0.84	7.43	0.75	8.67	0.96	8.3
RANDOM FOREST'		0.95	7.07	0.68	8.99	0.9	7.16	0.6	8.85
KNN'		0.73	8.24	0.74	6.62	0.9	7.86	0.75	9.54
SVM'		0.87	8.21	0.7	6.83	0.69	8.1	0.66	9.14
NEURAL NETWORK'		0.97	9.52	0.75	9.5	0.97	9.29	0.63	9.95
GAUSSIAN SVM'	85%	0.88	8.48	0.96	7.96	0.67	7.23	0.96	7.23
LINEAR SVM'		1	7.05	0.82	8.45	0.94	7.81	0.77	7.06
LOGISTIC REGRESSION'		0.8	9.68	0.73	9.05	0.76	9.67	0.82	9.73
NAVIE BAYES'		0.71	8.81	0.86	9.87	0.91	6.89	0.63	9.65
RANDOM FOREST'		0.88	8.23	0.85	6.66	0.73	6.61	0.84	7.79
KNN'		1	9.37	0.63	6.87	0.96	8.9	0.76	6.93
SVM'		0.68	7.47	0.75	6.91	0.95	7.51	0.85	6.95
NEURAL NETWORK'		0.77	9.58	0.98	9.3	0.65	9.99	0.61	6.47
GAUSSIAN SVM'	80%	0.66	8.34	0.97	7.02	0.87	7.1	0.61	9.32

LINEAR SVM'		0.7	6.77	0.7	8.32	0.66	8.42	0.78	9.38
LOGISTIC REGRESSION'		0.76	8.17	0.7	8.97	0.76	8.55	0.93	9.7
NAVIE BAYES'		0.73	7.59	0.91	9.97	0.82	8.22	0.88	7.69
RANDOM FOREST'		0.98	8.38	0.76	8.62	0.82	8.09	0.82	8.07
KNN'		0.78	9.85	0.96	7.48	0.91	6.99	0.68	8.74
SVM'		0.81	7.75	0.78	8.41	0.9	7.37	0.86	8.84
NEURAL NETWORK'		0.8	6.54	0.8	8.37	0.97	9.02	0.61	9.92
GAUSSIAN SVM'	75%	0.93	9.08	0.75	9.65	0.79	9.99	0.89	8.63
LINEAR SVM'		0.92	7.6	0.75	6.7	0.76	8.42	0.77	9.75
LOGISTIC REGRESSION'		0.64	8.81	0.78	9.95	0.98	8.74	0.81	9.75
NAVIE BAYES'		0.85	8.22	0.89	9.67	0.63	8.87	0.97	10
RANDOM FOREST'		0.91	9.5	0.89	9.79	0.82	9.84	0.9	7.4
KNN'		0.71	9.11	0.73	8.84	0.73	7.75	0.99	9.81
SVM'		0.71	7.86	0.79	9.03	0.98	8.12	0.91	6.84
NEURAL NETWORK'		0.79	8.28	0.75	6.78	0.85	6.9	0.77	9.54
GAUSSIAN SVM'	70%	0.75	8.52	0.62	6.93	0.79	6.73	0.78	6.83
LINEAR SVM'		0.79	8.25	1	9.33	0.96	8.38	0.8	9.64
LOGISTIC REGRESSION'		0.77	9.06	0.84	9.45	0.9	8.84	0.69	8.71
NAVIE BAYES'		0.7	9.04	0.78	6.81	0.84	9.45	0.78	8.58
RANDOM FOREST'		0.65	7.14	0.99	9.14	0.64	8.58	0.65	8.11
KNN'		0.7	8.06	0.88	7.59	0.6	6.67	0.81	8.27
SVM'		0.74	8.29	0.8	7.29	0.63	9.07	0.76	8.38
NEURAL NETWORK'		0.7	6.98	0.74	9.43	0.78	6.99	0.62	8.9
GAUSSIAN SVM'	65%	0.6	9.23	0.77	7.95	0.67	6.75	0.74	8.45
LINEAR SVM'		0.78	7.03	0.68	8.79	0.63	6.62	0.64	7.17
LOGISTIC REGRESSION'		0.97	8.21	0.82	6.57	0.73	6.73	0.85	7.78
NAVIE BAYES'		0.86	7.37	0.99	9.08	0.69	7.2	0.69	8.69
RANDOM FOREST'		0.65	8.66	0.67	10	0.73	7.83	0.83	8.28
KNN'		0.7	8.07	0.91	7.81	0.82	7.47	0.73	6.9
SVM'		0.92	7.35	0.94	8.33	0.99	6.98	0.9	9.4
NEURAL NETWORK'		0.89	8.35	0.97	9.24	0.69	8.7	0.92	9.43
GAUSSIAN SVM'	60%	0.93	6.77	0.85	7.02	0.76	6.85	0.64	7.6
LINEAR SVM'		0.98	8.44	0.84	9.65	1	9.06	0.77	7.01
LOGISTIC REGRESSION'		0.67	9.48	0.91	7.55	0.78	9.61	0.65	9.5
NAVIE BAYES'		0.81	9.03	0.71	8.86	1	8.02	0.64	7.58
RANDOM FOREST'		0.71	8.36	0.6	8.08	0.8	7.9	0.81	7.4

KNN'		0.76	8.84	0.71	8.81	1	7.92	0.67	9.51
SVM'		0.69	9.51	0.81	8.88	0.82	9.57	0.89	9.3
NEURAL NETWORK'		0.74	8.41	0.69	7.31	0.65	8.32	0.61	7.38
GAUSSIAN SVM'	55%	0.66	9.69	0.97	8.38	0.73	6.9	0.67	9.64
LINEAR SVM'		0.6	8.43	0.64	7.03	0.89	7.2	0.94	6.89
LOGISTIC REGRESSION'		0.76	9.5	0.7	9.67	0.77	7.04	0.94	9.49
NAVIE BAYES'		0.71	8.06	0.79	8.48	0.97	7.71	0.68	9.38
RANDOM FOREST'		0.67	7.02	0.87	9.45	0.62	7.74	0.79	8.06
KNN'		0.84	7.29	0.94	9.92	0.75	8.41	0.81	6.9
SVM'		0.89	8.62	0.86	9.28	0.83	6.91	0.68	6.74
NEURAL NETWORK'		0.89	7.99	0.93	8.42	0.91	8.97	0.97	7.07
GAUSSIAN SVM'	50%	0.8	8.61	0.83	9.55	0.9	7.71	0.71	9.03
LINEAR SVM'		0.73	7.29	0.8	7.4	0.75	8.75	0.98	8.18
LOGISTIC REGRESSION'		0.78	9.77	0.92	9.8	0.97	7.96	0.88	8.25
NAVIE BAYES'		0.72	8.95	0.78	9.97	0.67	9.14	0.85	9.72
RANDOM FOREST'		0.88	8.85	0.62	9.71	0.96	7.31	0.82	7.21
KNN'		0.97	7.85	1	6.58	0.7	7.12	0.91	6.93
SVM'		0.71	8.75	0.9	8.53	0.76	6.99	0.69	8.02
NEURAL NETWORK'		0.79	8.04	0.83	7.94	0.6	9.97	0.64	7.24

Breast Cancer dataset is analyzed by the PCA feature selection method which is shown in Table 6 and the data processing is done for the mentioned eight various classifiers. Each classifier generates different output in terms of different accuracy and distinguished. In the table, light pink colour indicates the row-wise highest accuracy for each of the eight classifiers used, middle pink colour is for the highest accuracy of PCA data in each feature selection % taken for different ratios and the dark pink colour cell is the highest accuracy out of the entire breast cancer data analysis.

Then, the shades of blue colour indicate the AUC for PCA data. For each classifier, light blue colour is the highest accuracy; the middle blue colour is for the highest accuracy in each feature selection % done for different ratios and the dark blue colour is the highest AUC of the complete result analysis of breast cancer data.

Table 6: PCA results for individual ratio of dataset

PCA									
Ratio		**09:01**		**08:02**		**07:03**		**06:04**	
Method		**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
GAUSSIAN SVM'	RAW	0.63	0.845	0.6	0.801	0.6	0.793	0.65	0.813
LINEAR SVM'		0.67	0.78	0.84	0.975	0.98	0.801	0.79	0.841
LOGISTIC REGRESSION'		0.81	0.812	0.62	0.856	0.67	0.751	0.99	0.855
NAVIE BAYES'		0.99	0.94	0.6	0.919	0.88	0.91	0.82	0.8
RANDOM FOREST'		0.91	0.803	0.75	0.973	0.95	0.848	0.73	0.9
KNN'		0.97	0.977	0.84	0.83	0.94	0.858	0.97	0.753
SVM'		0.81	0.928	0.67	0.831	0.67	0.827	0.76	0.885
NEURAL NETWORK'		0.61	0.886	0.71	0.806	0.69	0.784	0.99	0.984
GAUSSIAN SVM'	95%	0.93	0.931	0.67	0.837	0.67	0.745	0.72	0.924
LINEAR SVM'		0.85	0.884	0.78	0.818	0.8	0.939	0.91	0.892
LOGISTIC REGRESSION'		0.9	0.91	0.65	0.874	0.74	0.768	0.66	0.795
NAVIE BAYES'		0.87	0.855	0.88	0.81	0.6	0.881	0.71	0.987
RANDOM FOREST'		0.97	0.845	0.61	0.916	0.94	0.993	0.62	0.86
KNN'		0.83	0.92	0.89	0.911	0.89	0.84	0.83	0.774
SVM'		0.62	0.995	0.71	0.897	0.99	0.792	0.67	0.832
NEURAL NETWORK'		0.98	0.845	0.71	0.783	0.76	0.84	0.65	0.856
GAUSSIAN SVM'	90.0%	0.63	0.902	0.6	0.891	0.92	0.805	0.78	0.89
LINEAR SVM'		0.62	0.872	0.86	0.801	0.94	0.993	0.94	0.874
LOGISTIC REGRESSION'		0.71	0.936	0.69	0.99	0.85	0.898	0.67	0.768
NAVIE BAYES'		0.7	0.964	0.97	0.924	0.89	0.803	0.83	0.952
RANDOM FOREST'		0.76	0.998	0.63	0.827	0.8	0.76	0.89	0.887
KNN'		0.81	0.957	0.95	0.946	0.73	0.86	0.9	0.773
SVM'		0.64	0.814	0.81	0.993	0.89	0.824	0.71	0.962
NEURAL NETWORK'		0.97	0.908	0.7	0.767	0.94	0.894	0.98	0.76

GAUSSIAN SVM'	85%	0.83	0.817	0.93	0.793	0.78	0.845	0.93	0.918
LINEAR SVM'		0.68	0.826	0.65	0.916	0.83	0.788	0.66	0.866
LOGISTIC REGRESSION'		0.97	0.886	0.61	0.758	0.93	0.86	0.75	0.947
NAVIE BAYES'		0.74	0.881	0.89	0.968	0.73	0.911	0.99	0.764
RANDOM FOREST'		0.84	0.85	0.72	0.812	0.91	0.999	0.67	0.944
KNN'		0.68	0.999	0.92	0.853	0.89	0.872	0.93	0.836
SVM'		0.63	0.896	0.97	0.794	0.77	0.936	0.61	0.987
NEURAL NETWORK'		0.91	0.888	0.67	0.872	0.81	0.999	0.95	0.991
GAUSSIAN SVM'	80.0%	0.87	0.848	0.98	0.867	0.69	0.846	0.88	0.887
LINEAR SVM'		0.91	0.999	0.99	0.881	0.99	0.774	0.62	0.822
LOGISTIC REGRESSION'		0.83	0.88	0.96	0.883	0.77	0.883	0.89	0.749
NAVIE BAYES'		0.92	0.781	0.79	0.81	0.75	0.914	0.66	0.816
RANDOM FOREST'		0.68	0.794	0.73	0.97	0.79	0.848	0.67	0.993
KNN'		0.76	0.961	0.85	0.841	0.95	0.945	0.79	0.953
SVM'		0.96	0.854	0.73	0.897	0.96	0.924	0.75	0.933
NEURAL NETWORK'		0.99	0.883	0.82	0.824	0.62	0.791	0.63	0.863
GAUSSIAN SVM'	75%	0.6	0.979	0.86	0.745	0.61	0.798	0.78	0.777
LINEAR SVM'		0.6	0.931	0.74	0.944	0.77	0.856	0.62	0.757
LOGISTIC REGRESSION'		0.63	0.897	0.69	0.96	0.95	0.991	0.8	0.801
NAVIE BAYES'		0.69	0.882	0.91	0.833	0.78	0.908	0.97	0.786
RANDOM FOREST'		0.89	0.892	0.77	0.971	0.76	0.79	0.85	0.904
KNN'		0.73	0.95	1	0.996	0.65	0.804	0.6	0.9
SVM'		0.64	0.849	0.96	0.885	0.75	0.798	0.78	0.989
NEURAL NETWORK'		0.65	0.865	0.95	0.756	0.88	0.995	0.71	0.779
GAUSSIAN SVM'	70.0%	0.88	0.977	0.85	0.975	0.67	0.938	0.74	0.852
LINEAR SVM'		0.66	0.954	0.85	0.934	0.93	0.762	0.98	0.872
LOGISTIC REGRESSION'		0.9	0.935	0.94	0.785	0.78	0.903	0.98	0.958

NAVIE BAYES'		0.96	0.894	0.83	0.963	0.61	0.971	0.76	0.754
RANDOM FOREST'		0.9	0.784	0.65	0.9	0.7	0.827	0.76	0.849
KNN'		0.75	0.901	0.66	0.793	0.63	0.827	0.91	0.804
SVM'		0.9	0.922	0.93	0.956	0.72	0.824	0.81	0.828
NEURAL NETWORK'		0.94	0.952	0.82	0.812	0.87	0.804	0.78	0.843
GAUSSIAN SVM'	65%	0.82	0.998	0.9	0.995	0.69	0.88	0.62	0.938
LINEAR SVM'		0.84	0.964	1	0.982	0.76	0.745	0.82	0.798
LOGISTIC REGRESSION'		0.68	0.828	0.63	0.936	0.9	0.884	0.73	0.958
NAVIE BAYES'		0.82	0.99	0.96	0.836	0.82	0.833	0.85	0.948
RANDOM FOREST'		0.9	0.777	0.93	0.751	0.76	0.932	0.92	0.839
KNN'		0.9	0.973	0.69	0.778	0.69	0.834	0.71	0.982
SVM'		0.62	0.896	0.66	0.959	0.66	0.873	1	0.835
NEURAL NETWORK'		0.61	0.799	0.76	0.83	0.69	0.984	0.88	0.991
GAUSSIAN SVM'	60.0%	0.77	0.985	0.6	0.901	0.92	0.804	0.98	0.94
LINEAR SVM'		0.93	0.891	0.92	0.829	0.69	0.824	0.83	0.957
LOGISTIC REGRESSION'		0.71	0.848	0.95	0.902	1	0.797	0.93	0.918
NAVIE BAYES'		0.7	0.866	0.76	0.898	0.92	0.771	0.93	0.96
RANDOM FOREST'		0.74	0.855	0.83	0.924	0.9	0.939	0.75	0.854
KNN'		0.99	0.891	0.94	0.815	0.85	0.895	0.99	0.766
SVM'		0.8	0.878	0.63	0.976	0.96	0.857	0.92	0.783
NEURAL NETWORK'		0.85	0.811	0.78	0.961	0.68	0.822	0.79	0.831
GAUSSIAN SVM'	55%	0.92	0.997	0.66	0.805	0.88	0.841	0.99	0.993
LINEAR SVM'		0.86	0.965	0.76	0.906	1	0.888	0.98	0.929
LOGISTIC REGRESSION'		0.79	0.908	0.96	0.795	0.76	0.998	0.76	0.913
NAVIE BAYES'		0.96	0.999	0.86	0.772	0.61	0.903	0.83	0.991
RANDOM FOREST'		0.9	0.914	0.81	0.811	0.99	0.883	0.61	0.923
KNN'		0.81	0.76	0.96	0.829	0.69	0.774	0.72	0.803
SVM'		0.86	0.761	0.71	0.817	0.96	0.858	0.9	0.899

NEURAL NETWORK'		0.92	0.774	1	0.962	0.62	0.864	0.73	0.906
GAUSSIAN SVM'	50.0%	0.69	0.893	0.84	0.898	0.78	0.754	0.81	0.849
LINEAR SVM'		0.64	0.862	0.78	0.886	0.93	0.924	0.95	0.758
LOGISTIC REGRESSION'		0.69	0.862	0.99	0.947	0.78	0.83	0.62	0.934
NAVIE BAYES'		0.8	0.796	0.77	0.788	0.9	0.839	0.98	0.749
RANDOM FOREST'		0.93	0.905	0.82	0.911	0.89	0.769	0.95	0.748
KNN'		0.72	0.791	0.97	0.762	0.83	0.908	0.86	0.966
SVM'		0.62	0.954	0.81	0.922	0.68	0.884	0.88	0.989
NEURAL NETWORK'		0.78	0.766	0.62	0.906	0.92	0.921	0.74	0.987

The breast cancer data set is processed using the ReliefF feature selection method. The Table 7 displays the individual data for each training-testing ratio of data and the feature selection % used for the analysis purpose. Different classifier gives different accuracy and AUC which can be studied from the given table. In the table, light pink colour indicates the row-wise highest accuracy for each of the eight classifiers used, middle pink colour is for the highest accuracy of ReliefF data in each feature selection % taken for different ratios and the dark pink colour cell is the highest accuracy out of the entire breast cancer data analysis.

Then, the shades of blue colour indicate the AUC for ReliefF data. For each classifier, light blue colour is the highest accuracy; the middle blue colour is for the highest accuracy in each feature selection % done for different ratios and the dark blue colour is the highest AUC of the complete result analysis of breast cancer data.

Table 7: ReliefF results for individual ratio of dataset

ReliefF									
Ratio		**09:01**		**08:02**		**07:03**		**06:04**	
Method		**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
GAUSSIAN SVM'	Raw	0.76	7.9	0.88	8.19	0.95	8.4	0.75	8.51
LINEAR SVM'		0.65	7.13	0.97	7.67	0.81	6.94	0.9	6.98
LOGISTIC REGRESSION'		0.81	7.14	0.98	8.55	0.89	8.03	0.85	7.91
NAVIE BAYES'		0.82	6.98	0.7	7.27	0.76	6.95	0.69	9.62
RANDOM FOREST'		0.88	9.49	0.66	9.61	0.89	9.56	0.97	9.49
KNN'		0.95	9.76	0.75	8.47	0.88	8.4	0.75	8.38
SVM'		0.6	9.22	0.69	9.4	0.72	8.53	0.75	8.08
NEURAL NETWORK'		0.88	7.01	0.61	7.04	0.63	8.33	0.73	6.98
GAUSSIAN SVM'	95%	1	9.04	0.67	7.56	0.85	9.51	0.97	8.19
LINEAR SVM'		1	7.13	0.69	8.5	0.79	8.25	0.79	8.47
LOGISTIC REGRESSION'		0.78	8.08	1	6.88	0.99	9.07	0.8	7.66
NAVIE BAYES'		0.74	9.49	0.86	8.82	0.73	9.5	0.92	7.93
RANDOM FOREST'		1	6.91	0.64	7.09	0.88	6.85	0.65	7.07
KNN'		0.79	8.63	0.93	7.08	0.79	7.99	0.68	9.42
SVM'		0.72	8.72	0.89	9.33	0.7	7.46	0.67	8.85
NEURAL NETWORK'		0.75	9.12	0.89	6.65	0.79	9.54	0.6	7.11
GAUSSIAN SVM'	90%	0.6	7.36	0.82	8.52	0.85	6.68	0.95	8.62
LINEAR SVM'		0.8	8.4	0.75	8.95	0.78	8.62	0.6	8.96
LOGISTIC REGRESSION'		0.65	9.78	0.74	9.66	0.63	7.67	0.6	6.55
NAVIE BAYES'		0.72	9.83	0.63	7.48	0.99	7.41	0.94	8.3
RANDOM FOREST'		0.7	8.6	0.68	9.51	0.68	8.47	0.91	7.37
KNN'		0.6	9.33	0.63	7.9	0.63	6.92	0.95	8.25
SVM'		0.91	7.57	0.91	10	0.82	6.93	0.98	9.91
NEURAL NETWORK'		0.87	7.41	0.62	8.64	0.66	9.77	0.6	9.36
GAUSSIAN SVM'	85%	0.76	7.97	0.84	9.86	0.6	7.64	0.67	9.5
LINEAR SVM'		0.8	9.11	0.71	6.58	0.82	7.15	0.88	9.64
LOGISTIC REGRESSION'		0.85	8.1	0.81	7.09	0.85	6.7	1	7.51
NAVIE		0.65	9.95	0.99	6.84	0.76	9.69	0.83	9.58

BAYES'									
RANDOM FOREST'		0.71	9.47	0.88	7.11	0.83	9.85	0.83	7.35
KNN'		0.94	7.59	0.76	9.57	0.63	7.72	0.98	7.13
SVM'		0.95	7.08	0.64	8.64	0.8	6.79	0.88	7.11
NEURAL NETWORK'		0.62	9.49	0.77	9.14	0.63	9.63	0.73	9.51
GAUSSIAN SVM'	80%	0.8	9.51	0.82	8.38	0.7	7.81	0.73	8.03
LINEAR SVM'		0.94	6.82	0.69	7.14	0.67	7.01	0.98	7.82
LOGISTIC REGRESSION'		0.97	9.15	0.6	9.15	0.61	8.73	0.72	9.41
NAVIE BAYES'		0.92	6.94	0.82	7.06	0.95	6.58	0.61	7.18
RANDOM FOREST'		1	8.61	0.96	9.51	0.62	6.68	0.83	9.64
KNN'		0.83	9.27	0.75	9.41	0.79	8.36	0.86	8.08
SVM'		0.62	7.2	0.93	9.11	0.84	9.71	0.9	7.03
NEURAL NETWORK'		0.63	6.56	0.66	9.01	0.74	7.04	0.99	7.4
GAUSSIAN SVM'	75%	1	9.31	0.84	8.31	0.78	6.89	0.72	8.3
LINEAR SVM'		0.78	6.58	0.94	9.69	0.92	6.49	0.91	9.19
LOGISTIC REGRESSION'		0.96	9.08	0.73	7.44	0.93	9.16	0.71	7.82
NAVIE BAYES'		0.97	7.65	0.76	6.84	0.72	9.02	0.96	6.6
RANDOM FOREST'		0.99	7.93	0.76	7.71	0.7	9.49	0.87	9.96
KNN'		0.91	7.13	0.83	6.86	0.75	7.33	0.71	8.31
SVM'		0.98	9.82	0.65	8.6	0.66	8.63	0.73	9.45
NEURAL NETWORK'		0.87	6.93	0.64	7.73	0.9	6.62	0.89	9.62
GAUSSIAN SVM'	70%	0.89	8.12	0.65	8.49	0.62	6.85	0.67	8.89
LINEAR SVM'		0.65	9.48	0.78	8.65	0.63	7.68	0.78	8.29
LOGISTIC REGRESSION'		0.63	7.32	0.79	9.01	0.98	8.13	0.74	8.02
NAVIE BAYES'		0.64	7.98	0.88	8.43	0.83	6.73	0.68	7.8
RANDOM FOREST'		0.75	9.98	0.74	8.05	0.95	8.87	0.81	6.56
KNN'		0.99	9.75	0.72	7.68	0.75	8.32	0.94	7.21
SVM'		0.74	6.6	0.61	6.66	0.74	8.28	0.79	6.75
NEURAL NETWORK'		0.98	8.39	0.98	8.03	0.71	9.79	0.85	9.28
GAUSSIAN SVM'	65%	0.63	6.45	0.63	7.23	0.63	6.91	0.68	7.88
LINEAR SVM'		0.64	9.03	0.98	9.03	0.86	8.76	1	6.62
LOGISTIC		0.8	8.65	0.69	9.09	0.98	9.46	0.6	7.57

REGRESSION'									
NAVIE BAYES'		0.97	7.77	0.63	7.37	0.82	7.68	0.85	9.55
RANDOM FOREST'		0.65	7.49	0.89	6.46	0.87	6.69	0.93	9.76
KNN'		0.65	6.64	0.9	9.24	0.97	6.49	0.76	7.96
SVM'		0.73	8.94	0.96	9.96	0.83	6.47	0.96	9.53
NEURAL NETWORK'		0.78	9.62	0.84	8.61	0.97	7.49	0.8	8.16
GAUSSIAN SVM'	60%	0.6	9.03	0.79	7.72	0.98	7.22	0.99	9.68
LINEAR SVM'		0.76	8.53	0.73	9.88	0.71	7.52	0.84	7.34
LOGISTIC REGRESSION'		0.67	9.83	0.78	9.55	0.98	8.56	0.87	8.32
NAVIE BAYES'		0.95	7.73	0.78	7.99	0.93	8.45	0.97	7.32
RANDOM FOREST'		0.82	6.77	0.84	6.99	0.81	7.22	0.78	9.43
KNN'		0.95	8.18	0.9	8.76	0.93	7.84	0.85	8.32
SVM'		0.62	7.03	1	7.02	0.88	7.65	0.68	7.69
NEURAL NETWORK'		0.8	7.36	0.93	7.58	0.95	9.1	0.82	7.36
GAUSSIAN SVM'	55%	0.65	6.45	0.71	6.97	0.83	7.76	0.64	7.05
LINEAR SVM'		0.77	9.16	0.94	7.02	0.87	7.95	0.71	8.58
LOGISTIC REGRESSION'		0.78	9.16	0.81	7.34	0.61	8.48	0.82	9.33
NAVIE BAYES'		0.73	6.5	0.6	8.16	0.91	8.54	0.67	9.47
RANDOM FOREST'		0.97	6.89	0.91	7.97	0.67	6.55	0.71	9.45
KNN'		0.89	6.73	0.94	9.63	0.82	6.77	0.69	7.28
SVM'		0.86	6.72	0.69	7.47	0.81	8.23	0.82	8.25
NEURAL NETWORK'		0.98	8.55	0.78	7.51	0.72	9.71	0.82	8.83
GAUSSIAN SVM'	50%	0.89	8.21	0.96	8.46	0.88	7.33	0.85	6.77
LINEAR SVM'		0.81	8.2	0.63	9.18	0.99	8.77	0.6	8.38
LOGISTIC REGRESSION'		0.86	8.75	0.8	9.67	0.82	6.68	0.67	7.51
NAVIE BAYES'		0.9	9.44	0.69	7.75	0.7	6.5	0.79	7.8
RANDOM FOREST'		0.95	6.51	0.64	7.53	0.6	8.3	0.6	8.35
KNN'		0.95	9.45	0.78	8.59	0.91	8.77	0.79	7.96
SVM'		0.69	9.19	0.61	8.19	0.6	7.85	0.63	7.27
NEURAL NETWORK'		0.72	7.47	0.69	7.32	0.7	7.8	0.7	6.55

After analyzing the results of data processing by CFS feature selection method, the cells showing the best result are extracted and mentioned in Table 8. The different combination of feature selection % ratio with different training testing ratio of different classifiers with CFS feature selection method gives best accuracy that is 1 as well as gives best AUC for 60% of feature selection ratio with 60:40 (training – testing ratio) with Logistic Regression classifier which is 9.99.

Random Forest method gives the highest accuracy with different training- testing ratios for CFS feature selection method.

Table 8: CFS Results

CFS							
Ratio	**Feature selection % ratio**	**Classifier**	**Accuracy**	**Ratio**	**Feature selection % ratio**	**Classifier**	**AUC**
08:02	95%	Random Forest	1	06:04	60%	Logistic Regression	9.99
07:03	85%	Linear SVM	1				
09:01	85%	Random Forest	1				
08:02	80%	SVM	1				
09:01	75%	Linear SVM	1				
06:04	75%	Naive Bayes	1				
08:02	75%	SVM	1				
07:03	70%	Logistic Regression	1				
08:02	65%	Random Forest	1				
08:02	60%	Random Forest	1				

Table 9 indicates the analysis of Gain Ratio feature selection method, the cells showing the best result are extracted and mentioned in Table 9. The different combination of feature selection % ratio with

different training testing ratio of different classifiers with Gain ratio feature selection method gives best accuracy that is 1 as well as gives best AUC for 75% of feature selection ratio with 60:40 (training – testing ratio) with Logistic Naïve Bayes and 65% of feature selection ratio with 80:20 (training – testing ratio) with Random Forest which is 10.

KNN method gives the highest accuracy with different training- testing ratios for Gain Ratio feature selection method.

Table 9: Gain Ratio Results

Gain Ratio							
Ratio	**Feature selection % ratio**	**Classifier**	**Accuracy**	**Ratio**	**Feature selection % ratio**	**Classifier**	**AUC**
08:02	Raw	Gaussian SVM	1	06:04	75%	Naive Bayes	10
09:01	85%	Linear SVM	1	08:02	65%	Random Forest	10
09:01	85%	KNN	1				
08:02	70%	Linear SVM	1				
07:03	60%	Linear SVM	1				
07:03	60%	Naive Bayes	1				
07:03	60%	KNN	1				
08:02	50%	KNN	1				

Studying the result table of PCA feature selection method, accuracy and AUC are highest for more than single combination of training-testing ratio, feature selection % and classifiers used. All such maximum values for PCA data are arranged in Table 10 for the ease of analysis.

The different combination of feature selection % ratio with different training testing ratio of different classifiers with PCA feature selection method gives best accuracy that is 1 as well as gives best AUC for different combination of feature selection % ratio with different training testing ratio of different

classifiers with PCA feature selection method which is 10.

Linear SVM method gives the highest accuracy with different training- testing ratios for Gain Ratio feature selection method.

Table 10: PCA Result

PCA							
Ratio	**Feature selection % ratio**	**Classifier**	**Accuracy**	**Ratio**	**Feature selection % ratio**	**Classifier**	**AUC**
08:02	75%	KNN	1	07:03	85%	Random Forest	0.999
08:02	65%	Linear SVM	1	09:01	85%	KNN	0.999
06:04	65%	SVM	1	07:03	85%	Neural Network	0.999
07:03	60%	Logistic Regression	1	09:01	80%	Linear SVM	0.999
07:03	55%	Linear SVM	1	09:01	55%	Naive Bayes	0.999
08:02	55%	Neural Network	1				

Table 11 indicates the analysis of RelifF feature selection method, the cells showing the best result are extracted and mentioned in Table 11. The different combination of feature selection % ratio with different training testing ratio of different classifiers with RelifF feature selection method gives best accuracy that is 1 as well as gives best AUC for 90% of feature selection ratio with 80:20 (training – testing ratio) with SVM which is 10.

Table 11: ReliefF Results

ReliefF							
Ratio	**Feature selection % ratio**	**Classifier**	**Accuracy**	**Ratio**	**Feature selection % ratio**	**Classifier**	**AUC**
09:01	95%	Gaussian SVM	1	08:02	90%	SVM	10
09:01	95%	Linear SVM	1				
08:02	95%	Logistic Regression	1				
09:01	95%	Random Forest	1				
06:04	85%	Logistic Regression	1				
09:01	80%	Random Forest	1				
09:01	75%	Gaussia n SVM	1				
06:04	65%	Linear SVM	1				
08:02	60%	SVM	1				

The model consists of a large number of gene data points in the training set as well as the testing set. Evaluation is based on four distinct feature selection methods and 10 cross-fold validations are done at a later stage. Eight different classification methods, linear as well as non-linear methods are executed. However, considering the statistics, Linear SVM can be evaluated as the best method.

245 different features are selected spanning the methylation, histone, human genome as well as CHIP-Seq data. Initially, the relationship between these four is derived. Based on that, the clustering of selected data is obtained. In the entire process, CpG islands are proved to be important for the gene expression prediction. Also, a collinear relation is observed between methylation and histone as some of the methylation features are found to be clustering with histone modification features [11].

The entire data is classified based on eight different classifiers namely, Gaussian SVM, Linear SVM,

Naïve Bayes, Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Neural Network.

6.2 Discussion

Though various approaches figure out the modification in epigenetic as well as genetic expression, a well-structured quantitative approach that feature the accurate prediction of up and down- regulation of gene expression is still lacking. It has been noted quite often that reliable epigenetic data is obtained but genetic data is missing. Epigenetics measurement is possible in several data sets for which genetic quantification is difficult. In such cases, a predictive method can effectively provide the required information. Apart from providing the prediction of gene expression, this model also expresses the relative importance of genome data and their genomic location. Certainly, CpG methylation data consists of more predictive values for various genetic expressions. Although various histone modification data can be used for the study, they are quite expensive as compared to CpG methylation. All the parameters in this model are useful in extracting some sort of information at the genetic level. Many features obtained from methylation and histone modification are based on the annotations from Illumina 450K for DNA methylation.

CFS: By using the Random Forest classifier, for the ratio 80:20, the highest accuracy is observed with feature selection ratio to be 95% and 85% for 90:10 ratios. Whereas for the Linear SVM classifier, 85% feature selection ratio is observed for the 70:30 training-testing ratio. For the SVM classifier, 80% feature selection is found for the ratio 80:20. For the 70:30 ratios, using Logistic Regression 70% feature selection is observed. Maximum AUC (9.99) is found for 60:40 ratios and feature selection is 60% using the Logistics Regression method.

Table 12: CFS Result Analysis

CFS							
Ratio	**Feature selection % ratio**	**Classifier**	**Accuracy**	**Ratio**	**Feature selection % ratio**	**Classifier**	**AUC**
09:01	Raw	Linear SVM	0.99	07:03	Raw	Linear SVM	9.84
08:02	95%	Random Forest	1	09:01	95%	KNN	9.89
07:03	90%	Gaussian SVM	0.98	09:01	90%	Logistic Regression	9.98
07:03	85%	Linear SVM	1	09:01	85%	SVM	9.87
09:01	85%	Random Forest	1	09:01	80%	SVM	9.85
08:02	80%	SVM	1	06:04	75%	Logistic Regression	9.97
09:01	75%	Linear SVM	1	06:04	70%	Naive Bayes	9.84
06:04	75%	Naive Bayes	1	06:04	65%	Logistic Regression	9.9
08:02	75%	SVM	1	06:04	60%	Logistic Regression	9.99
07:03	70%	Logistic Regression	1	09:01	55%	Logistic Regression	9.96
08:02	65%	Random Forest	1	09:01	50%	Neural Network	9.96
08:02	60%	Random Forest	1				
07:03	55%	Naive Bayes	0.99				
09:01	50%	Neural Network	0.99				

Gain Ratio: Table 13 indicates the analysis of Gain Ratio feature selection method. For each classifier, the best combination of ratio along with the feature selection % is mentioned. For Feature selection ratio (Raw, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, 50%) with different training- testing ratio (90:10, 80:20, 70:30, 60:40) with different classifiers (Gaussian SVM, Linear SVM, KNN, Naïve Bayes, Random forest) the highest accuracy to be observed is 1.

Table 13 also shows the analysis of different combination of ratio along with the feature selection % for each classifier. The highest AUC to be observed in this table is 10 for 80:20 training-testing ratios along with 65% of feature selection ratio with Random Forest Classifier and also for 60:40 training-testing ratios along with 80% of feature selection ratio with Naïve Bayes Classifier.

Table 13: Gain Ratio Result Analysis

Gain Ratio							
Ratio	**Feature selection % ratio**	**Classifier**	**Accuracy**	**Ratio**	**Feature selection % ratio**	**Classifier**	**AUC**
08:02	Raw	Gaussian SVM	1	06:04	Raw	Naive Bayes	9.93
06:04	95%	Gaussian SVM	0.99	08:02	95%	SVM	9.83
09:01	90%	Naive Bayes	0.99	06:04	90%	Neural Network	9.95
09:01	85%	Linear SVM	1	07:03	85%	Neural Network	9.99
09:01	85%	KNN	1	08:02	80%	Naive Bayes	9.97
09:01	80%	Random Forest	0.98	06:04	75%	Naive Bayes	10
06:04	75%	KNN	0.99	06:04	70%	Linear SVM	9.64
08:02	70%	Linear SVM	1	08:02	65%	Random Forest	10
08:02	65%	Naive Bayes	0.99	08:02	60%	Linear SVM	9.65
07:03	65%	SVM	0.99	08:02	55%	KNN	9.92
07:03	60%	Linear SVM	1	08:02	50%	Naive Bayes	9.97
07:03	60%	Naive Bayes	1	07:03	50%	Neural Network	9.97
07:03	60%	KNN	1				
08:02	55%	Gaussian SVM	0.97				
08:02	50%	KNN	1				

PCA: 100% accuracy is observed for five times while analyzing data using PCA. Maximum accuracy is observed for the 80:20 ratios, when feature selection ratio is 75% and KNN classifier is used; when feature selection is 65% and Linear SVM is used and when feature selection is 55% and neural network is used. Also, for 60:40 ratio and 65% feature selection ratio using Linear SVM, 100% accuracy is found. Best AUC (0.999) is recorded five times for ratio of 90:10 and 70:30.

Table 14: PCA Result Analysis

PCA							
Ratio	**Feature selection % ratio**	**Classifier**	**Accuracy**	**Ratio**	**Feature selection % ratio**	**Classifier**	**AUC**
09:01	Raw	Naive Bayes	0.99	06:04	Raw	Neural Network	0.984
06:04	Raw	Logistic Regression	0.99	09:01	95%	SVM	0.995
06:04	Raw	Neural Network	0.99	09:01	90%	Random Forest	0.998
07:03	95%	SVM	0.99	07:03	85%	Random Forest	0.999
06:04	90%	Neural Network	0.98	09:01	85%	KNN	0.999
06:04	85%	Naive Bayes	0.99	07:03	85%	Neural Network	0.999
07:03	80%	Linear SVM	0.99	09:01	80%	Linear SVM	0.999
08:02	80%	Linear SVM	0.99	08:02	75%	KNN	0.996
09:01	80%	Neural Network	0.99	09:01	70%	Gaussian SVM	0.977
08:02	75%	KNN	1	09:01	65%	Gaussian SVM	0.998
06:04	70%	Linear SVM	0.98	09:01	60%	Gaussian SVM	0.985
06:04	70%	Logistic Regression	0.98	09:01	55%	Naive Bayes	0.999
08:02	65%	Linear SVM	1	06:04	50%	SVM	0.989
06:04	65%	SVM	1				

07:03	60%	Logistic Regression	1				
07:03	55%	Linear SVM	1				
08:02	55%	Neural Network	1				
08:02	50%	Logistic Regression	0.99				

ReliefF: For the testing-training ratio of 90:10 and 95% feature selection, highest accuracy is noted thrice - once using Gaussian SVM, next using Linear SVM and then for Random Forest classifier. For 80:20 and 95% feature selection ratio, the highest accuracy is noted using the Logistic Regression classifier. Also, when the logistic regression classifier is used for 60:40 ratio and 85% feature selection, the highest accuracy is recorded. Maximum AUC (10) is noted for 80:20 ratios using a 90% feature selection ratio and SVM classifier is shown in Table 15.

Table 15: ReliefF Result Analysis

ReliefF							
Ratio	**Feature selection % ratio**	**Classifier**	**Accuracy**	**Ratio**	**Feature selection % ratio**	**Classifier**	**AUC**
08:02	Raw	Logistic Regression	0.98	09:01	Raw	KNN	9.76
09:01	95%	Gaussian SVM	1	07:03	95%	Neural Network	9.54
09:01	95%	Linear SVM	1	08:02	90%	SVM	10
08:02	95%	Logistic Regression	1	09:01	85%	Naive Bayes	9.95
09:01	95%	Random Forest	1	07:03	80%	SVM	9.71
07:03	90%	Naive Bayes	0.99	06:04	75%	Random Forest	9.96
06:04	85%	Logistic Regression	1	09:01	70%	Random Forest	9.98
09:01	80%	Random Forest	1	08:02	65%	SVM	9.96
09:01	75%	Gaussian SVM	1	08:02	60%	Linear SVM	9.88
09:01	70%	KNN	0.99	07:03	55%	Neural Network	9.71
06:04	65%	Linear SVM	1	08:02	50%	Logistic Regression	9.67
08:02	60%	SVM	1				
09:01	55%	Neural Network	0.98				
07:03	50%	Linear SVM	0.99				

Figure 12 represents the comparison of various feature selection techniques. The data with 100% accuracy are selected to compare the techniques. Different techniques are represented with different colours for a comprehensive pictorial analysis.

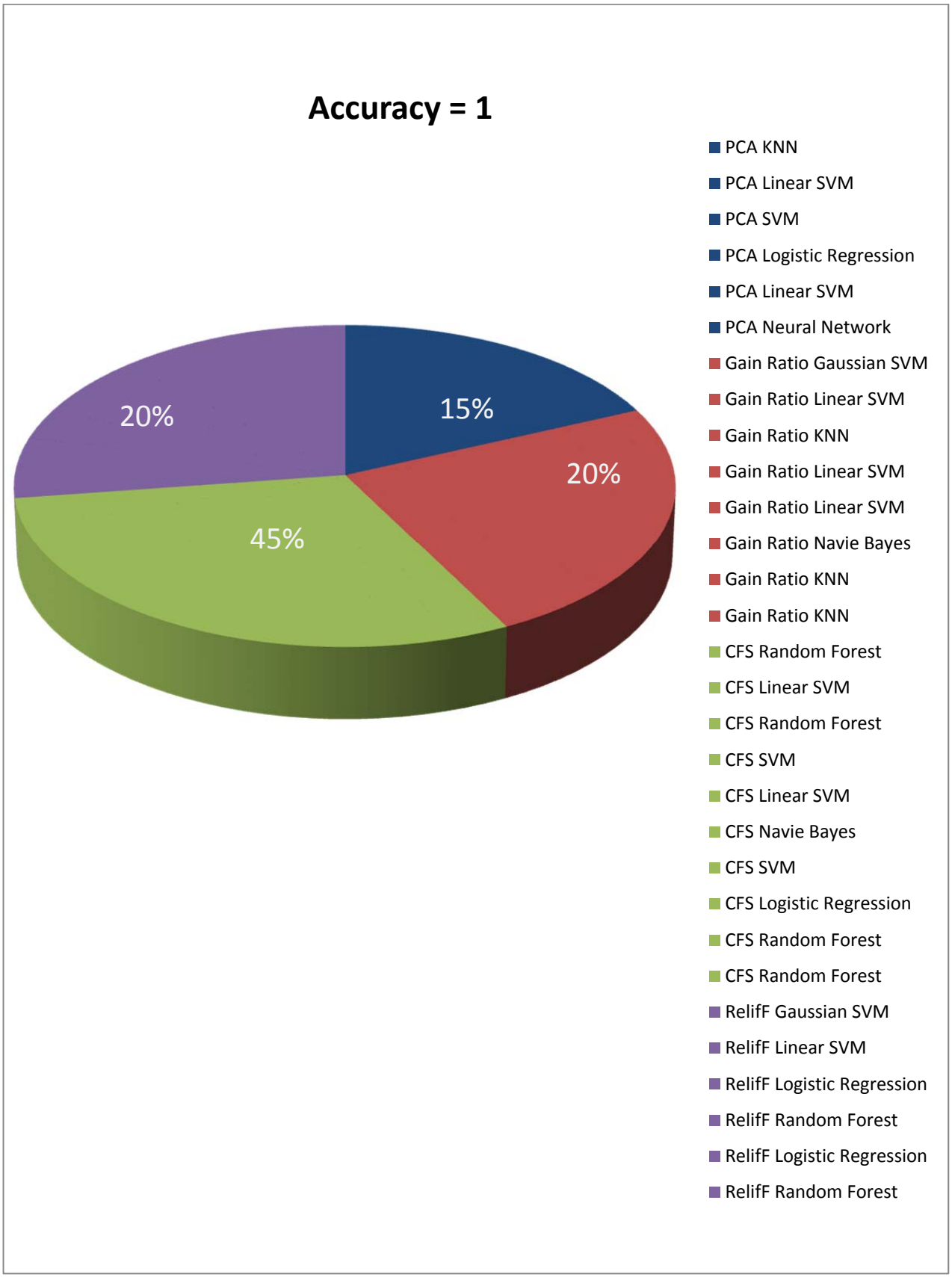


Figure 12: Feature Selection Comparison

Figure 13 is the graphical view of CFS output data. For data of 100% accuracy, the values have been taken for different classifiers and a graph has been plotted for detailed analysis.

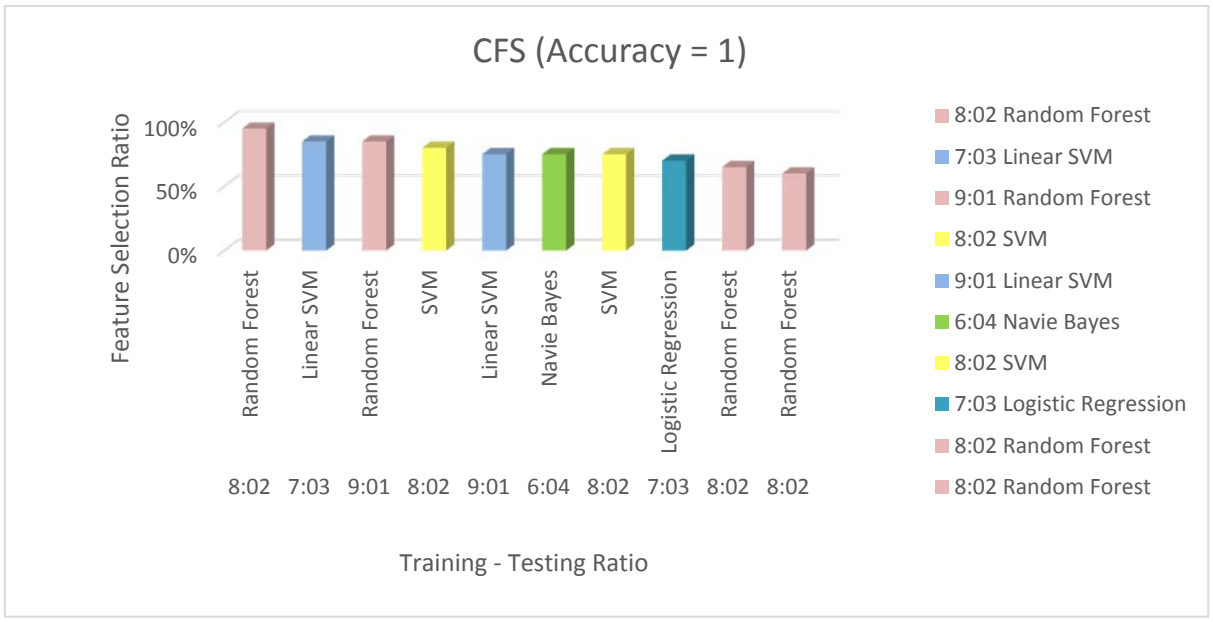


Figure 13: Graphical Representation of CFS Output Data

For analytical study of the output of gain ratio, a graph has been plotted considering the data for different classifiers when accuracy is 1. Figure 14 represents the output graph taken with data at different training-testing ratios.

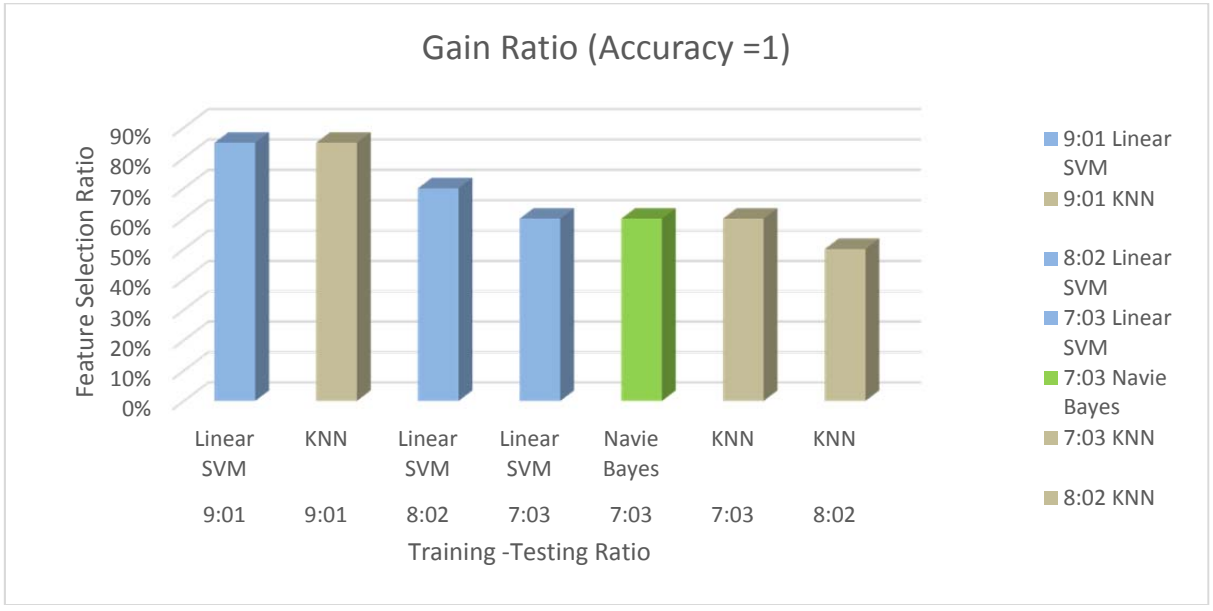


Figure 14: Gain Ratio Output Data Graph

Figure 15 displays the comprehensive graph for PCA at various training-testing ratios. The data is chosen for which accuracy is recorded to be 100%. Using such data values for all the chosen classifiers, the graph in Figure 14 has been plotted.

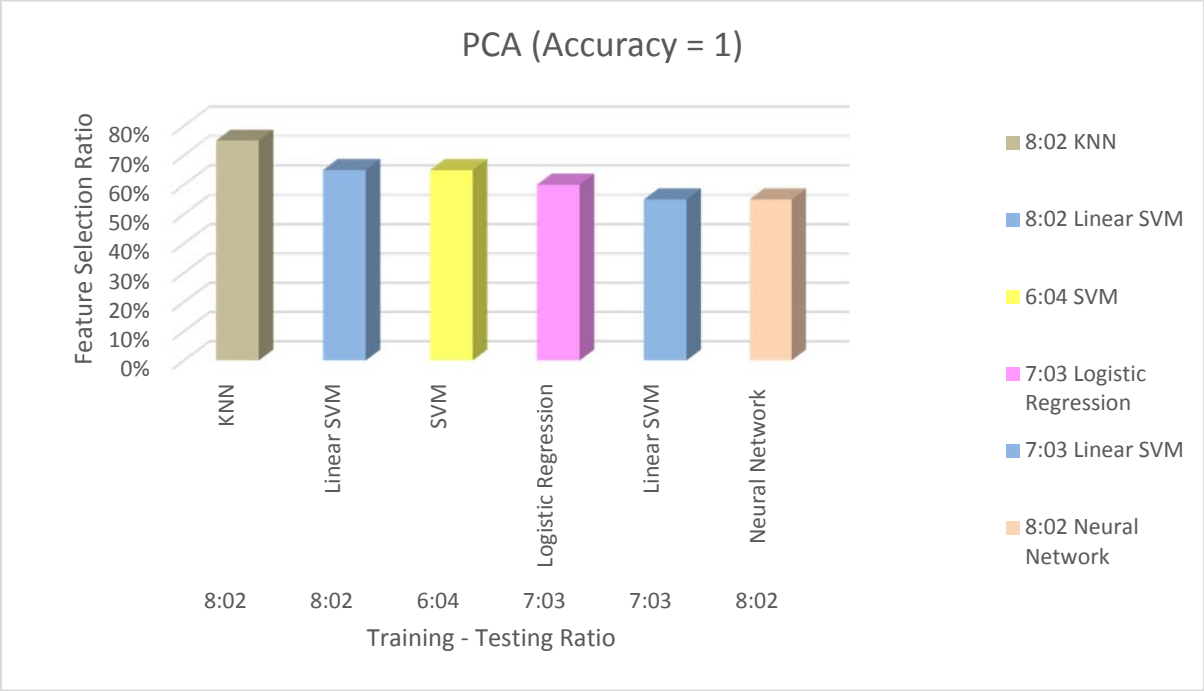


Figure 15: Output Graph for PCA

Data for various training-testing ratios have been taken for ReliefF when 100% accuracy has been noted. All such data values for various classifiers have been generated and an analytical graph has been plotted. Figure 16 represents the graph for ReliefF for 100% accuracy.

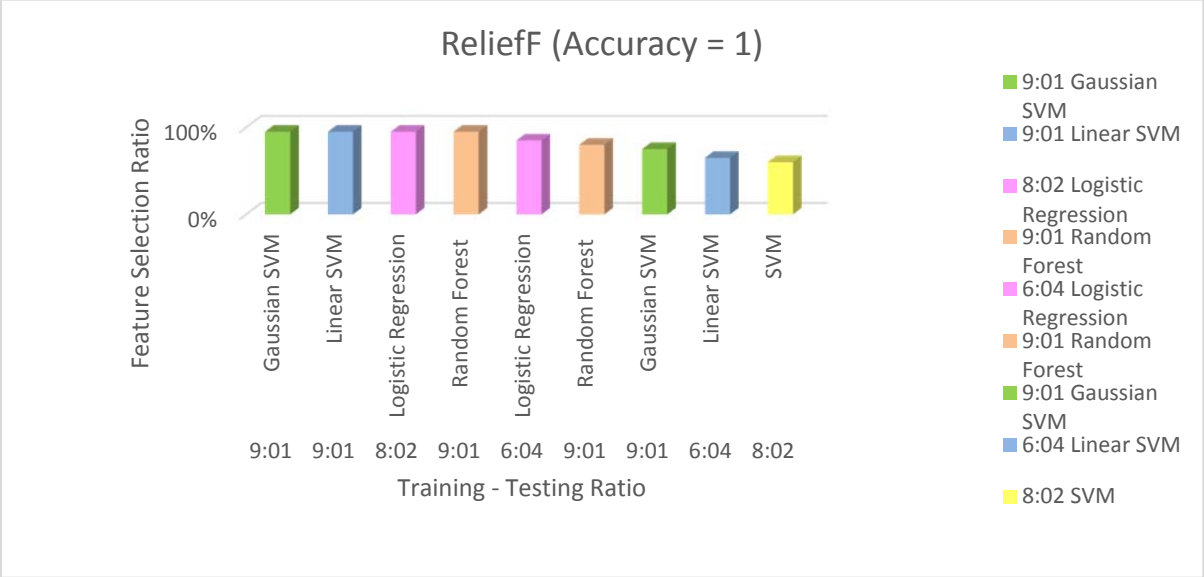


Figure 16: ReliefF Graph

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In the context of breast cancer, a new machine-based learning method of gene expression prediction is being developed. Using the Illumina Infinium HumanMethylation450 K Bead chip CpG methylation range, this technique utilizes information from combined pulmonary disease and neighboring ordinary organs in the Cancer Genome Atlas (TCGA) and histone alteration marker CHIP-Seq from the ENCODE initiative. It sees a comprehensive list of characteristics covering the four classes of CpG methylation, histone H3 methylation alteration, nucleotide structure, and conservation.Different techniques of choice of features and classification are contrasted in the training-testing ratio to select the best model over 10-fold cross-validation.

7.2 Future Work

In future, we can work on other cancer cell diagnosis such as liver cancer, bone cancer, blood cancer, oral cancer etc. The accuracy level of our research helps us to work more on such dataset. We should point out that our present model does not include all data on histone modification, but only three commonly used methylation markers on histone H3 (H3K4Me3, H3K27Me3, and H3K36Me3). In addition, the histone H3 data is taken from the ENCODE cell lines, as the TCGA samples do not have such data. The heterogeneity of the sample resources could influence the model's accuracy. When more histone marker data combined with DNA methylation and RNA-Seq data becomes openly accessible for breast cancer, we can include them to obtain a stronger model. In the ideal setting, we would like to build a predictive model with multiple types of epigenomics data obtained from the same samples.

References

[1] Portela A, Esteller M: Epigenetic modifications and human disease. Naturebiotechnology 2010, 28(10):1057-1068.

[2] Bock C, Lengauer T: Computational epigenetics. Bioinformatics 2008,24(1):1-10.

[3] Laird PW: Principles and challenges of genomewide DNA methylationanalysis. Nature reviews Genetics 2010, 11(3):191-203.

[4] Lim SJ, Tan TW, Tong JC: Computational Epigenetics: the new scientificparadigm.Bioinformation 2010, 4(7):331-337.

[5] Gardiner-Garden M, Frommer M: CpG islands in vertebrate genomes.Journal of molecular biology 1987, 196(2):261-282.

[6] Daura-Oller E, Cabre M, Montero MA, Paternain JL, Romeu A: Specific genehypomethylation and cancer: New insights into coding region featuretrends. Bioinformation 2009, 3(8):340.

[7] Wild L, Flanagan JM: Genome-wide hypomethylation in cancer may be a passive consequence of transformation. Biochimicaetbiophysica acta2010, 1806(1):50-57.

[8] Figueroa ME, Chen SC, Andersson AK, Phillips LA, Li Y, Sotzen J, Kundu M,Downing JR, Melnick A, Mullighan CG: Integrated genetic and epigeneticanalysis of childhood acute lymphoblastic leukemia. The Journal ofclinical investigation 2013, 123(7):3099-3111.

[9] Rhee JK, Kim K, Chae H, Evans J, Yan P, Zhang BT, Gray J, Spellman P,Huang TH, NephewKP, et al: Integrated analysis of genome-wide DNAmethylation and gene expression profiles in molecular subtypes of breast cancer. Nucleic acids research 2013, 41(18):8464-8474.An integrated encyclopedia of DNA elements in the human genome. Nature 2012, 489(7414): 5774.

[10] Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D,Kent WJ: The UCSC Table Browser data retrieval tool. Nucleic acidsresearch 2004, 32 Database: D493-496.

[11] Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2.Nature methods 2012, 9(4):357-359.

[12] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G,Abecasis G, Durbin R, Genome Project Data Processing S: The SequenceAlignment/Map format and SAMtools. Bioinformatics 2009, 25(16):2078-2079.

[13] Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparinggenomic features.Bioinformatics 2010, 26(6):841-842.

[14] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K,Clawson H, Spieth J, Hillier LW, Richards S: Evolutionarily conservedelements in vertebrate, insect, worm, and yeast genomes. Genomeresearch 2005, 15(8):1034-1050.

[15] Love MI, Huber W, Anders S: Moderated estimation of fold change anddispersion for RNA-Seq data with DESeq2. bioRxiv 2014.

[16] Holm S: A simple sequentially rejective multiple test procedure.Scandinavian journal of statistics 1979, 65-70.

[17] Smyth GK: Limma: linear models for microarray data. Bioinformatics andcomputational biology solutions using R and Bioconductor Springer; 2005,397-420.

[18] Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparinggenomic features.Bioinformatics 2010, 26(6):841-842.

[19] Pages H, Aboyoun P, Gentleman R, DebRoy S: String objects representingbiological sequences, and matching algorithms. R package version 2009, 2(2).

[20] Hall MA, Smith LA: Feature Selection for Machine Learning: Comparing aCorrelation-Based Filter Approach to the Wrapper. FLAIRS Conference:1999 1999, 235-239.

[21] A.A. Helal, K.I. Ahmed, M.S. Rahman, S.K. Alam, Breast cancer classifi- cation from

ultrasonic images based on sparse representation by exploiting redundancy, in: 16th International Conference of Computer and Informa- tion Technology, Khulna, 2014, pp.92-97.

[22] N. Karianakis, T.J. Fuchs, S. Soatto, Boosting convolutional features for robust objectproposals, in: Tech. Rep. arXiv:1503.06350, University of California Los Angeles, Mar. 2015.H. Kong, Z. Lai, X. Wang, F. Liu, Breast cancer discriminant feature anal- ysis for diagnosis via jointly sparse learning, in: Neurocomputing, Volume 177, 2016, pages 198-205, ISSN 0925-2312.

[23] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, R. Monczak, Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images, in: Computers in Biology and Medicine. 2013, 43 (10):1563–1572.

[24] C. Loukas, S. Kostopoulos, A. Tanoglidi, D. Glotsos, C. Sfikas, D. Cavouras, Breast cancer characterization based on image classification of tissue sec- tions visualized under low magnification, in: Computational and mathe- matical methods in medicine, Aug 31, 2013.

[25] M. Macenko, M. Niethammer, J.S. Marron, D. Borland, J.Y. Woosley, X. Guan, A method for normalizing histology slides for quantitative analy- sis, in: Proceedings—2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009. Boston, Massachusetts, 2009. p. 1107–1110.

[26] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Man´e, M. Schuster, R. Monga,

[27] S.Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V.Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke,Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[28] M.T. McCann, J.A. Ozolek, C.A. Castro, B. Parvin, J. Kovacevic, Auto- mated histology

analysis: Opportunities for signal processing, in: IEEE Signal Processing Magazine. 2015, 32(1):78.

[29] N. Nayak, H. Chang, A. Borowsky, P. Spellman, B. Parvin, Classification of tumor histopathology via sparse feature learning, in: 2013 IEEE 10th International Symposium on Biomedical Imaging, San Francisco, CA, 2013, pp.410-413.

[30] A. Pˆego, P. Aguiar, Bioimaging 2015, available from: http:// www.bioimaging2015.ineb.up.pt/dataset.html, 2015.

[31] G. Quellec, K. Charri`ere, Y. Boudi, B. Cochener, M. Lamard, Deep image mining for diabetic retinopathy screening, in: Medical Image Analysis, 39, 178–193, Jul. 2017.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, in: International Journal Computer Vision, 115 (3), 211–252L, Apr. 2015.

[33] H. Schwenk, Y. Bengio, Boosting neural networks, in: Neural Computing, 12 (8), 1869–1887, Aug. 2000.

[34] Y. Song, J.J. Zou, H. Chang, W. Cai, Adapting fisher vectors for histopathology image classification, in: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, pages 600–603.

[35] F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, Breast cancer histopathological image classification using convolutional neural networks, in: International Joint Conference on Neural Networks (IJCNN 2016), Van- couver,2016.

[36] F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, A dataset for breast cancer histopathological image classification, in: IEEE Transactions on Biomedical Engineering (TBME), 63(7):1455-1462, 2016.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, Inception-v4, inception-resnet and the impact of residual connections on learning, in: CoRR , abs/1602.07261,2016.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna. Rethinking the inception architecture for computer vision, in: CVPR, 2016.

[39] J. Tang, R.M Rangayyan, J. Xu, I.E. Naqa, Y. Yang, Computer-aided detection and diagnosis of breast cancer with mammography: recent ad- vances, in: IEEE Transactions on Information Technology in Biomedicine, 2009, 13(2):236–251.

[40] T. Tieleman, G. Hinton, Divide the gradient by a running average of its re- cent magnitude, in: COURSERA: Neural Networks for Machine Learning, 4, 2012. Accessed: 2015-11-05.

[41] M. Veta, J.P. Pluim, P.J van Diest, M.A Viergever, Breast cancer histopathology image analysis: A review, Biomedical Engineering, in: IEEE Transactions on. 2014 May, 61(5):1400-11.

[42] Jeffery Li, Travers Ching, Sijia Huang, Lana X Garmire: Using epigenomics data to predict gene expression in lungcancer. Li et al. BMC Bioinformatics 2015, 16(Suppl 5):S10http://www.biomedcentral.com/1471-2105/16/S5/S10