SHINEEKS
DEEPER KNOWLEDGE PUBLISHERS

Analysis and Prediction of Cancer Using Genome by Applying Data Mining Algorithms

Dr. Tejal Upadhyay

Analysis and Prediction of Cancer using Genome by Applying Data Mining Algorithms

Authors

Dr. Tejal Upadhyay
Assistant Professor, Department of Computer Science and Engineering, Nirma University, S G Highway, Ahmedabad, Gujarat, India

ISBN: 978-9-69-339255-5

Published: July 2023
Printed: July 2023

Published by SHINEEKS Publishers
304 S. Jones Blvd #2521, Las Vegas NV 89107, USA

SHINEEKS
DEEPER KNOWLEDGE PUBLISHERS

Notice

Statements and opinions expressed in the book are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

Additional hard copies can be obtained by orders

@https://www.info@shineeks.com

Contents

Preface

In today's era, a lot of data is available, but without proper utilization, there is no use. Data mining is a technique that can be used to retrive some fruitful information from a huge amount of data. Different types of algorithms are used to retrieve knowledge from the data. Sometimes we may derive some interesting patterns also, which can be used for future prediction.

Cancer is a group of conditions where the body's cells begin to grow and reproduce in an uncontrolled manner. These cells can then invade and destroy healthy tissues. The cells become cancerous or malignant because of DNA damage. Collaborative study of Medical Science and Computer Science serves society in a better way. This book is focused on that, the algorithmic methods of Computer Science and the Dataset on which the methods are applied in Medical Science.

The methods like Classification - Supervised Learning and Clustering - Unsupervised learning are the two major topics of Data Mining. In this book, cancer data sets are taken and applied to classification and clustering parameters.

Acknowledgment

Writing a book is harder than I thought and more rewarding than I could have ever imagined. None of this would have been possible without my surrounding communities. Being a Teacher, daily new faces come across me, and my contributions toward their future matter a lot.

I'm eternally grateful to my student community who can be an inspiration for me. I am very much thankful to my whole family without their support this work is not possible for me.

I would like to express my gratitude to all those who have supported me in the creation of this book. Thank you my family and friends for their unwavering encouragement throughout the writing process. I am also grateful to my editor and publisher for their guidance and support. Special thanks to the reviewer for their valuable feedback and suggestions. Finally, I extend my thanks to all the readers who will benefit from this book. Special thanks to my husband and my son.

Abstract

The Genome is the complete sequence of DNA which has all genes. To build and maintain any organism, every genome contains the entirety of the data required. In the human body, more than 300 crores of DNA base pairs are maintained within all cells and their nucleus. The complete study of the genome is called genomics.

Information mining is the course toward finding structures in huge educational collections including techniques at the convergence motivation behind bits of information and database frameworks. Information mining is an interdisciplinary sub field of programming building and estimations with a general goal to discard in- formation (with sharp methods) from an educational assortment and changing the information into a coherent organization for extra use. There are two functionalities- arrangement and grouping which can be applied on information to digest information from the huge dataset.

In this research work, a genome study is done and in that study, we have applied a few data mining techniques like supervised and unsupervised learning on cancer data sets. Getting the dataset is a challenging job and got support from the website Bioconductor and R packages. The implementation detail is divided into four different parts. The first part of the research is based on classification where we have identified Leukemia types by applying classification algorithms. The second part is to identify the subtypes of cancer using nonsupervised learning clustering. The third and fourth parts are focused on different types of clustering methods which can be applied to genome study and also perform some fusion of clusters.

As a part of preprocessing techniques, outliers need to be removed so the applied method to clean the data, then transform and reduce data is applied. In the existing algorithms of data mining, there are many short comings. In this work, an attempt is made to overcome the disadvantages offered by existing algorithms by applying some mix and modified approaches and trying to improve the prediction of the cancer disease from the genomics.

Introduction

CHAPTER ONE

Author
Dr. Tejal Upadhyay

*Corresponding author: Dr. Tejal Upadhyay, Assistant Professor, Department of Computer Science and Technology, Nirma University, S G Highway, Ahmedabad, Gujarat, India, Email: tejal.upadhyay@nirmauni.ac.in

The first part of this Chapter provides a brief description of the Genome of the human body and its functionality and importance. The next two parts describe the Need Analysis and Motivation respectively. The next part explains the objectives, definition, and software used for the present research work. The last part provides a brief outline of the thesis.

Background

A genome is a creature's arrangement of finished DNA, including the total of its qualities. Every genome contains the total of the data likely to fabricate and retain that life form. The genome consists of more than 300 crore DNA sequences and is controlled in all cells that have a nucleus.

The DNA is converted into protein which is a useful product. When the data is stored in our DNA and converted into the process of protein creation, it is called gene ex-expression. Gene expression is a strongly controlled process that allows a cell for answering to its varying atmosphere. (Genomic 2016). Malignancy refers to any of the many uncountable infections described by the progress of uneven cells that isolate enthusiastically and can enter and eliminate normal build tissue.

The tumor often can spread all over the body. Cancer growth is carried about by changes (transformations) to the DNA secret cells. The DNA secret of a cell is bundled into an enormous amount of separate qualities, every one of which covers a lot of directions mentioning to the cell whatever capacities to accomplish, just as how to grow and isolate. Mistakes in the directives can cause the cell to halt its normal purpose and may allow a cell to become Tumorous.

Viruses can be harmful to the computer, similarly, a tumor can be called a virus of different gene expression. The total moment of the cell is controlled by protein initiation and silencing movement of genes. Another part is called gene mutation or alter- ation in regulations in which a gene is not usually switched on and expressed at high stages. Examples of Gene Mutations are Epigenetic, Transcription, Post-transcription, Translation, and post Translation, etc. Cancer is the second-leading reason of death in the world. But survival rates are successful for many types of Tu- more due to advancements in Tumor screening and Tumor treatment.

To study cancer types, a lot of parameters to be studied first. Genome study is also one of those parameters and without the study of Biology, Genome study is almost impossible. In sort to study cancer types, we need to study genome study, for genome study, we need Biology. To make better predictions of cancer Genome studies, Bi-ology studies can find new methods of diagnosing and treating diseases.

For example, cancer-causing genetic changes and epigenetic changes in tumor identification plays an important role. Day by day new methods are identified and used by the therapist to diagnose and cure the patients. By discovering new methods, society is benefited and more and more researchers of Bioinformatics can focus on it.

Over the past few years, research projects on a large scale have begun which re- quires to survey and categorize the genomics modifications associated with a different type of cancer. These initiatives have uncovered unexpected genetic commonalities among various tumor forms. For example, mutations in the HER2 gene have been found in different types of cancers, including breast, bladder, pan-creatic, and ovarian.

Additionally, studies have demonstrated that a single cancer type, such as breast, lung, or stomach cancer, may include several molecular subtypes. Before re- searchers started profiling the genomes of tumor cells, some subtypes of various cancers were unknown to exist.

The outcomes of these initiatives demonstrate the variety of genetic mutations that occur in cancer and lay a foundation for comprehending the underlying molecu- lar causes of this class of illnesses. National Cancer Institute, U S Department of Health and Human Services (2018)

The proposed method is to study and analyze the gene expression details to identify and predict cancer. Information mining is the process of extracting practical knowledge from massive data sets. The large data set of gene expression is used to identify and predict the disease called cancer. This work focuses on supervised learning - classification and un supervised learning -clustering techniques on the genome data sets to identify and predict the disease cancer for early diagnosis. (The National Center for Biotechnology Information 2020).

Need Analysis

Malignancy is a classification of sickness described by uncontrolled cell development and multiplication. The characteristics overseeing the cell advancement and partition should be changed for the creation of malignant growth, these changes are then kept up through coming about cell divisions and are truly present in every risky cell. Gene Expression profiling is a technique used in sub-nuclear science to scrutinize the statement of thousands of characteristics simultaneously. Concerning dangerous development, Gene Expression profiling has been used to even more described Tumors. The information derived from quality ver- banalization profiling as often as possible aides in foreseeing the patient's clinical outcome.

Fast enhancements in genomics and proteomics lately have formed a lot of organic information. Complex computational examinations are required to reach determinations from this information. Bioinformatics or computational science is the in-interdisciplinary study of translating and investigating organic information utilizing data innovation and computational strategies. This region has developed enormously lately because of the hazardous development of organic data produced by established researchers. Bioinformatics is the study of overseeing, mining, coordinating and deciphering data from natural information at the proteomic, phylogenetic, and genomic cell or entire life form levels.

Genome sequencing initiatives have resulted in exponential growth in complete and incomplete arrangement databases, increasing the need for bioinformatics equipment and expertise. As we continue to generate and incorporate massive volumes of genomic, proteomic, and other data, the implications of this new era of bioinformatics will increase.

Information mining is the operation of computerized data scrutiny actions to expose already hidden connections among information effects. Information mining often includes the inspection of information put away in an information distribution center. Information mining helps researchers and explorers remove valuable data from the tremendous sum of organic information within reach by giving advanced procedures. It is a method applied for the revelation of

examples covered up in enormous informational indexes, concentrating on issues identified with their practicality, helpfulness, adequacy, and versatility. A blend of delicate registering (which manages data handling) and information mining in a productive way can viably be utilized for information disclosure in enormous databases.

A precise active zone of exploration in bio-informatics is the idea and enhancement of data mining procedures to tackle natural issues. Breaking down enormous organic data-sets requires understanding the data by gathering arrangements or assumptions from the information. Instances of this kind of examination include protein building prediction, quality order, malignancy arrangement dependent on micro-array data, bunching of value articulation data, truthful show casing of protein-protein connections, and so forth. Hence, there is a possibility to shape the connection between data mining and bio-informatics.

Motivation for the research work

Collaborative study of the medical field and information technology is a trending area in research. Different types of data mining techniques can be applied to cancer research data, which can also be used to forecast the initial stage of cancer. Studying the details of cancer identification and prediction requires data pre-processing like data cleaning, data transformation, and data reductions. Development of recent tools and technologies hence opened new windows of research in this areas. Heteroge- nous medical data can be processed and evaluated by data mining technologies. Statistical analysis plays an important role while processing the data and abstracting knowledge from it. (Upton & Cook 2014).

The unprecedented large-scale views of gene expression can be provided by DNA micro-arrays and as a result, there is a requirement of fundamental measurement tools to study diverse fields of biological systems. Since thousands of individual genes are measured with only a little amount of replication, many typical data analytic methodologies are not applicable in major part.

Empirical Bayes methods offer a natural method for analyzing microarray data since they can greatly reduce the dimensionality of an inference issue while accounting for only a small number of replicates.

The suggested empirical Bayes modeling method permits the replication of expression patterns under various situations. The hierarchical mixture model takes into consid- eration measurement fluctuations, the differential expression for a given gene across cell types, and variations in the average levels of expression of distinct genes. (Up- ton & Cook 2014). The features ranking is based on the Parametric Empirical Bayes method. To extract the feature from it and to estimate the generalization error, Cross-validation (Barrier et al., 2005) method is used.

In this study, Genome data is taken and that data is used for classification and clustering methods. Some samples of Leukaemia are considered and classified into different types of leukemia, similarly clustering algorithms like consensus clustering and fusion clustering is used for the identification of early stages of cancer.

The goal of identifying and predicting cancer is done by applying different algorithms of data mining. Data mining can have different approaches like supervised learning and unsupervised leaning. If a large data set is given and we need to categorize it with some predefined sections, it is called supervised learning classification algorithms. If a large data set is given and we need to identify their category by their characteristics, then we can apply unsupervised learning called clustering algorithms.

The research work aims to apply classification and clustering on cancer genomics data sets and modify existing techniques and/or develop new methods for identifying and predicting cancer with the help of gene expression profiles.

Objectives

- To enhance the link between Computer Studies and Medical Science studies to investigate the utility and effectiveness of Data Mining algorithms that will help society.
- To identify and predict the subtype of blood cancer from the Gene expression by choosing and appropriate classification technique which will help to detect early-stage of cancer.
- To identify cancer sub types, similarity index among clusters, and survival ratio of a cancer patient by applying Data Mining Algorithm.
- To get the optimized number of clusters on large data sets, distributed across many data sets
- To get the optimized result of Silhouette analysis.

Problem Definition

Analysis and Prediction of disease Cancer using Genome by Applying Data Mining Algorithms.

Cancer is a major cause of death worldwide, and an early-stage diagnosis is very helpful to society to decrease the death ratio as well as to cure the disease. In today's era, a lot of data is available for study yet it is not properly utilized. In this study, to get knowledge from the data, different methods are proposed. Class- densification and Clustering are two major parts of Data Mining methods. Leukaemia is one of the types of cancer which can have sub types based on acute or chronic diseases.

TCGA - The Cancer Genome Atlas, a program on cancer genomics, has more than 20,000 molecular characteristics with 33 cancer types. (National Cancer Institute, U S Department of Health and Human Services 2018).

National Cancer Institute - NCI and National Human Genome Research Institute started a joint effort in 2006 to promote researchers from different disciplines and multiple in-institutions.

In this study, a method is proposed to classify the type of Leukaemia from its Gene types. Input is taken as an Expression set matrix with 20172 features of 60 samples. The main process is to filter noisy data and calculate gene ranking. The correlations between genes are calculated and based on cross-validation a training model is generated.

Another part is about applying clustering, in which cancer subtypes are identified for which dataset is taken from TCGA. There are 590 observations and 17815 variables which are converted to a large matrix of 10510260 elements. From these elements, most likely genes and their gene expression values are segregated. The distribution pattern is identified by applying statistics like mean, variance, and also- lute derivation.

Noise cancellation is done by applying an imputation method like mean. The feature selection is applied to find the important features from a huge data set. The features are selected based on most variance, principle component analysis, most variant Median Absolute Deviation (MAD), and Cox model. Finally, clustering techniques are applied and a comparison is generated.

The next part of the study is Monte Carlo Simulation which is a computer simulation technique used to estimate the possible outcomes and a strategy's viability. Clustering is the process of grouping things who are having similar characteristics without overlapping. On social media platforms, we can see the cluster of users to target the advertising of a product or marketing of the product. In this article, gene expressions having similar properties are identified and grouped to identify cancerous behavior. There are so many clustering methods which can have their own merits and demerits.

Clustering methods K-means or K-medoids are very much useful in today's era. These processes are started randomly with every run producing slightly different results or the initialization could provide biased results. The value of K also matters a lot which shows several clusters. The clustering method like Hierarchical cluster- ing not initialize several clusters at the beginning and allows to cut any desired number of clusters.

The concept of Cluster of the cluster is called Consensus clustering which gives more strong results by doing multiple iterations on the sub-samples of the dataset. The process of sampling with sub-sampling will produce more stability of the clusters and decision parameters of K.

The last part of the study is about Gene expression. The DNA is converted into a functional product which is called protein and other molecules. This conversion process is called Gene Expression, and similar gene expression patterns show biological similarity.

The article Wang et al., (2011) says that some times Time series analysis could have intrinsic noise and a B-Spline curve is used for the interpolation between two points. The method of combining a mixed-effects model with a non-parametric smoothing spline gives better results. This combination can robustly stratify genes by their complex time series patterns. These functionalities support the user to visualize and assessing the clustering results for choosing the optimal clustering method. (Pati et al., 2019).

The data points taken over time may have an internal structure like auto correction, and trend or seasonal changes may be accounted and that's why Time Series Analysis is used. The best tool for Time series modeling is Gaussian mixture models. A mixture model can estimate the density of the points by embedding the time series in a higher-dimensional space. For short-to-medium-term forecasting and missing value imputation, this model is used. For a more accurate model, some initial restrictions are applied to the mixture model. Time series forecasting experiments demonstrate that incorporating constraints in the training phase specifically lowers the danger of over fitting in difficult cases with missing values or a significant num- ber of Gaussian components.

The similarity of an object is to be measured and the parameter silhouette value is the best for that. The silhouette value shows how and object is similar to its cluster called cohesion compared to other clusters which are called separators. The range of silhouette is from −1 to 1. The high values show that the cluster is very much similar to its cluster and poorly matched with other clusters. The maximum clusters are having high values that show the appropriate clustering configuration. The law values show the clustering configuration is not so good or we can say poor.

Tools, technology, and dataset

Tools and Technology: Open Source Platform R programming is used. R Programming supports a lot of free methods for statistical analysis as well as genome study. The package Bioconductor and its methods are used in the study. Dataset: The data set for the Leukaemia classifier is Gene Expression Matrix with 20172 features used. The TCGA data set for the identification of cancer subtypes, which has 590 observations and 17815 variables with a large matrix of 10510260 elements are used. Different types of clustering including consensus clustering are performed on a matrix of normalized continuous expression data with the samples at columns and features at rows.

Organization of the book

The research work presented in the thesis is organized and structured in the form of six chapters. The first chapter outlines the introduction. The second chapter discusses the complete literature review with a comparison analysis of similar studies. The major work on classification and clustering are discussed in chapters 3 and 4 respectively. The study details and workflow with mathematical models and meth- ods are mentioned in chapter 4. The result analysis is discussed given in chapter 5 and chapter 6 details the conclusion with future scopes.

Literature Review

CHAPTER TWO

Author
Dr. Tejal Upadhyay

*Corresponding author: Dr. Tejal Upadhyay, Assistant Professor, Department of Computer Science and Technology, Nirma University, S G Highway, Ahmedabad, Gujarat, India, Email: tejal.upadhyay@nirmauni.ac.in

This chapter discusses the literature survey done for the research. A brief description of Data Mining in Biological Data and Medical Data is given in the initial sections of this chapter. The later parts show the detailing of Cancer Genomics, their opportunities in research, and their characteristics. The last part of this chapter showcases a comparative analysis done for the study.

Data Mining in Medical Data

So many organisations are using Data mining intensively and extensively and it is becoming very popular in the domain of healthcare. It is an interdisciplinary where doctors and engineers can work together and it will be of support to both fields.

All parties who are related to healthcare are using Data Mining applications extensively and it is beneficial to society also.

There are a lot of companies that are involved in healthcare insurance, and their major task is to detect fraud and abuse. Sometimes healthcare organizations need support to make customer relationship managing decisions. Data mining ap- plications can give better results and better business too, they are involved everywhere either on the customer side or the employer side. Through these applications, so- city get befitted by getting better and more affordable healthcare services. Physicians are also using data mining applications for diagnosing and treating patients in a better way. The data generated by healthcare services are too com- plex and not in the same format, some reports are text reports, and some are images also. Properly using these data is a challenging task for any company/Healthcare institutions and Data Mining methodologies are the solutions for these types of utilities.

Another application of Data Mining is to find the previously unnoticed patterns and trends in databases and apply that knowledge to create prediction models. (Kin-cade 1998). Alternately, it can be described as the procedure of data selection exploration and model building that makes use of massive data repositories to find patterns that were previously unknown (Milley 2000).

Financial companies/institutions are also using data mining applications intensively and extensively to calculate credit scores or to detect frauds type of plications. Marketers are using data mining applications for direct marketing and cross-selling or up-selling. The retailers are using it for market segmentation and store layout and the manufacturers are using it for quality control and manufacturing schedules.

This section examines the use of data mining in the healthcare industry and its key applications, including the assessment of treatment efficacy, healthcare management, customer relationship management, and the identification of fraud and abuse.

It also gives an example of a healthcare data mining application that entails finding risk factors associated with the onset of sickness to contain losses through the use of data mining methods to assist in the capture of criminals. (Koh et al., 2011). Applications for data mining are frequently used in the business sphere to detect fraud (Christy 1997).

Another issue is that the enormous amounts of data produced by healthcare transactions are too complicated and substantial to be handled and analyzed using conventional techniques. By identifying trends and patterns in vast amounts of complex data, data mining can enhance decision-making. (Biafore 1999) Financial constraints have made it more important than ever for healthcare organizations to base their decisions on the examination of clinical and financial data. While maintaining a high level of care, data mining insights can affect cost, revenue, and operational efficiency (Silver et al., 2001).

Data mining helps healthcare organizations be more prepared to meet future demands. Organizations in the healthcare industry can benefit much from data, but first, they must be converted into information. (Benko et al., 2003). The knowledge that data mining may produce information that is helpful for all parties engaged in the healthcare industry is yet another driving force for the adoption of data min- ing applications in healthcare. For example, By discovering effective therapies and best practices, data mining tools can also help patients and healthcare providers including hospitals, clinics, and doctors. (Gillespie 2000) (Koh et al., 2011)

Other factors boost the popularity of data mining. Applications for data mining can considerably improve the healthcare sector. By first looking at data mining methodology and approaches, categorizing prospective data mining applications in healthcare, and then showing a healthcare data mining application, the goal of this article is to examine relevant data mining applications. Data Analysis and Data mining is a technique and technology that has only lately been established. (Chung & Gray 1999) It searches through massive data sets in search of patterns that are too subtle or complicated for people to recognize to find genuine, innovative, possibly helpful, and intelligible correlations and patterns in data.

According to the Cross-Industry Standard Process for Data Mining, the approach for data mining should include business comprehension, data preparation, interpretation, modeling, evaluation, and implementation. (Koh et al., 2011). Because it establishes the company objectives and serves as the benchmark for data mining projects' success, business understanding is essential. Further, Data is essential to mining, as the phrase "data mining" suggests, mining is impossible without data. Online analytical processing, conventional statistical techniques like cluster analysis, discriminant analysis, and regression analysis, as well as novel statistical techniques like neural networks, decision trees, link analysis, and association analysis, are all features found in the majority of data mining software. Given that data mining has been seen as the child of three different disciplines, namely database management, statistics, and computer science, including artificial intelligence and machine learning, this wide range of methodologies is not surprising.

By employing a standard yardstick, such as lift charts, profit charts, or diagnostic classification charts, the evaluation step permits the comparison of models and outcomes from any data mining model. Finally, Deployment is the process of putting data mining models into practice and making them more logical. Depending on what they are capable of, data mining techniques can be broadly categorized into three groups: description and visualisation, association and grouping, and classification and estimation, or predictive modeling. Understanding a data set, especially a large one, and finding hidden patterns in data, particularly sophisticated data comprising complex and non-linear relationships can be considerably aided by description and visualization.

Finding out which variables go together is the goal of the association. One illustration is market-basket analysis. Data mining is growing in popularity, if not need, in the healthcare industry." (Koh et al., 2011) (the most popular form of association analysis) refers to a method that produces probabilistic sentences like, "Patients receiving treatment A have a 0.35 chance of exhibiting symptom Z." Investigating associative linkages in healthcare can benefit from such information. The goal of clustering is to organize items, like patients, into groups where objects that belong to the same cluster are similar and objects that belong to separate clusters are dis- tinct. To better characterize and comprehend such individuals, readmitted patients are grouped using clustering (Koh et al., 2011).

Predictive modeling is perhaps one of data mining's most widespread and significant uses. forecasting a category target variable, such as forecasting healthcare fraud vs. non-fraud, is referred to as classification. Traditional statistics, such as multiple discriminant analysis and logistic regression analysis, are among the data mining approaches that are frequently employed for predictive modeling. They are also com- prise novel techniques created in the fields of machine learning and artificial intelligence. Neural networks and decision trees are the two most significant models among them. about a bioinformatics learning platform, you can find additional in- formation about data mining methods (Learning Resource Platform 2019).

Biological Data Mining

The briefing about how Data Mining can be helpful for Biological Data.

Introduction

The meaning of Data Mining is to dig the knowledge from a huge data set. It is de- fined as "the process of discovering meaningful new interrelationships, patterns, and modes by finding into large amounts of data stored in a data set". As data sets have gotten bigger and more complicated, data mining is sometimes referred to as the knowledge discovery repository in databases. Data gathering and its organization have become much simpler and easier to manage because of the new technologies of networks, computers, and sensors. However, to make the recorded data more usable, it must be transformed into information and knowledge. The complete process of using computer-based techniques, including fresh approaches for knowledge-based data finding, is known as data mining. Since biological data mining is data-rich, data mining techniques seem to be the perfect fit, but the molecular theory of life's organization is not fully understood. The vast databases of biological data present opportunities and challenges for developing new knowledge-based database discovery techniques. Massive data sets in biology and related fields of the life sciences, like medicine and neuroscience, can be usefully mined to extract knowledge. (Medicine & Neuroscience 2013).

Sequence mining

- Finding sequential patterns among a huge dataset is known as sequence mining. It extracts patterns from a dataset and identifies common substrings. Numerous sectors are getting interested in mining their databases for sequential patterns since enormous amounts of data are being continuously collected.
- One of the most popular techniques, sequential pattern mining has numerous applications, including web-based analysis, customer purchase behavior analysis, and medical record analysis. Sequential patterns can be gleaned from client transaction records in the retail industry.
- For example, after purchasing a bread packet, a customer returns the following time to purchase a butter and milk packet. All of this data may be used by the vendor to analyze client behavior, identify their interests, meet their needs, and anticipate their wants.
- Sequential patterns of symptoms displayed by patients for any disorders, when used to detect strong symptom/disease correlations, can be a very useful source of data for medical diagnosis and preventative medicine. When followed, these sequential patterns provide enormous earnings and raise customer loyalty.
- Sequential data mining aims to find often-occurring yet distinct patterns. Allowing for some noise in the matching process is difficult when trying to find such patterns. The initial stage in developing such a method is to determine the concept of a pattern, followed by the similarity between two patterns. The two patterns' resemblance can be defined differently depending on the application.

DNA sequence mining

It is necessary to study the structure and function of the DNA which is called DNA sequence. The alphabets used by DNA sequences are A, C, G, and T representing the four nitrogenous bases Adenine, Cytosine, Guanine, and Thymine. The DNA sequence of human (The Homo Sapiens) AX829174 (Hoffman et al., 1997) starts with TTCCTCCGCGA and contains 10,011 characters. Finding short repetitive sequences, often of lengths 6 to 15, that regularly appear in a given collection of DNA sequences is a challenge in the sub-sequence mining problem, which is of special significance in biology. The locations of so-called "restrictive regions," which are significant repeating sequences in the biological dataset, can be inferred from these short sequences.

DNA sequencing for human health

Large lengths of DNA, 1 million bases or more, from various individuals, can now be compared by researchers swiftly and affordably. Such comparisons can reveal a wealth of knowledge regarding the role of inheritance in illness susceptibility and response to environmental factors. Additionally, there is a huge potential for diagnoses and treatments due to faster and more

affordable genome sequencing. Even though widespread DNA sequencing in medical offices is still many years away, certain major hospitals have started using sequencing to identify and treat select disorders. For instance, in cancer, doctors are increasingly able to de- termine the specific type of malignancy a patient has using sequence data. The doctor can then choose better remedies as a result. Moreover, DNA sequencing is being used by the National Cancer Institute (NCI)-funded Cancer Genome Atlas project (TCGA) to elucidate the genetics of about 30 different cancer types. Other National Institutes of Health initiatives look at the regulation of genes in various organs and their impact on disease. The development of common and complicated diseases, such as heart disease and diabetes, as well as inherited diseases that cause physical malformations, developmental delays, and metabolic diseases, are all being studied using DNA sequencing in ongoing and future large-scale initiatives.

Cancer Genomics

The Connection between Cancer and DNA

Deoxyribonucleic acid, or DNA, is the substance that carries the instructions that tell cells what to perform. Cells may not behave appropriately when those instructions are incorrect, including proliferating out of control and developing cancer. The so-called mutations that cause cancer can be inherited, but the majority are picked up throughout life. They may result from lifestyle decisions like smoking or environmental causes like toxins. Cancer typically develops through several distinct mutations. To better understand the many types of cancer and dis- cover novel methods to treat them, several new research projects will search for and characterize all the changes in cancer cells.

Genome and its importance

Imagine the DNA in just one of our cells. Our genome refers to this entire collection of DNA instructions. One pair of chromosomes from our mother and one set from our father make up the genome in the majority of cells. Each of these chromosomes contains 6 billion different DNA letters.

There are 26 letters: A to Z in the language English, similarly, there are 4 alphabets in the genome A, G, G, and T. The letters are used to prepare words and words preparing stories of a book, similarly that 4 letters preparing genome sequences. Similarities we can see are that a small change in the letter will change the meaning of the sentences, and a small change in the genome sequence will change the meaning of the genome. The sections of our genome known as genes provide the instructions needed to create the molecules, such as proteins, that carry out the majority of the work in our cells. Genes can be turned on or off by other regions of our genome. Even though all of our cells have the same genes, depending on what they do, different cells will employ different genes. This is significant. For instance, muscle cells use the genes required to produce muscle proteins but not liver proteins.

DNA alterations can lead cells to create the incorrect quantity or an improperly shaped protein that does not function as intended. These modifications may result in health issues since numerous proteins regulate how cells behave. These alterations in cancer cause cells to persist and proliferate out of control, harming neighboring organs.

The Cancer Genome Atlas (TCGA)

The National Cancer Institute and the National Human Genome Scientific Institute at the National Institutes of Health are supporting the groundbreaking scientific project The Cancer Genome Atlas (TCGA). The genetic alterations in more than 20 different forms of human cancer will be found by TCGA experts. Researchers can spot DNA mutations unique to a given tumor by contrasting samples of normal tissue and malignant tissue extracted from the same patient. For each cancer kind, TCGA is examining hundreds of samples. Researchers will have a better knowledge of what distinguishes one cancer from another cancer by examining several samples from numerous different individuals. This is crucial since even patients with the same form of cancer might have drastically varied outcomes or treatment responses. Researchers will be able to create more effective, individualized methods of treating each cancer patient by linking specific genetic changes with specific outcomes.

Role Of TCGA

We shall learn more about how a normal cell becomes a cancer cell with the aid of TCGA. We have previously discovered that there are specific regions of the genome that are frequently impacted in various types of tumors by comparing DNA from normal and malignant tissue. These modifications frequently have an impact on genes that regulate cellular processes that enable cells to divide and survive when they would otherwise perish. We can distinguish one type of cancer from another by identifying specific changes, sometimes known as signatures. These hallmarks aid in the diagnosis of particular cancer kinds, each of which may react differently to various therapies or have a varied prognosis.

Information beyond the sequence

Sometimes, changes don't occur in the DNA's actual sequence. Some of the letters have a distinctive chemical added to them to designate them. These markers alter how these sequences are seen by a cell. They may even alter how well a cell reads them. These markers can therefore alter which proteins are produced in particular cells. The epigenome is the collective name for all of the chemical alterations to your DNA. The regions of the genome with these tags may now be located by scientists. These techniques can be used to determine whether the DNA of a cancer cell contains more or fewer of these marks or whether they are located in different locations. We can start to detect common modifications by cataloging all the variations in the genomes and epigenomes across numerous samples of various cancer types. These alterations might indicate regions of the genome that particular medications can target.

Others will demonstrate a connection between a certain change and the rate at which a disease will advance or the likelihood that the tumor will return following therapy. We are investigating genomes now in the hopes of gaining fresh insights on how to assist patients in the future. The majority of the data we are gathering will not directly impact patients, while some early data are already causing changes in patient care. To transform genomics data into patient care, several steps must be taken, including figuring out which genetic abnormalities are actually to blame for each malignancy and finding or inventing medicines to mitigate their effects.

Health surveys offer a plethora of data on a variety of topics, including the occurrence and frequency of diseases, the prevalence of healthy and unhealthy behaviors, exposure to risk factors, nutritional intake, population physiology measurements, expenses, and the use of health services. Repeated surveys of the same population can be used to determine population patterns since effective statistical techniques can make the findings of an analysis of survey data representative of the population sampled. In addition, following up with people who participated in a baseline survey enables one to track change at the individual level and link pre-existing risk factors to the emergence of diseases. (Kauffman et al. 2017). The literature survey with comparative analysis is given in **Table 2.1.**

Sr. No	Title of the Paper	Methods used
1	Wavelet Analysis in Current Cancer Genome Research: (Meng et al., 2013)	Potentially valuable patterns can be mined by the usage of signal processing methods within the sequence. **Input:** Protein Sequence / DNA sequence **Process:** Numerical Protein Signal (Total Amino Acids 20: (A, R, D, N, C, E, Q, H, G, I, K, L, F, M, P, S, T, V, W, Y) and DNA (C, A, T, G- as binary 1 and 0) **Wavelet Transforms:** Applied Fourier transforms and wavelengths on the biological sequence and finally offers a list of families widely used in biological sequence wavelets. **Output:** Cancer Diagnosis, Mutation Classification, Structure Analysis.
2	A DNA microarray survey of gene expression in normal human tissues (Shyamsundar et al., 2005)	**Input:** 26000 distinct human genes are present in 115 human tissue samples that represent 35 different categories of human tissue. **Process:** Clustering is applied and relevant gene expression patterns with related biological functions are identified. **Output:** The dataset offers a starting point for comparing healthy tissues to diseased tissues and will help identify functions unique to different tissues.
3	Predicting cancer type from tumor DNA signatures (Soh et al., 2017)	**Input:** DNA sequence with gene alternations for 6640 tumor samples with 28 types of cancers. **Process:** AI methods, specifically straight help vector machines with the recursive component choice choose a little subset of quality modifications that are generally instructive for malignant growth type expectation. **Output:** Linear Support Vector Machine (SVM) is the most prescient model of malignant growth type from quality modifications.

4	Functional genomics data sets for the prognosis of cancer (Das et al., 2014)	**Input:** Diverse useful genomics informational collections to recognize atomic marks that can be utilized to anticipate prognostic results for different human tumors. **Process:** 3 methods are used: • Prediction of prognosis using profiles of the expression • Combining expression data with interactive networks to predict prognoses • Other functional genomics data sets for prognosis prediction **Output:** A mix of various layers of data helps expectation precision. Endeavors, for example, the TCGA Pan-Cancer Analysis Project are as of now attempting to join transformation, duplicate number, quality articulation, methylation, micro RNA, proteomic exhibits, and clinical information to recognize drivers from traveler mutations.
5	Cancer tissue sample categorization and validation using support vector machines and microarray expression data (Bernhard et al., 2002)	**Input:** DNA microarray tests quality articulation estimations and cell tests in regards to quality articulation contrasts that will be valuable in diagnosing the ailment. **Process:** Support Vector Machines (SVM) is used to investigate both arrangement of the muscle tests and the information for mismarked or sketchy tissue results. **Output:** Shown the technique in detail on tests comprising ovarian malignant growth tissues, ordinary ovarian tissues, and other typical tissues. The dataset comprises articulation test results for 97802 cDNAs for each tissue. To show the vigor of the SVM technique, two recently distributed datasets from different kinds of tissues or cells are examined. The outcomes are similar to those recently gotten.

6	Analyzing clinical breast cancer data with a hybrid intelligence model that combines feature selection and clustering approaches (Chen & Chien-Hsing 2014)	**Input:** Gene expression clinical data of Breast cancer patients. The target of this examination is to choose notable highlights that can be utilized to recognize fascinating groups in the investigation of bosom malignancy analysis. **Process:** For feature selection, a hybrid intelligent model and wrapper-based methods are used. **Output:** • The clusters worked by a subset of striking highlights are more viable and interpreted-tive than those worked by the entirety of the highlights. • The clustering results give clinical specialists comprehension of the setting of clinical bosom malignant growth analysis

Table 2.1: Survey and comparative analysis.

Opportunities in Cancer Genomics Research

- Although there are constant genetic alterations that lead to new developments in cancer diagnostics and therapy. New progressions and the data got from past genomics exercises could be used to describe the total course of action of driver changes and various innovations to DNA and RNA in various threatening developments. Instructions that take a gender at genomics information from tumors and normal tissue from a comparative patient license pros to discover genomics changes that may drive threatening development. (National Cancer Institute, US Department of Health and Human Services 2018).
- One greater open door is to develop the current usage of genomics procedures to analyze the nuclear reason for trial phenotype. This strategy may help investigators with perceiving hereditary deviations that may perceive amazing tumors from lethargic ones, for test. Systems could be applied to consider the sub-nuclear thought of reaction to a given treatment, similarly as instruments of security from treatment.
- The bounty of data growing up out of threatening improvement genome focuses continuously will be synchronized with patients' clinical narratives and test data. These consolidated outcomes could be utilized to deliver progressively practice-fitted strategies to manage threatening development findings and activity, similarly as to improve techniques for foreseeing contamination risk, perception, and response to activity.
- Genomics gadgets will correspondingly be fundamental for separating results from precision drug test primers, for instance, those present coordinated by NCI's National Clinical Trials Network.

Finding From Literature Survey

- Clinical analysis of malignant growth genomes showed a significant number of various varieties of inherited variants from the normal which is observed in solitary tumors. Additionally, repeated genetic modifications within these tumors occur frequently with only a small number of cases. The challenges for the field are to distinguish which hereditary changes begin to advance malignant growth and discover uncommon genetic modifications that initiate diseases in this way.
- Another challenge is increasing best-grade regular models which are mandatory for genomics, particularly for malignant growth types that are phenomenal or not ordinary, or those not compensated essentially by clinical strategy.
- Developing cell layouts and animal mock-ups that get the varying assortment of human ailments is likewise a deserted need. Mock-ups of exceptional dangerous improvement subtypes may be missing or underrepresented, and there are no models for some irregular hereditary injuries in human illness.
- Managing and looking at the gigantic activities of data drawn in with genomics are extra issues for the field. This zone of assessment needs a beneficial bioinformatics structure and bit-by-bit incorporation guaranteeing data and capacity from cross-disciplinary sets.

NCI's role in cancer genomic research

The discovery of inherited malignant growth institutions is an important piece of NCI's research efforts. The NCI Cancer Genomics Center (CCG) focuses on investigating how modified qualities advance malignant growth. CCG uses high-throughput approaches to identify and consider modifications, immense genome enhancements, increase and decrease in the number of DNA duplicates, material alterations to DNA, and changes in RNA and protein declarations NCI underpins various disease genomics research and related endeavors to interpret these discoveries into clinical advances for patients.

The Foundation also promotes collaborative efforts to promote malignancy genomics exploration and conversation of circumstances and requirements for analysis that could promote new knowledge on etiology, outcomes, and disease threat factors. In October 2018, for example, NCI gathered experts to discuss potential headings on the depiction of mutational marks in malignant growth research.

Characterizing cancer genomes

NCI examiners break down the DNA and RNA of malignant growth cells using trend-setting innovations, such as cutting-edge DNA sequencing to plan the malignant genome scene and find new infection-related changes. NCI thinks generally use diverse genomic methodologies. Planning the results from a

couple of assessments helps scientists with expanding a prevalent appreciation of harm, much like combining red, cyan, and yellow inks can make dynamic concealing prints.

The National Human Genome Research Institute and the National Cancer Institute collaborated to create the Cancer Genome Atlas (TCGA), and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) has portrayed countless genomes and facilitated common models. This enormous number is crucial for identifying DNA, RNA, and protein variants that differ from the norm and cause illnesses in fewer persons.

The Cancer Genome Characterization Initiative (CGCI) additionally considers malignancy genomes, incorporating diseases related to HIV contamination.

CCG depicts sickness genomes through its Genome Characterization Pipeline, which changes over tissue tests given by patients into the great gauge, unreservedly open genomics data.

TCGA and TARGET demonstrated the assessment to organize the essentials of getting patients' clinical data together with illness genomics data, inciting NCI programs that join ironic genomics and scientific datasets (BioConductor 2018).

Novelties Proposed

From the challenges in the field of Cancer Genome Research, sometimes clinically and inherently studies are different. Sometimes experience also helps to identify and predict the disease. If different types and formats of data are available and unless and until not getting or retrieved any knowledge from it, the data is useless. Data Mining is the field of knowledge discovery and that knowledge can be applied to the real-time data of the medical field. The medical field is still emerging, day by day. Innovations and analysis are done to support that innovations and different techniques of Data Mining can be applied to it.

The proposed work is to study and analyze the different genome fields for the identification and prediction of cancer. Supervised and Unsupervised learning is also applied on the dataset which will help us to extract the knowledge from the data. This work can also be extended for different categories of different diseases.

Classification of Leukemia is performed on the Gene Expression dataset. The input Gene Expression CEL...files are prepossessed such that direct gene names can be identified. After that Correlation and interaction between genes are calculated. The Training model is generated by applying an 8-fold cross-validation method. The data set is of all 4 types of Leukemia and one type without Leukemia is used to train the model. Out of 60 samples with 20172 features, 50 samples (10 from each type) are taken to create a training model and remaining 10 samples are used for testing (2 from each type). The Support Vector Machine classifier is used to avoid the difficulties of using linear functions in the high-dimensional feature space.

Clustering to identify sub-types of cancer is done on the data set provided by TCGA. The pre-processing is applied on 590 observations with 17815 variables which cover the large matrix of 10510260 elements. The statistical approach of mean, variance, and absolute deviation is applied to know the distribution pattern, with missing values imputed by mean. Feature selection is done based on top most variant selection. Survival Analysis is also an important feature that is done by the Cox model. A well-known statistical method for examining the association between a patient's survival and various explanatory variables is the Cox model. GeneExp information, days for survival, parameter time, and censored status, either 0 or 1. (0 indicates that Satus is not aware) are chosen. We can set the cut-off threshold for gene expression values to 0.05. After using the Cox model, the similarity matrix between two separate data sets with the use of the clustering algorithm SNF (Similarity Network Fusion is calculated. The similarity index is graphically represented by a Silhouette plot - A measure of how close each point in one cluster is to a point in the neighboring clusters. -Ve silhouette width shows the dissimilarity of the subtypes.

Consensus Clustering is robust clustering which is an agreement between multiple clustering. Monte Carlo reference-based consensus clustering work on Relative Clustering Stability Index (RCSI) and P value to decide how many clusters are needed. Input for this method is a data frame matrix of normalized continuous expression data in which columns are samples and rows are features. Multiple clusterings like k-means, Partition Around Medois, and Hierar- chemical clusterings are implemented. A cluster's P-value, which ranges from 0 to 1, represents how well the cluster is supported by the data. The feature correlation structure of the input data is preserved by the Monte Carlo simulations. Then, to test the null hypothesis that K=1, an empirical p-value is produced for each value of K and used to compare the reference scores with the real values. As a metric for choosing K, the Relative Cluster Stability Index (RCSI) is obtained from a comparison with the reference mean. Based on that, k values different clusters which are plotted for tumor identification.

The EM (Expectation-Maximization) algorithm begins with an initial estimate of a function (such as random) and then iterative updates that function until convergence is observed. An E-step and an M-step make up each iteration.

Numerous longitudinal studies that assess gene expression have as their goal the stratification of the genes according to distinct temporal behaviors. Similar gene expression patterns may represent biologically relevant functional responses. The underlying noise in these observations, however, makes it challenging to cluster their time series. Time series analysis is used to take into consideration the possibility that data points collected over time may have an inherent structure like as auto-correction, trend, or seasonal fluctuation. A useful technique for time series modeling is the Gaussian mixture model.

A mixture model can estimate the density of the points by embedding the time series in a higher-dimensional space. The main components of the suggested method include loading gene expression data into a data frame, showing the data's time series plot, and applying clustering parameters. The clusters can then be validated with silhouette analysis before stability analysis is run.

Leukaemia Classification

CHAPTER THREE

Author
Dr. Tejal Upadhyay

*Corresponding author: Dr. Tejal Upadhyay, Assistant Professor, Department of Computer Science and Technology, Nirma University, S G Highway, Ahmedabad, Gujarat, India, Email: tejal.upadhyay@nirmauni.ac.in

The supervised technique of Data Mining is called Classification in which the data set is categorized in the predefined labels. In this chapter, the classification is applied to the Genome dataset and tested to identify the subtypes of Leukaemia. Leukemia has four subtypes: Acute Myeloid Leukaemia, Acute Lymphocytic Leukaemia, Chronic Myeloid Leukaemia, and Chronic Lymphocytic Leukaemia.

Background

Gene Expression: It is a process by which the information in our DNA gets transformed into a useful product, such as a protein or RNA.

Gene Expression Profiles: When thousands of genes are measured simultaneously to provide a global picture of cellular function, this process is known as profiling.

Classifier: It is a comprehensive package for automatically training and validating a multi-class, Support Vector Based (SVM) based gene expression data.

Types of Classifier: Classifier is a supervised learning algorithm of Data Mining and it has different types based on the data sets and requirements. The detailed classifiers and their pros and cons are shown in **Table 3.1**.

Sr. No	Types of Classification	Advantages	Disadvantages
1	**Logistics Regression:** Using a logical function, the probabilities defining the potential outcomes of a single experiment are simulated.	Most practical for understanding and influence of several on a single outcome variable, independent variables	Effective when the variable is binary and assumes data is free from missing values.
2	**Naïve Bayes:** Based on the Bayes theorem and assuming that each pair of attributes is independent. **Ex:** Document classification and spam filtering	A little amount of training data, and an incredibly quick estimation of required parameters.	Known to be a bad estimator
3	Stochastic Gradient Descent: A quick and effective method for fitting linear models.	Utilized when there are many samples. Different loss functions and classification penalties are supported.	Requires a large number of hyper parameters and is delicate to feature scaling.
4	K-Nearest Neighbors:it is lazy leaning, and merely-maintains instances of the training data rather than attempting to build a comprehensive internal model. Classification derived from the simple majority vote of each point's k-nearest neighbors.	Simple to implement Robust to noisy training data effectively if the training data is large.	Cost is high due to the K Computation distance of each instance to all the training samples.
5	**Decision Tree:** Creates a series of classification rules that may be applied to the data.	Simple to comprehend and visualize, with the ability to handle both categorical and numerical data.	can produce complicated trees that are difficult to generalize.
6	**Random Forest:** Meta estimator that uses average to increase prediction accuracy after fitting number ous decision trees to various subsamples.	More accurate than a decision tree, reduction in overfitting.	Complex algorithm, slow real-time prediction

Table 3.1: Types of classifier.

Types of Leukaemia

Four major types of Leukaemia (Su et al. 2015):

- Acute Chronic Leukaemia - AML,
- Acute Lymphocytic Leukaemia - ALL,
- Chronic myeloid Leukaemia - CML.
- Chronic Lymphocytic Leukemia - CLL

Leukaemia and lymphomas are both included in a larger group of tumors known as tumors of the hematopoietic and lymphoid tissues that affect the blood, bone marrow, and lymphoid system.(Vardiman et al., 2009) (Baba & Câtoi 2007).

In 2015, More than three million people were infected and Leukaemia caused 353,500 deaths. (Kendziorski et al. 2003) (Morris 1983). In 2012, new 352,000 people were infected with Leukaemia. (WHO 2019) Leukaemia is exceptionally regular in children additionally, with 3 fourth of threatening neoplastic sickness cases in children being the intense lymphoblastic sort, (Su et al., 2015) however, with AML and CLL being the most common in adults, adults were diagnosed with around one-ninth of all malignant neoplastic diseases (Su et al., 2015). It occurs frequently and unremarkably in the developed world. WHO (2019).

Algorithm for Leukaemia Classifier

Based on genome-wide expression data, this approach is intended to construct transparent classifiers utilizing geNetClassifier and the related gene networks. (Aibar et al., 2013) Figure 3.1 show the overall graphical representation of the algorithm.

Input: TA sample's expression set or expression matrix serves as the classifier's input.

Output: Three alternative output categories-Gene Ranking, Multiple Classifier, and Gene Networks linked with each class-are shown.

Gene Ranking: Genes and probe sets are entered into the expressionSet and used as classification features. These characteristics are assessed and put into the best class for them.

Classifier: The input has very high dimensions, so we have used a Support Vector Machine classifier. Further, features like optional features and discriminatory power For a particular gene, classifier generalization error, and statistics are found.

Network: Calculated and evaluated are the mutual information, interaction, and co-expression (Correlations) between the genes. This analysis section helps us build gene networks by estimating the degree of association between the genes.

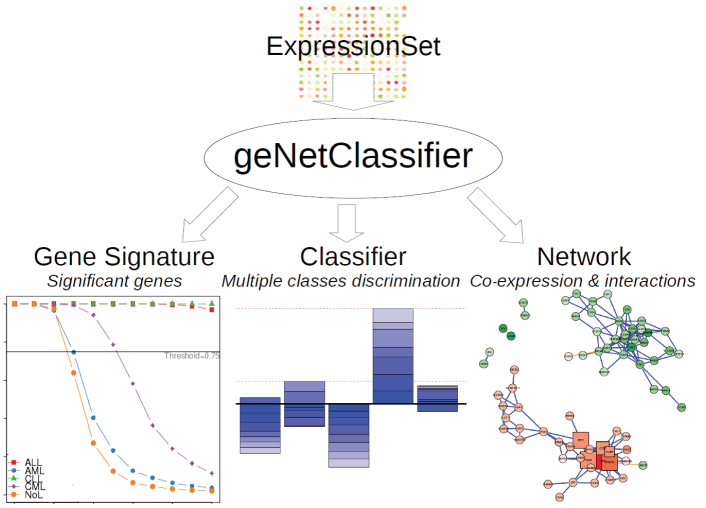


Figure 3.1: Expression set (BioConductor 2018).

Method and algorithmic steps

The algorithm is a combination of machine learning techniques and statistical methods. A Parametric Empirical Bayes approach (PEB) and Double-nested internal cross-validation (CV) are used to pick the features) (Barrier et al., 2005). A multi-class Support Vector Machine (SVM) is the machine learning technique used in the classifier. (Meyer et al. 2008). The relationships produced from gene-to-gene co-expression analysis and the interactions derived from gene mutual information analysis are calculated to create the gene networks (Winkler et al., 1999).

Algorithm: Classifier

Input: LeukemiasEset.

Output: Categories of Leukemia.

A. Preprocessing

1. Input: LeukemiasEset produced by MILE (\1) project
2. Large Matrix X = Load the Expression profile with 20172 features of 60 samples
3. Gene Matrix Y[ ] = CDF_Func(X)
4. For (i=0; i<= n; i++)
5. {

Gene_ID Z[i] = Y[i]
Gene_Symbols (A[i]) = Gene_Labels(Z[i])
For (j=0; j<=Gene labels; j++)
Annotation Details (B[j]) = Abstract Gene table information (A[i])
}

6. Protein coding P[] = Protein_Gene(B[])

B. Gene Ranking – Classifier (Training samples = 50 and Testing samples = 10)
Q[] = Func_Classifier(P)
Total Genes = 26 (Classwise: ALL-9, AML-5, CLL-1, CML-5, NoL-6)
Post. Prob. of Each Gene class pair= Func_Expectation Max Methods(Q[ ])
For (Type = ALL, AML, CLL, CML and NoL)
{
Posterior Probability[Type] >95%
Significant Genes[type] = Gene names
}
Plotting Significance gene diagram with
Gene Rank on X Axis and Posterior Probability on Y-Axis.

C. Classifier: Support Vector Machine is used (Effective in high dim. data)
For (Type = ALL, AML, CLL, CML, NoL)
GeneSelectionProcedure() {

Each cross-validation starts with the first-ranked gene of each class
and added one more gene in every iteration.

Apply 8-fold cross-validation several times
}

D. Estimation of Performance and Generalize Error
Calculations of the following parameters {
Sensitivity TP -
Specificity TN
Matthews Correlation coefficient (MCC)
}

Figure 3.2: Algorithm 1: classifier.

Work flow

In addition to returning the genes ranking and additional information about the selected genes, geNetClassifier constructs the Classifier and the gene network associated with each class. The advancement internally is followed by the subsequent steps:

- The first step is Filtering information and calculating the gene ranking.
- The next step is to Calculate correlations between genes.
- The third step is to Calculate the interactions between genes.
- Through 8-fold cross-validation, we have chosen the subset of genes to train the classifier on. The classifier is trained with the entire set of samples using the selected genes.
- Performance estimation: It determines the classifier's generalization error and, consequently, the statistics regarding the genes, adding 5-fold cross-validation around the classifier's construction (nested cross-validation).
- Building the gene networks: Using the pairwise gene-to-gene correlations and interactions, a factor network is built for each of the classes.

Experimental results

Gene Ranking: Based on the analysis of the expression signal, first we need to determine the gene ranking for each class. The task of a classifier is to assign the class for each gene. The classifier's next step is to assign the most appropriate class to each gene such that separates each class and is optimized. Finally, each gene can be filled with only the class most appropriate for that gene.

Dataset: Leukaemia Eset, The patient's bone marrow was used to collect a total of 60 microarray samples, representing the various types of leukemia: AML, ALL, CLL, CML and NoL.

60 input samples: Twelve of each type and one with no Leukaemia- are chosen for the experiment (12*5=60). Twelve are selected, ten from each category are to be used as a training set, and two are to be used as testing samples. A total of 50 samples are used for training, and 10 samples are used to test the classifier.

Input for classifier: For several samples and SampleLabels, an Expression Set-a genome with a broad expression matrix-is used. (A vector with each sample's class name or an object with this information)

geNet Classifiier: The classifier is constructed using the gene network linked to each class, and it also provides the ranking of the genes and further details about the genes that were chosen.

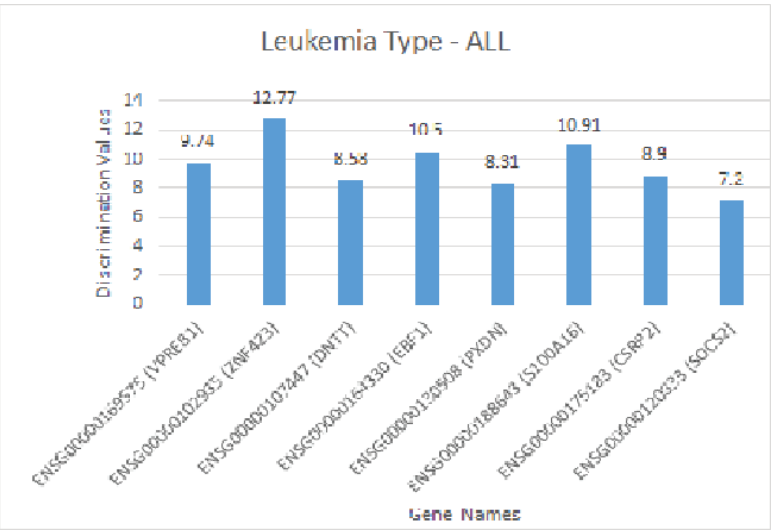
No of trained samples	Number of total Genes	AML	ALL	CML	CLL	No Leukemia
50	34	5	8	4	3	14

Table 3.2: Classifier Sample.

Leukaemia classifier discriminant Power: A parameter called discriminant power that measures the importance that the classifier internally gives to each gene which shows the characterization of the genes to separate different classes (i.e. different diseases or diseases subtypes compared). The overall discriminant power values of all categories like AML, ALL, CML, CLL, and NoL are given in figures 3.4, 3.3, 3.6, 3.5, and 3.7 respectively.

Table 3.3 shows the detailed gene names of **Figure 3.3**. **Table 3.4** shows the detailed gene names of **Figure 3.4**.

For each category of Leukaemia types, significance genes play a major role and that role is represented in the diagrams of Gene names Vs Discriminant power values.



Name of Gene	Discriminant Power
ENSG00000169575 (VPREB1)	9.74
ENSG00000107447 (DNTT)	8.58
ENSG00000102935 (ZNF423)	12.77
ENSG00000130508 (PXDN)	8.31
ENSG00000164330 (EBF1)	10.5
ENSG00000120833 (SOCS2)	7.2
ENSG00000188643 (S100A16)	10.91
ENSG00000175183 (CSRP2)	8.98

Table 3.3: Subtype: ALL.

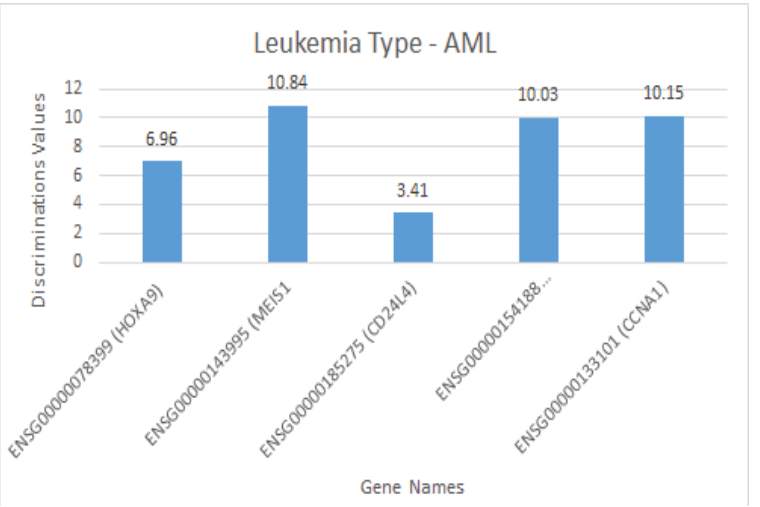


Figure 3.4: Subtype: AML.

Name of Gene	Discriminant Power
ENSG00000185275 (CD24L4)	3.41
ENSG00000078399 (HOXA9)	6.96
ENSG00000143995 (MEIS1)	10.84
ENSG00000133101 (CCNA1)	10.15
ENSG00000154188 (ANGPT1)	10.03

Table 3.4: Subtype: AML.

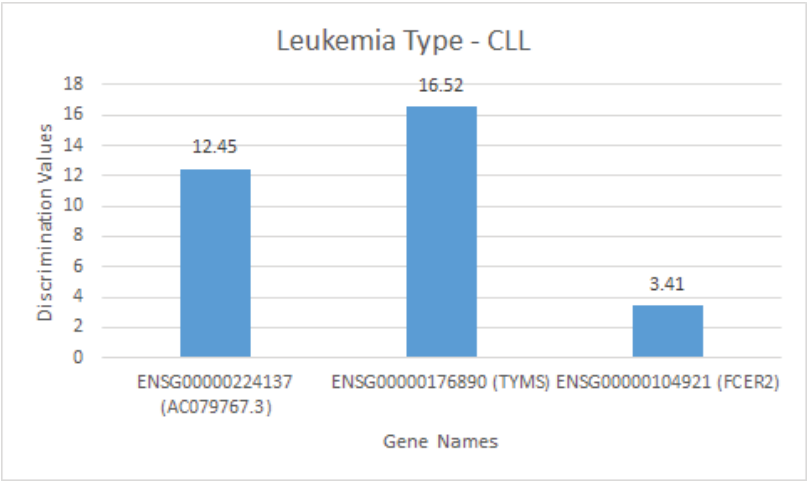


Figure 3.5: Subtype: CLL.

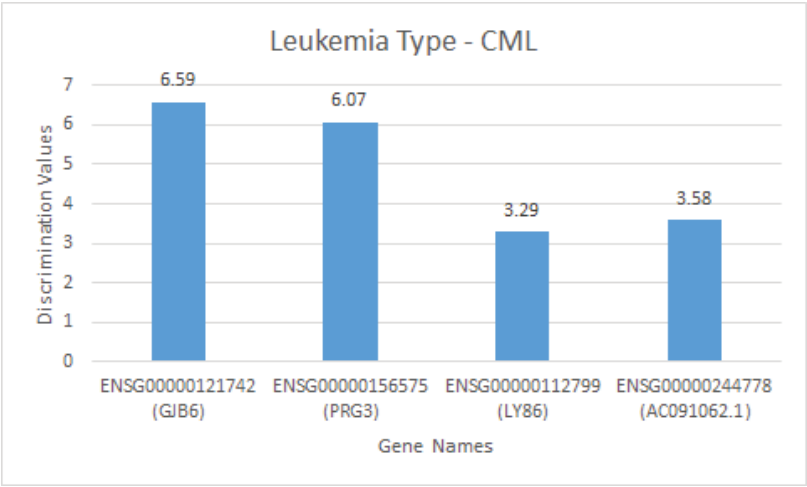


Figure 3.6: Subtype: CML.

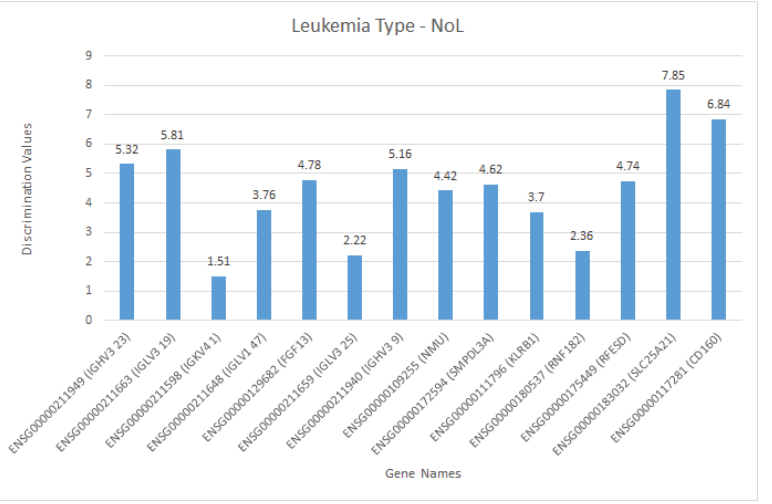


Figure 3.7: Subtype: NoL.

Name of Gene	Discriminant Power
ENSG00000211663 (IGLV3 19)	5.81
ENSG00000211949 (IGHV3 23)	5.32
ENSG00000211648 (IGLV1 47)	3.76
ENSG00000211598 (IGKV4 1)	1.51
ENSG00000129682 (FGF13)	4.78
ENSG00000211659 (IGLV3 25)	2.22
ENSG00000180537 (RNF182)	2.36
ENSG00000109255 (NMU)	4.42
ENSG00000172594 (SMPDL3A)	4.62
ENSG00000111796 (KLRB1)	3.7
ENSG00000117281 (CD160)	6.84
ENSG00000175449 (RFESD)	4.74
ENSG00000183032 (SLC25A21)	7.85
ENSG00000211940 (IGHV3 9)	5.16

Table 3.5: Subtype: NoL.

Overall Output of Classifier

The overall output is divided into three different sections: Gene Ranking, Classifier, and Gene Networks., Table 3.6 shows the Top Ranking Genes for all the classes. The classifier shows how many genes are in each category including No Leukaemia, based on the name of genes we have categories. **Table 3.7** shows which are the significant genes that are taking part to identify the Leukaemia type, and Table 3.8 shows the count of significant genes for each category. Gene Networks will create clusters of genes whose functionalities are almost similar to each other.

	ALL	AML	CLL	CML	NoL
1	VPREB1	HOXA9	AC079767.3	GJB6	IGHV3-23
2	EBF1	ANGPT1	NUCB2	AC091062.1	IGLV147
3	DNTT	CD24L4	FCER2	LY86	IGKV41
4	ZNF423	MEIS1	TYMS	PRG3	IGLV319
5	S100A16	TRBVB	PNOC	TRIM22	IGLV325
6	PXDN	CCNA	RRAS2	ABP1	FGF13
7	SOCS2	HOXA5	RRM2	LPXN	NMU
8	CSRP2	ZNF521	C6orf105	NLRC3	IGHV3-9
9	COL5A1	NKX23	UHRF1	TNS3	KLRB1
10	CTGF	DEPDC6	KIAA0101	GBP3	SMPDL3A

Table 3.6: Top ranking genes for the class.

ALL	AML	CLL	CML	NoL
883	236	1683	746	174

Table 3.7: Number of rank significance genes.

ALL	AML	CLL	CML	NoL
2551	3274	3025	2738	3262

Table 3.8: Genes(LeukaemiasClassifier@genesRanking.

Number of ranked significant genes (posterior probability greater than 0.95): The the complete number of qualities in the positioning for each class can be questioned utilizing the capacity of the function num Genes(). These numbers incorporate all the qualities that have some capacity to recognize classes, albeit just the best ones which are truly critical. The significant genes are plotted in **Figure 3.8**.

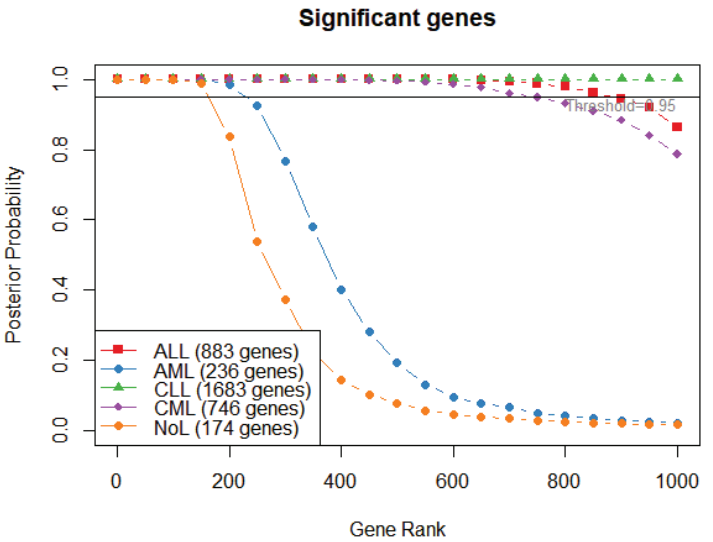


Figure 3.8: Significant genes:(BioConductor 2018).

Classifier: The ideal scope of genes to mentor the classifier is picked by assessing the classifiers prepared by expanding the scope of genes. This should be possible through the exploitation of numerous emphases of the Eight-fold cross-validation technique. The emphases of cross-validation begin with the essential hierarchical qualities of each class: at that point, it prepares an encased classifier with these genes, and assesses its performance.

One more gene is included in each progression to those classifications that an 'amazing forecast' isn't accomplished. The genes are pulled from each category's gene ranking to determine which category has the greatest variety of genes (maxGenesTrain=100) or till zero error is reached (continue Zero Error=FALSE).

Gene mistakes are discovered but cannot be preserved during each cycle of the cross-validation loop. The base assortments of genes per class are saved once the loop's execution is complete to reduce errors.

The three slots that Classifier uses to store data are Classifier, Classification Genes, and Generalisation Error. Classifiers use SVMs, and Classification Genes choose the final genes that are typically used to create classifiers and calculate generalsation errors.

The SVM classifier, which may later be utilized to conduct research, is contained in the classifier slots. Classification genes contain a definitive gene that is chosen to make the classifier. The classification qualities even have data concerning the discriminant power of the qualities. The discriminant power might be a boundary that was derived from the classifier's Support vectors looks like the office of each factor to check the qualification between classifications.

Gene selection Procedure: By comparing classifiers trained with an increasing range of genes, which may be accomplished by numerous cycles of the Eight-fold cross-validation approach, the best range of genes to teach the classifier is selected. The principal hierarchical component for each class is the starting point for each cross-validation iteration. Next, the classifier is trained using genes, and its ef- effectiveness is assessed. To achieve the highest level of stability within the intended gene range, the cross-validation is repeated several times with fresh samples.

Estimation of performance and generalization error procedure: In each cycle of this loop, numerous tests have neglected the instructing and utilized as check tests. This step allows us to estimate and supply insights and metrics concerning the norm of the classifier and hence the genes which are elite for classification.

The last decision is done to uphold the genes to choose everything about iterations. By selecting the best set of genes while excluding outliers chosen during cross-validation cycles, the highest hierarchical genes for each category are determined. This enables the detection of a steady range of genes while considering sampling defenses. The following **Table 3.9** shows calculations on the number of genes selected for each class in the 5 runs of the 5-fold cross validation which is applied for the estimation of performance. These numbers allow us to explore the number of genes that are used per class.

Gene Networks: Every category's factor networks are built to support the association parameter between genes. These association parameters, which are calculated on all the samples of each category of the analyzed dataset, square the factor-to-factor co-expression. They are obtained from factor-to-factor interactions and correlations. Table 3.9 shows the attribute outline of the GeneNetwork.

Type	Number of Nodes(genes)	Number of Edges (Relationships)
ALL	883	1511
AML	236	184
CLL	1683	16280
CML	746	4007
NoL	174	1018

Table 3.9: Attribute outline of the GeneNetwork.

Two text files are produced as the output, one for each form of leukemia. Both files are text files and one has information about nodes while the other contains information about edges. The nodes that interact (gene1, gene2), the type of link (correlation or interaction), and the value of the relationship are all included in the edges file. The networks can be exported to external software for additional processing using these flat text files.

Novelties Proposed

Every iteration of the EM-based pre-processing algorithm ensures that the likelihood function can be enhanced and that a local maximum can be reached.

EM algorithm - the first step in seeking expectations(Expectation Step), known as the E step;the second step for maxima(Maximization Step), known as a step-by-step M.

EM algorithm is used to calculate the principle based on incomplete data and maximum likelihood estimation.

Each gene-class pair's posterior probability is determined, and it shows how much each gene sets one class apart from the others.

The best value is represented by 1 and the worst by 0. By contrasting the data of one class with all the other samples, the posterior probability enables the identification of the genes that exhibit meaningful differential expression.

The Classifier A support vector machine (SVM) is a supervised machine learn- ing model that uses classification algorithms for two-group classification problems.

They have two key advantages over other methods, such as neural networks: greater speed and improved performance with a small number of samples (in the thousands). As a result, the approach works well for classification issues where it's typical to only have access to a dataset with a few thousand tags on the samples.

Summary

This implementation work has given three main results:

- Gene Ranking.
- Classifier.
- Gene Network.

Gene ranking: It can provide a wide range of subclasses of genes for cancer. The ranking is determined by allocating each component to the category in which it is the most straightforward to rate. In this method, the categorization is optimized, allowing the technique to select the genes that best distinguish any of the groups first. This strategy ensures that each factor only ranks in one category, even if a factor is shown to be related to too many categories during the expression analysis.

Classifier: By using an OneVsOne (OvO) technique and the Support Vector Machine, this solution enables multi-class classification. All binary categoryifications are unit fitted, and the correct class is discovered to have supported an electoral system. Sensitivity (True Positive), Specificity (True Negative), Matthews coefficient of correlation (MCC), a life of the standard of binary classifications, and international Accuracy are the classifier's live parameters. The chosen samples included a part of accurate results.

Gene Network: The factor networks are created using association parameters be- tween genes for each category. These association parameters, which are factor-to-factor co-expression, are computed using correlation and factor to factor interactions produced from mutual data (MI) analysis, each is computed on all samples of each category of the examined dataset.

Clustering Focusing on Cancer Study

CHAPTER FOUR

Author
Dr. Tejal Upadhyay

*Corresponding author: Dr. Tejal Upadhyay, Assistant Professor, Department of Computer Science and Technology, Nirma University, S G Highway, Ahmedabad, Gujarat, India, Email: tejal.upadhyay@nirmauni.ac.in

In this chapter, the basics of unsupervised learning - clustering is explained. The initial part shows the types of clustering and their comparison. The later part of this chapter focuses on how the advanced clustering technique Monte Carlo Clustering is engaged in cancer studies with some experimental results.

Introduction

Clustering is the process of grouping the data with almost similar properties. A major part of this chapter is how clustering is helpful to identify cancer sub-types from the genome; what is the concept of consensus clustering, and how For example, studies on cancer can use many types of clustering.

Analysis and Prediction of Cancer using Genome by Applying Data Mining Algorithms

Clustering and Types Of Clustering

In essence, it is a kind of unsupervised learning technique. A technique called "unsupervised learning" allows us to make inferences from data sets that just contain input data without any labeled replies. Typically, it is employed as a method to identify the groups, generative properties, and meaningful structures present in a collection of cases.

The objective of clustering is to divide the population or set of data points into several groups so that the data points within each group are more similar to one another and different from the data points within the other groups. It is essentially a grouping of objects based on how similar and unlike they are to one another.

Types of Clustering

- **Hierarchical clustering:** Using the data points' top-to-bottom hierarchy, clusters are formed.
- **Partitioning methods-K means:** Data points are grouped into clusters based on centroids and their distance from the cluster centroid.
- **Distribution-based Clustering:** Clusters are created using a variety of metrics, such as mean, variance, etc., based on the probability distribution of the data.
- **Fuzzy Analysis Clustering:** This algorithm uses fuzzy cluster assignment as its clustering mechanism. The way this algorithm works is virtually identical to that of k-means, which assigns data points to clusters based on distance. However, as was already indicated, this approach allows for the placement of a data point in more than one cluster.
- **Gaussian Mixed Models (GMM):** It is a probabilistic strategy for classifying the observations in a fuzzy manner. Each cluster's probability of being in each observation is calculated, and the most likely cluster is typically assigned to each observation to produce a classification. The interpretation of alleged categories can also be done using these probabilities.

Comparison of clustering methods

Clustering plays an important role in this study **Table 4.1** shows the clustering methods and their advantages and disadvantages:

Sr.No	Types Clustering	Advantages	Disadvantages
1	Hierarchical Clustering	Simple to use, the number of clusters need not be determined beforehand, dendrograms are simple to read.	Cluster assignment is rigid and cannot be changed, it is also time-consuming and in-effective for larger datasets.
2	Partitioning methods	Easy to construct, quick processing, works with larger data sets, straightforward to read results	The number of centroids must be determined a priori because the newly formed clusters have irregular sizes and densities. noise- and outlier-affected.
3	Distribution-based Clustering	Working with real-time data, the number of clusters need not be predetermined, and the metrics are simple to comprehend and adapt.	Slow and complex algorithms are incapable of scaling to larger data sets.
4	Fuzzy Analysis Clustering	can be used with data that are heavily overlapping, higher rate of convergence	The number of centroids must be specified a priori, Affected by outliers and noise, with a slow approach that is not saleable.
5	Gaussian Mixed Models	Lot more flexiblein terms of cluster co-variance, it is a generalization of the K- Means / Partitioning method.	

Table 4.1: Comparison of clustering methods.

Cancer Subtypes Using Clustering

The main goal of the suggested approach is to use genome data to use unsupervised learning to identify the subtypes of a disease cancer.

Method

The Cancer Genome Atlas (TCGA) data Portal provided the data set for this investigation. Clinical details, genomics characterization data, and high-level sequence analyses of the tumor genomes are also included. These data sets are freely accessible for use in research.

The package used in R programming is called Bioconductor, using R programming, we can download other packages and apply various processing to the data. Results may be gathered and presented. The process is shown in **Figure 4.1.**

- **Step 1:** Download the TCGA data set.
- **Step 2:** To transform data sets from TCGA that may be used by R tools, use the RTGCA package. 590 observations and 17815 variables make up this data collection.
- **Step 3:** We transformed it into a big matrix with 10510260 elements to load and process 590 observations and 17815 variables.
- **Step 4:** Out of 10510260 items, we selected the most probable gene names and their accompanying Gene Expression values, and preprocessed the data on those.
- **Step 5:** We used the dataset's mean, variance, and absolute derivation distribution to determine its distribution pattern.
- **Step6:** To deal with missing values, preprocessing has been used. It is not advised to reject records with partial missing values in microarray gene expression data. The typical approach is to infer the appropriate values. Imputation by mean and imputation by microarray are two of the three frequently used imputation techniques.
- **Step 7:** On the data set's imputed values, feature selection is used.
- **Step 8:** The following four feature section techniques are used:
  - The most variance.
  - The most variant Median Absolute Deviation (MAD).
  - Principal Component Analysis.
  - Cox regression model.
- **Step 9:** Different Cluster algorithms are applied:
  - Using consensus clustering to identify cancer subtypes.
  - Consensus Identification of cancer subtypes using non-negative matrix factorization.

- Integrative clustering for identifying cancer subtypes.
- Identification of cancer subtypes using a Similarity Network.
- SNF and CC ensemble approach for cancer subtype identification.
- Weighted Similarity network fusion for the identification of cancer sub-types.

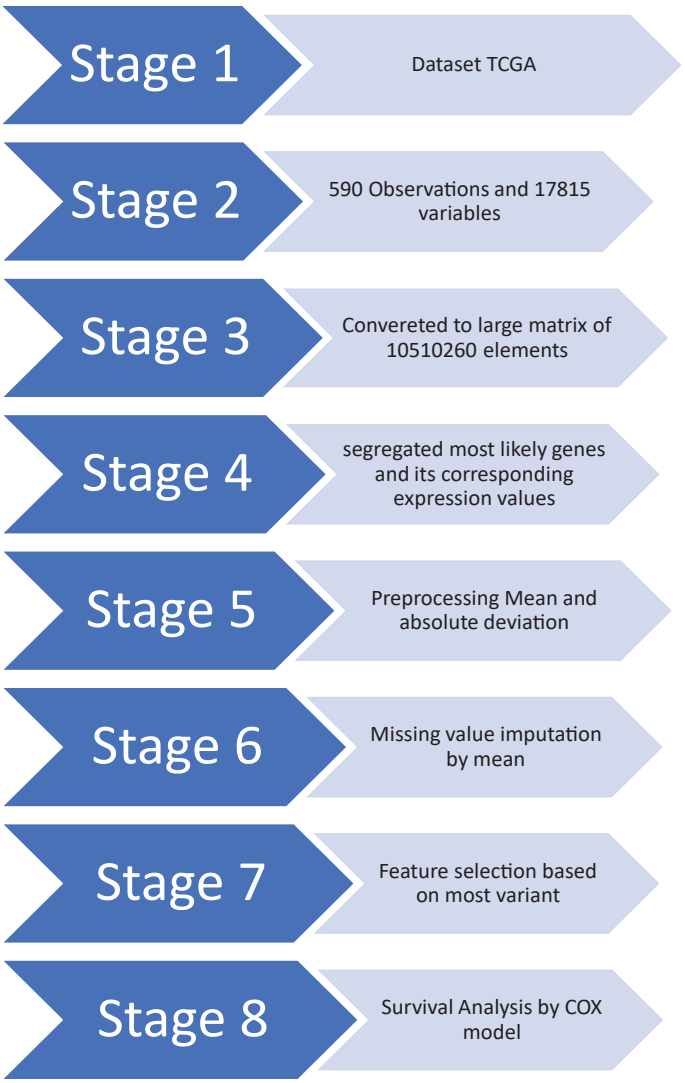


Figure 4.1: Cancer subtypes identification.

Algorithm for Cancer Subtypes Identification

Figure 4.2 shows the basic algorithmic steps of cancer subtypes identification.

```
---------------------------------------------------------------------------
Algorithm2: Cancer Subtypes using Clustering
---------------------------------------------------------------------------
Input: The Cancer Genome Atlas (TCGA) portal.
Output: Measuring Parameters:
        Silhouette Analysis
        Survival Analysis
        Statistical Significance
        Differential Expression Analysis

TCGA = Download the data from TCGA Portal
TCGA' = Func_RTCGA(TCGA) (500 observations and 17815 Variables)
Large Matrix of 10510260 elements A[][] = Load TCGA'
Most likely gene names and Gene expression values C [] []= Func_mostlikely_Gene
and _Expression Values (A[] [])
D [] [] = Preprocessing (C [] [] )
   {
        Imputation by mean(), Variance() and Absolute Derivation  distribution()
  }
 E [] [] = Func_feature Selection(D [] [])
   {
        Variance()
        MAD ()
        PCA()
        Cox_Model()
   }
F [][] = Func_Clustering()
   {
        Consensus Clustering()
        Consensus Non-Negative Matrix Factorization()
        Integrative Clustering ()
        Similarity Newwork Fusion()
        Ensemble method SNF() CC()
   }
```

Figure 4.2: Algorithm 2: Cancer subtypes identification.

Data Preprocessing

A data processing method called ”WeighData pre-processing” may entail reformatingting data in an understandable manner. Real-world knowledge typically has gaps, is inconsistent, lacks specific behaviors or trends, and contains several inaccuracies. The method for decomposing these issues could be established as knowledge pre-processing.

There are four feature selection methods of pre-processing as shown in **Table 4.2** utilized on the same dataset. The input and output formats are identical across all information processing techniques.

Variance-Analysis,	MAD-Median Absolute Deviation	An estimate of the treatment effect on survival - COX model	PCA-Principal Component Analysis
Its purpose is to practice value reduction and price management. They are gauges of unfolding, but they also suffer from extra-typically low values. The quality deviation is often the simplest method for determining to unfold when the information is conventional.	It is a solid example of non-normality and information being displayed together. If you have non-traditional knowledge, one data item we will employ instead is the MAD.	It is a regression model, which is a type of applied mathematics typically used in medical analysis to examine the relationship between a patient's survival time and one or more predictor factors.	A statistical technique known as main parts transforms a set of observations of supposedly linked variables into a set of values of unrelated, linearly orthogonal variables.

Table 4.2: Analysis techniques.

There are three methods for feature selection and implantation:

- Based on the majority of the variance.

  The top 1000 features with the most variation The features with variance > 0.5.

- Considering the majority of possibilities Absolute Median Deviation The top 1000 features with the most variation.

  The features with variance > 0.5.

- Based on Principal Component Analysis, dimension reduction, and extraction.

Data Clustering

Unsupervised learning is done using the clustering technique from data mining. Consensus Clustering, also known as (Monti et al., 2003), is a technique for quantifying the number and membership of potential clusters within a dataset, such as microarray gene expression. This technique has grown in prominence in the field of cancer genomics, where new disease molecular subtypes have been identified. (Hayes et al., 2006), (Verhaak et al., 2010).

The Consensus Clustering method determines the clustering of specified cluster counts (k) by subsampling from a set of elements, such as microarrays. The proportion of times two items appeared in the same subsample that they occupied the same cluster is then determined using pairwise consensus values and recorded in a symmetrical consensus matrix for each k. The consensus matrix is condensed into a variety of graphical displays that let the user choose an appropriate cluster size and membership. Consensus Clustering is openly accessible (Gene Pattern 2017). For a formal description (Monti et al., 2003), Consensus ClusterPlus (Wilkerson & Hayes 2010) carries out Consensus Clustering in R and enhances it with new functions and graphical outputs to help users find classes.

The following clustering techniques are implemented in this article to identify cancer subtypes: The comparative analysis of various clusters is displayed in **Table 4.3**.

- CC - Consensus Clustering.
- CNNF - Consensus Non-negative matrix factorization.
- IC - Integrative Clustering.
- SNF - Similarity Network Fusion.
- Ensemble method of SNF and (CC).
- WSNF - Weighted similarity network fusion.

Variance-Analysis,	**Consensus Non-negative matrix factorization (CNNF)**	**Integrative Clustering (IC)**	**Similarity Network Fusion (SNF)**	**Ensemble of SNF and CC**	**Weighted similarity network fusion (WSNF)**
Frequently employed and highly beneficial method with numerous successful applications in genomic studies (Monti et al., 2003)	Utilizing efficient dimension reduction techniques for high-dimensional genomic knowledge, unique molecular patterns (Magidson 2013)	A combined latent variable model for many omics data types. (Ritchie et al., 2015)	A fusion similarity network approach for aggregating data from many omics (Bersanelli et al., 2016)	Combining CC and SNF to provide a novel way of cancer detection (Wang et al., 2014)	A technique using knowledge of the gene regulatory network

Table 4.3: Comparative study of clustering algorithms.

Experimental Results

The computationally determined cancer subtypes should display distinctive molecular patterns and correspond to biological implications. The following parameters are used to evaluate the outcomes.

Silhouette Width (Rousseeuw 1987)

- The resulting clusters' separation distance can be examined using silhouette analysis. By showing a measure of how close each point in one cluster is to a point in the surrounding clusters, the silhouette plot provides a visual method to assess variables like the number of clusters.
- A high number shows that the sample is closely matched, and the purpose of this is essentially to demonstrate the similarity between two datasets.
- In the silhouette plot, each horizontal line corresponds to a single sample.
- The silhouette width refers to the line's length.
- The silhouette plot will be in the opposite direction if the cancer types are not matched, hence, samples with negative silhouette width are said to be dissimilar to one another.

Choice of parameters in the proposed methods: **Table 4.4** shows the measuring parameters to identify the subtypes of cancer.

Silhouette Width	Survival Analysis	Statistical significance of the clustering	Differential expression analysis
Whether a sample and its known sub-types are comparable or not, used to live. A high price suggests that the sample and the data are matched.	used to evaluate the various survival patterns among subtypes	It is purely a statistical distribution of subtype-specific differential data.	Examining the expression distinction between each subtype and a reference cluster is customary.

Table 4.4: The measuring parameters.

Monte carlo clustering

This topic describes the details of Monte Carlo and Consensus Clustering. Genome data is widely used for clustering and the crucial part is to select several clusters. The Monte consensus clustering algorithm uses stability selection to estimate K. Based on simulated and actual expression data from The Cancer Genome Atlas, M3C has been shown to correct the inherent bias of consensus grouping. (TCGA). This section displays the gene expression. Smoothing and clustering using Gaussian Mixed Effects Models.

Unsupervised learning has a crucial characteristic called clustering. Consensus clustering is a unique type of clustering that is utilized when multiple clustering techniques are used and it is desired to identify a single clustering that makes more sense than the other clustering techniques currently in use. (Strehl & Ghosh 2002).

This work uses microarrays to demonstrate consensus clustering, also known as aggregation of clustering.

Consensus Clustering

Consensus clustering is comparable to ensemble learning in supervised learning for unsupervised learning. All of the clustering methods now in use have some drawbacks. This could make it challenging to understand the results, especially if the number of clusters is unknown.

To count the number of clusters in the data and evaluate the stability of the found clusters, consensus clustering offers a method that expresses the consensus across numerous runs of a clustering algorithm.

To account for the clustering algorithm's sensitivity to the beginning conditions, the approach can also be used to express the consensus over several runs of a clustering algorithm with random restart (such as K-means, model-based Bayesian clustering, etc.).

Consensus A technique called clustering can be used to count and count the members of potential clusters within a dataset of gene expression. In cancer genomics, where novel molecular subclasses of the disease have been identified, this technique has grown in favor.

Monte Carlo Reference Base Consensus Clustering

When determining the ideal value for clustering, Monte Carlo Reference based consensus clustering generates all distributions of stability scores within the range of K using a Monte Carlo simulation. Genome Clustering Using Monte Carlo Rather Than Consensus A widely used method for group recognition using the standard of soundness determination is Data For Tumour Identification. (Antolin et al., 2020). M3C enhances the restrictions of accord bunching. M3C utilizes the Relative Clustering Stability Index (RCSI) and p esteems to choose the estimation of K and reject the invalid theory, k=1. It overcomes the drawbacks of consensus clustering by using a reference that preserves the correlation structures of the input feature. **Figure 4.3** shows the algorithmic steps.

Algorithm3: Monte Carlo Clustering

Input: A matrix of normalized continuous expression data taken from the TCGA glioblastoma (GBM) dataset

: Where columns are samples and rows are featuredOutput: CDF Plot

PAC score Vs. number of Clusters PlotRCSI Vs. K Plot

X = Load the Matrix

Data = Load the Object Mydata (Expression data from GBM)

Info = Load the object desx (Annotation data: Patient's age, sex, etc)

Y = Filtering(X) {Func_Variance();

Func_PCA(); //Adding label and categorical information into data}

Z = normalized the data in homoscedastic(Y) PCA Plot() //Shows the distribution of clusters

Func_M3C(Z) //default 100x Monte Carlo iterations and 100X innerGet RCSI = Max (RCSI-Relative Cluster

Stability Index values)

For RCSI (True)

{

Retrieve KPlot CDF

Plot PAC score Vs. number of ClustersPlot RCSI Vs. K

}

Figure 4.3: Algorithm 3: Monte carlo clustering.

Operations on M3C

- **Input:** A Matrix Data Frame of normalized continuous expression data (Mimicroarray), Columns shows sample and rows as features.
- **Preprocessing:** Dimensionality reduction is the process of filtering that is applied. Variance and p-value are used for unsupervised and supervised learning respectively.
- **Outliers:** They are detached using PCA (Principle Component Analysis) function.

- Clustering:
  - **K Means:** Samples can be separated into n clusters of equivalent variance Algorithm isolates a lot of N tests into k disconnect groups C and each is depicted by the methods.
  - **PAM:** Partition Around Medoids: Similar to K-means, but data points are centers instead of medoids.
  - **HC:** Hierarchical Clustering.
    - Nested Clustering by combining or splitting.
    - Tree Representation.
    - The root is a unique cluster.

Gene Expression Clustering With Gaussian Mixed Effects Models and Smoothing

Introduction

Gene expression data conceal important information that is necessary to comprehend the biological processes that occur in a certain organism in connection to its environment. Our comprehension of functional genomics can be greatly enhanced by uncovering hidden patterns in gene expression data. Due to the complexity of biological networks and the large number of genes present, it is challenging to comprehend and analyze the massive amount of data that results from these observations.

As a result, using clustering techniques is the first step in resolving these issues. Clustering techniques are crucial in the data mining process because they enable the discovery of natural structures and intriguing patterns in the underlying data. Understanding gene functions, cellular processes, and cell subtypes, as well as mining usable information from noisy data and understanding gene regulation have all been demonstrated to be possible thanks to the clustering of gene expression data.

A further advantage of clustering gene expression data is the discovery of homology, which is crucial for vaccine development. To identify and impart important knowledge about the best clustering technique that would provide stability and a high degree of accuracy in its analysis procedure, this paper looks at the many clustering algorithms suitable to gene expression data.

Depending on the kind of dataset, clustering can be carried out using genes, samples, or temporal variables. Gene expression data must be clustered to be meaningful, Genes form a cluster that shows related expression across conditions, whereas samples form a cluster that shows related expression across all genes.

In gene-based clustering, the samples are thought of as the features and the genes as the objects. With sample-based clustering, which treats the samples as objects and the genes as features, the samples can be divided into identical groups. The difference between sample-based and gene-based clustering is like the clustering tasks for gene expression data.

Clustering can be partial or complete, whereas complete clustering assigns each gene to a cluster, partial clustering does not. Given that gene expression data sometimes includes some unrelated genes or samples, partial clustering tends to be more appropriate for gene expressions. Partial clustering in gene expression permits some genes in the expression data not to belong to well-defined clusters because most of the time, genes in the expression data could represent noises, allowing their impact to be correspondingly less on the outcome. In addition, by not permitting some genes in the expression data to belong to well-defined clusters, it helps in neglecting quite a few irrelevant contributions.

By allowing membership of unrelated genes rather than imposing it, partial clustering helps prevent scenarios where an intriguing subset in a cluster is kept. Hard or overlapping clustering are two different types. While overlapping clusters give each input gene varying degrees of membership in different clusters, hard clustering places each gene during operation and output in a single cluster. Assigning each gene to the cluster with the dominant degree of participation will convert an overlapped clustering into a hard clustering.

The workflow

Figure 4.5 shows the workflow in which the first step is to load the gene expression data into the frame. The time series plot is designed for that data and clustering parameters are set. After designing the clusters, performance analysis is done with the help of silhouette analysis. Finally, the application of clustering is mentioned. **Figure 4.4** shows the algorithmic steps and **Figure 4.5** shows the flow chart.

Stability analysis and optimal clustering

Clustering is based on a statistical model, and the Expectation Maximisation (GlsEM) technique is used to conclude. The EM method may become trapped in a local optimum while executing the clustering algorithm. The likelihood of the local optimum solution is decreased, making it suboptimal. To determine how frequently the algorithm becomes trapped in local optima and how this local op- tima differ from the optimal clustering solution, it is highly advised to do a stability analysis of the clustering.

Algorithm4: Gene Expression Clustering

Input: Gene expression data

Output: = Optimal Clustering

X = Load the data from the SPEM package of Bioconductor

Funct_Set Time slots() // A time series gene expression data set contains an ensemble of time series vectors, each time series is associated with a gene.

No of Rows = no of time series vectors // Gene Names

No of Columns = Time Points // Set hours 2 hr, 4 hr etc

Data Frame = Time series Data Set (X)

Func_Clustering(Data Frame , number of clusters n) //Perform repeatedly for optimal result

Func_EM() // Expectation Maximization Technique)

Func_stability Analysis() // get the most like hood values

Func_Plot Clusters()

Func_Plot Silhouette Analysis() // for 3 clusters

Func_Plot Silhouette Analysis() // for 4 clusters

Z = Load the data //Input Yeast Time series Data

Func_Clustering(Z) // 10 times with k = 4 , time plots = 0, 24, 48, 63, 87….

Y = Optimal Clustering,

Figure 4.4: Algorithm 4: Gene expression.

Figure 4.5: Flow graph of clustering implementation.

Silhouette Analysis

The term "silhouette" describes a technique for analyzing and verifying consistency inside data clusters. The method gives a brief graphic representation of how accurately each object has been identified.

The silhouette value expresses an object's cohesion, or similarity to its cluster, about other clusters, or separation. A strong clustering coherence is indicated by a high average silhouette width. The visualization of a silhouette plot, which is described in the chapter on result analysis, is the easiest way to examine silhouette widths for the data points in a clustering.

EM algorithm

Clustering is based on Expectation Maximized (EM) Algorithm

- **E-Step:** Estimation of the lost variables in the dataset.
- **M-Step:** Maximize the parameter of the perfect in the occurrence of data.

Silhouette Analysis: It is a process of understanding and validation of stability within groups of data. Silhouette value is the amount of how comparable an item is to its cluster compared to other clusters, the range is from -1 to +1. -1 indicated poor matching and +1 indicates good matching.

Result Analysis

CHAPTER FIVE

Author
Dr. Tejal Upadhyay

*Corresponding author: Dr. Tejal Upadhyay, Assistant Professor, Department of Computer Science and Technology, Nirma University, S G Highway, Ahmedabad, Gujarat, India, Email: tejal.upadhyay@nirmauni.ac.in

This chapter describes and details the result analysis. The initial parts briefly discuss the Research outcomes. The later part shows what type of results are achieved in the overall research. The results and the outcomes of the following are also discussed. Leukemia subtypes Classification, Clustering for cancer subtypes, role of a Monte Carlo clustering in Genome and Gene Expression Clustering.

Research Outcomes

- Medical data is taken as input and Data Mining algorithms are applied to it, so the collaborative study allows Medical Science and Computer Science.
- Supervised and Unsupervised learning methods are used in such a way that where the categories are fixed. Classification algorithms are used and where the categories are not fixed, clustering methods are used based on the data characteristics.
- A gene expression data of project MILE is taken as input with 60 samples and 20172 features to classify the subtypes of leukemia.
- The support Vector Machine classifier is used to identify blood cancer types as it supports high dimensional data sets.
- Mean, Variance type of statistical data is used to fill up the empty data of large data sets.
- Future work can be to enhance this research work with new disease analyses like Covid-19
- The complete research work is done on open-source tools and technologies.
- The research enhancement and future work give a chance to young and dynamic researchers to see the new era of research.
- The classified and clustering results are visualized in diagrams.
- The measuring parameter Silhouette analysis shows whether the data is a part of that cluster or not, or how much it is similar which is given with the silhouette analysis diagram.
- The similarity index impact by different numbers of clusters on time series data is shown in the result analysis chapter.
- RCSI Reference Clustering Stability Index decides the optimal cluster count by applying Monte Carlo Simulation based clustering works on distributed data sets

Result Analysis

Classification

Genes are ranked for each class in the Classifier's first step based on an analysis of the expression signal. The classifier's final step is to place each gene in the gene class that is most appropriate for it. By doing this, the algorithm will select the first genes that best distinguish any of the classes, optimizing the separation between them. As a result of this procedure, a gene can be filtered to only one class that is the most appropriate for it if it is connected with multiple classes during expression analysis. The proposed algorithm takes input from different Genomics data of Leukemia patients and builds a classifier and the associated gene networks on genome-wide expression data. A total of 60 samples are taken, 12 of each class (ALL, AML, CLL, CML, and NoL). 10 samples from each

class will work as a training set and 2 will be as a testing set. Thus out of 60, a total of 50 samples are used for the training set and 10 samples for the testing set.

The workflow **Figure 5.1** shows the workflow to classify the leukemia type.

- **LeukemiasEset:** Leukemia Expression set- a large matrix data file can be downloaded, which is of size 20172 X 60 and has 17557 features in 60 samples.
- The format for this file is sample names: GSM330153.CEL GSM331677.CEL GSM330151.CEL. (60 total)

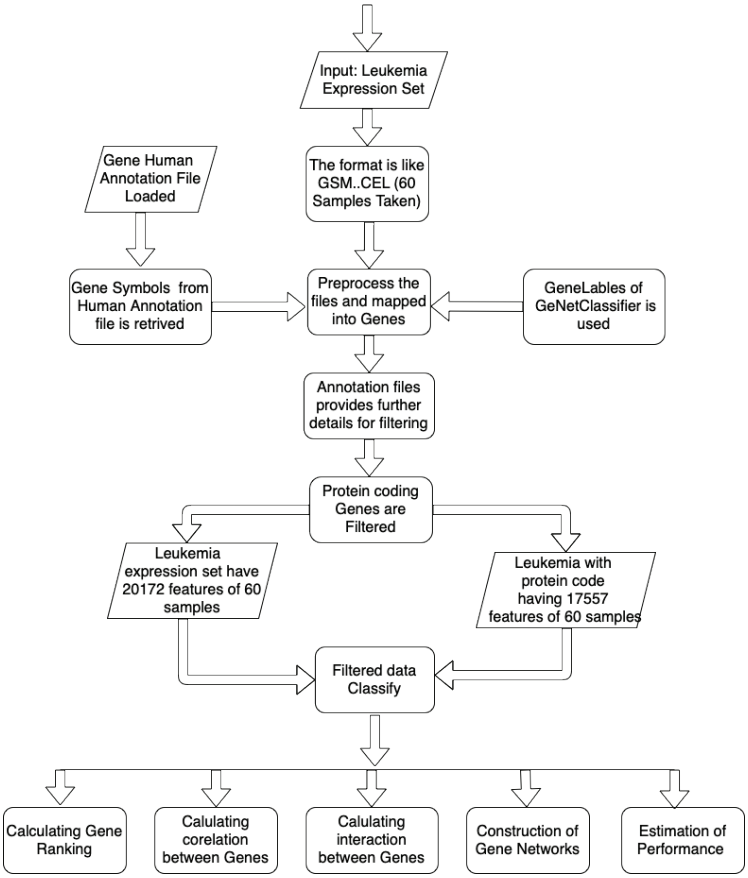


Figure 5.1: Workflow of leukemia classifier.

- These.CEL files are preprocessed and allow mapping the expression directly to genes (Ensembl IDs ENSG), for that gene labels of geNetClassifier are used.
- GeneSymbols can be retried by loading genes-Human-annotation.R
- The annotation files provide information on genes that can be filtered. We have retrieved only proteins coding genes that are required for the building of the classifier which we have used for the filtering.
- Out of a total of 60 samples, we have taken 50 samples as a training set in which a total of 34 different genes are identified. These 34 genes are categorized into different types of leukemia like AML, ALL, CML, CLL, and NoL. The results are shown in **Table 3.2**.

Tables 5.1, 5.2, 5.3, 5.4, and 5.5 are the genes which we have taken as and input for the category ALL, AML, CLL, CML, and NoL respectively.

Sr.No	Sample Name	Project Name	Tissue	Leukemia Type
1	GSM330153.CEL	MILE1	BoneMarrow	ALL
2	GSM330151.CEL			
3	GSM330157.CEL			
4	GSM330154.CEL			
5	GSM330174.CEL			
6	GSM330171.CEL			
7	GSM330182.CEL			
8	GSM330178.CEL			
9	GSM330186.CEL			
10	GSM330185.CEL			
11	GSM330201.CEL			
12	GSM330195.CEL			

Table 5.1: Gene expression sample data - ALL.

Sr.No	Sample Name	Project Name	Tissue	Leukemia Type
1	GSM330532.CEL	MILE1	BoneMarrow	AML
2	GSM330559.CEL			
3	GSM330546.CEL			
4	GSM330571.CEL			
5	GSM330566.CEL			
6	GSM330580.CEL			
7	GSM330574.CEL			
8	GSM330593.CEL			
9	GSM330584.CEL			
10	GSM330611.CEL			
11	GSM330603.CEL			
12	GSM330612.CEL			

Table 5.2: Gene expression sample data - AML.

Sr.No	Sample Name	Project Name	Tissue	Leukemia Type
1	GSM330934.CEL	MILE1	BoneMarrow	CLL
2	GSM330933.CEL			
3	GSM330979.CEL			
4	GSM330969.CEL			
5	GSM330982.CEL			
6	GSM330980.CEL			
7	GSM330999.CEL			
8	GSM330987.CEL			
9	GSM331009.CEL			
10	GSM331004.CEL			
11	GSM331048.CEL			
12	GSM331037.CEL			

Table 5.3: Gene expression sample data - CLL.

Sr.No	Sample Name	Project Name	Tissue	Leukemia Type
1	GSM331378.CEL	MILE1	BoneMarrow	CML
2	GSM331377.CEL			
3	GSM331382.CEL			
4	GSM331381.CEL			
5	GSM331386.CEL			
6	GSM331383.CEL			
7	GSM331388.CEL			
8	GSM331387.CEL			
9	GSM331390.CEL			
10	GSM331389.CEL			
11	GSM331393.CEL			
12	GSM331392.CEL			

Table 5.4: Gene expression sample data - CML.

Sr.No	Sample Name	Project Name	Tissue	Leukemia Type
1	GSM331661.CEL	MILE1	BoneMarrow	NoL
2	GSM331660.CEL			
3	GSM331666.CEL			
4	GSM331663.CEL			
5	GSM331670.CEL			
6	GSM331668.CEL			
7	GSM331672.CEL			
8	GSM331671.CEL			
9	GSM331674.CEL			
10	GSM331673.CEL			
11	GSM331677.CEL			
12	GSM331675.CEL			

Table 5.5: Gene expression sample data - NoL.

Each samples have so many genes. **Figure 5.2** shows Genes priority for each class which is calculated and a probability of 0.95 or more can be taken for that disease.

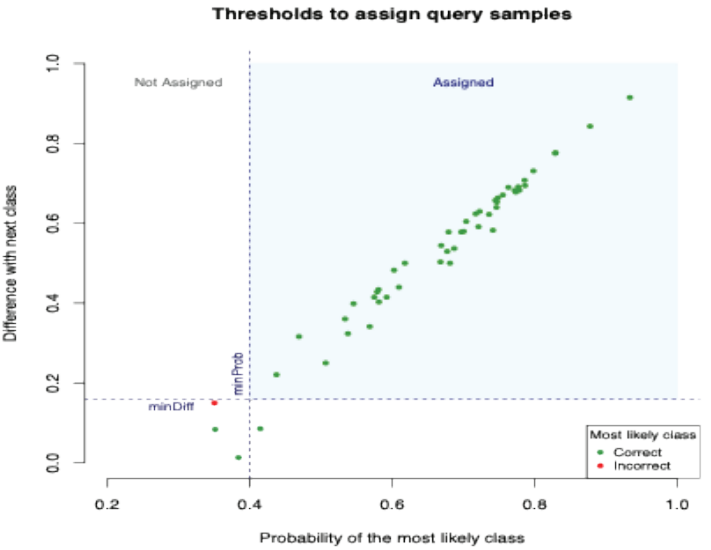


Figure 5.2: Most likely gene class.

As per **Figure 5.3**, Diseases A, B, and C have some common Genes, so for further classification we are using, the genes which are not overlapping and have maximum probability in their disease. The Blue color area is of significance genes selection area.

Classifier: Classifier Support Vector Machine is used. Inquiries can be made using the SVM classifier. Using a One-versus-One (OvO) technique, this implementation enables multi-class classification by fitting all of the binary classifications and identifying the appropriate class. Classifier will build the final number of genes as per the type of leukemia. (number of support vectors = 29).

8-fold cross-validation is used in multiple iterations of the gene selection process. The cross-validation is performed multiple times with fresh samples rather than simply once. A parameter number of iterations are passed to the classifier and based on that during every iterations we are getting, the max genes are selected with minimum error rate. **Figure 5.4** shows the iterations steps and **Figure 5.5** shows number of genes selected in each iteration.

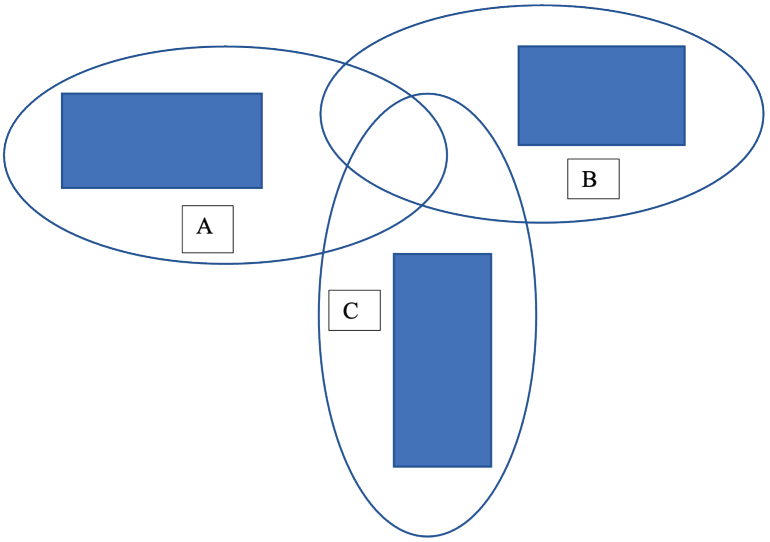


Figure 5.3: Gene selection area.

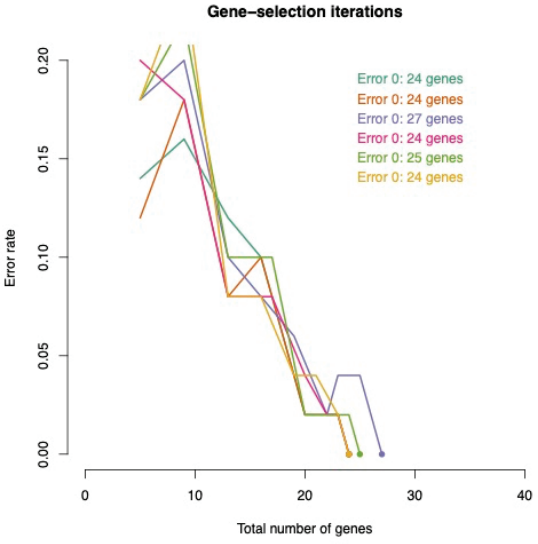


Figure 5.4: Gene selection iterations.

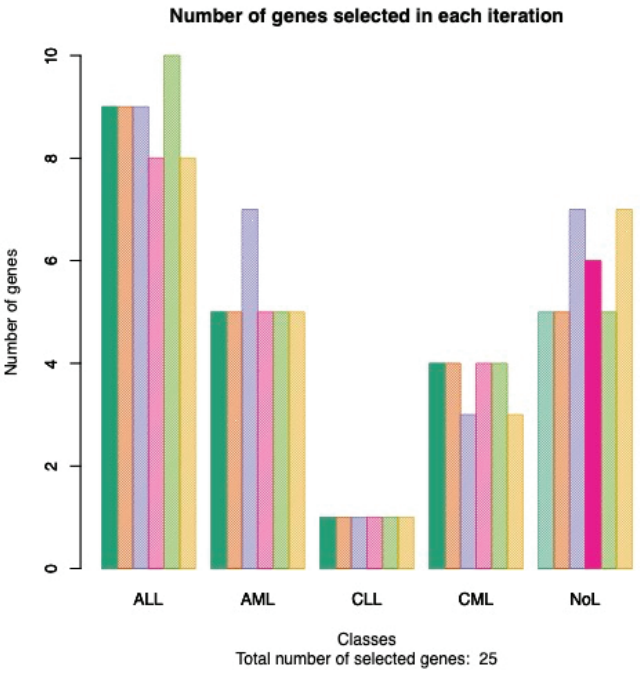


Figure 5.5: Number of Genes selected in each iteration.

The maximum number of genes selected with a minimum error rate would be what the support vectors count.

The Measuring parameters for Classifier Sensitivity: Percentage of samples belonging to a specific class that were correctly identified.

$$Sensitivity = \frac{TP}{(TP + FN)} = Truepositiverate$$

Specificity: Percentage of samples belonging to the class to which they were allocated. (True Negative Rate).

$$Specificity = \frac{TN}{(TN + FP)} = TrueNegativeRate$$

Matthews Correlation Coefficient (MCC): It measures the quality of binary classifications. It is a measurement that accounts for both real and imagined advantages and drawbacks.

$$MCC = \sqrt{\frac{(TPXTN - FPXFN)}{[(TP + FP)(TP + FN)(TN + FP)(TN + FN)]}}$$

Call Rate per class and Global Call Rate: Percentage of assigned samples in a class or across all of the predictions.

$$CallRate = \frac{Assigned}{((Assigned + NotAssigned))}$$

Table 5.6 shows Performance Estimation and Generalization Errors.

	Accuracy	Call Rate		
Global	**100 Sensitivity**	**90 Specificity**	**MCC**	**Call Rate**
ALL	100	100	100	90
AML	100	100	100	70
CLL	100	100	100	100
CML	100	100	100	100
NoL	100	100	100	90

Table 5.6: Performance estimation and generalization error.

Table 5.7 shows the Leukemia Subtypes probability. The probabilities of assignment to each class, which were determined using cross-validation, are shown in the figure. This probability matrix gives a decent idea of how simple or difficult it is to classify each sample. Additionally, it gives a hint as to how likely it is that one class will be confused with another.

	ALL	AML	CLL	CML	NoL
ALL	0.697	0.060	0.073	0.067	0.102
AML	0.697	0.060	0.073	0.067	0.102
CLL	0.697	0.060	0.073	0.067	0.102
CML	0.697	0.060	0.073	0.067	0.102
NoL	0.697	0.060	0.073	0.067	0.102

Table 5.7: Probability of all leukemia subtypes.

A confusion matrix is shown in **Table 5.8** This is used to swiftly visualize and assess a classification algorithm's performance. While the columns show the class that the samples were assigned to, the rows represent the samples' actual class. Therefore, the diagonal is where the appropriately allocated samples are located.

	ALL	AML	CLL	CML	NoL	Not Assigned
ALL	9	0	0	0	0	1
AML	0	7	0	0	0	3
CLL	0	0	10	0	0	0
CML	0	0	0	10	0	0
NoL	0	0	0	0	9	1

Table 5.8: Confusion matrix.

Clustering for Cancer Subtypes

The Cancer Genome Atlas (TCGA) data Portal provided the data set for this study. The tumor genomes' high-level sequencing analysis, genomic characterization data, and clinical information are all available through TCGA.

Workflow

In this work, the clustering methods comparisons are shown.

- Consensus Clustering (CC) is the most frequently used method for genomic data (Monti et al., 2003).
- Consensus with non-negative matrix factorization which gives effective results in case of dimension reduction (Magidson 2013).

- Integrative clustering is used when we have multiple types of omics data (Ritchie et al., 2015).
- Similarity Network Fusion (SNF) and SNF with weighted data are used for the aggregation of multi-omics data (Bersanelli et al., 2016).
- The combination of SNF and CC is used for cancer identification.

The computational methods used to identify cancer subtypes should be consistent with biological interpretations and show distinctive molecular patterns. Silhouette Width, Survival Analysis, statistical significance of clustering, and differential expression analysis are some of the characteristics used to evaluate the results.

- The separation distance between the generated clusters can be investigated using silhouette analysis. In essence, the silhouette plot illustrates similarities between two datasets by displaying a measure of how close each point in one cluster is to points in the neighboring cluster.

  In the silhouette plot, each horizontal line corresponds to a sample, and its length is referred to as the silhouette width.

  Figure 5.6 shows the similarities and dissimilarities of clusters. The positive width shows the similarity and the negative width shows dissimilarity.

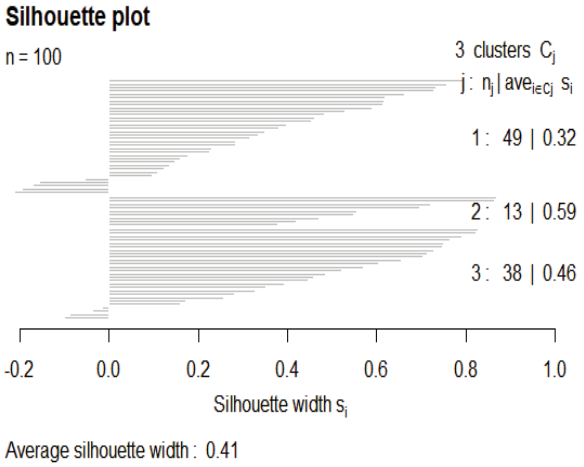


Figure 5.6: Silhouette analysis.

- The results of a survival analysis reveal various patterns of cancer subtype survival. It is employed to evaluate the various survival patterns among sub-types.

  A statistical technique called the chi-square test is used to compare actual outcomes with predictions. **Table 5.9** shows the results.

	N	Observed	Expected	$(O-E)^2/E$
Group 1	46	44	41.3	0.173
Group 2	40	39	27.3	4.990
Group 2	14	13	27.4	7.529

Table 5.9: Different survival patterns among subtypes.

This test's goal is to establish whether a discrepancy between actual and expected data is the result of chance, or if the variables under consideration are related in some way.

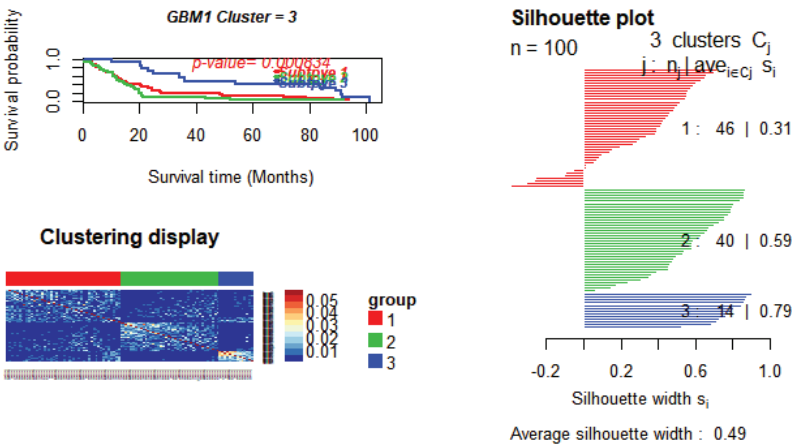


Figure 5.7: Survival analysis.

Survival Analysis is a combination of three parts as shown in **Figure 5.7** with three parts:

- Survival curves,
- Heat map of the sample similarity matrix,
- Silhouette width plots for the identified cancer subtypes.

  The detected cancer subtypes have been used to reorganize the samples in each plot. This type of graphic offers clear results that are simple to assess.

- Statistical significance shown in **Figure 5.8** is a purely statistical approach to test the significant difference in data distribution between subtypes.

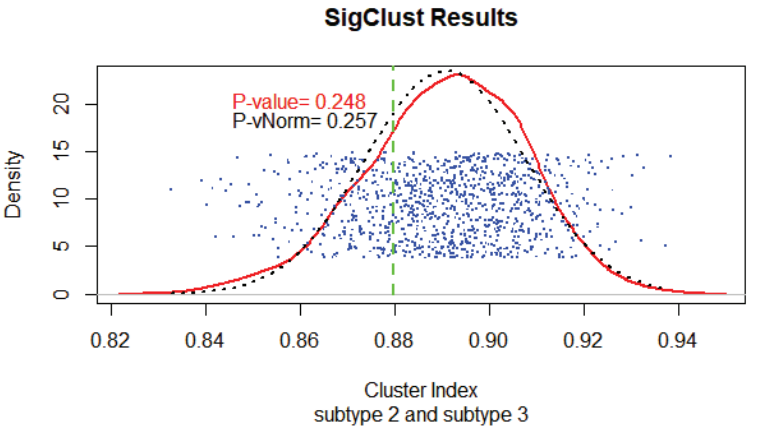


Figure 5.8: Statistical analysis.

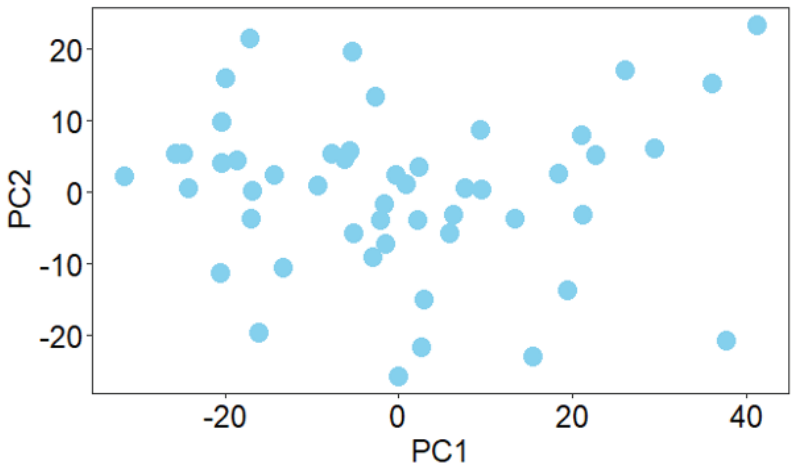


Figure 5.9: All data without clustering.

- The purpose of differential expression analysis is to compare the expression of each subtype to that of a control group (always a set of normal samples). Table 5.10 compares the 3 subtypes.

	Subtype 1	Subtype 2	Subtype 3
Subtype 1	1.000	0.013	0.000
Subtype 2	0.013	1.000	0.248
Subtype 3	0.000	0.248	1.000

Table 5.10: Subtypes comparision.

Monte Carlo Clustering

It is Over Consensus Clustering for Genomics Data For Tumor Identification. It is a generally applied technique to recognize the number of groups through the guideline of soundness determination (Antolin et al., 2020).

The commonly used Monti consensus clustering algorithm uses the stability selection principle to determine the number of clusters (K). This approach functions by resampling data for each K and clustering it using an NXN consensus matrix, where each member denotes the percentage of times two samples were combined. The graph of **Figure 5.9** shows the null data set in which clusters are not shown. M3C is a consensus clustering algorithm that improves performance by eliminating overestimation of K and can test the null hypothesis K = 1. It utilizes the Relative Cluster Stability Index (RCSI).

Monte Consensus Clustering is applied and Cumulative Distribution Function (CDF) is plotted. For K = 2 to 10 as shown in **Figure 5.10**. As the values of K increase a more stable matrix is found.

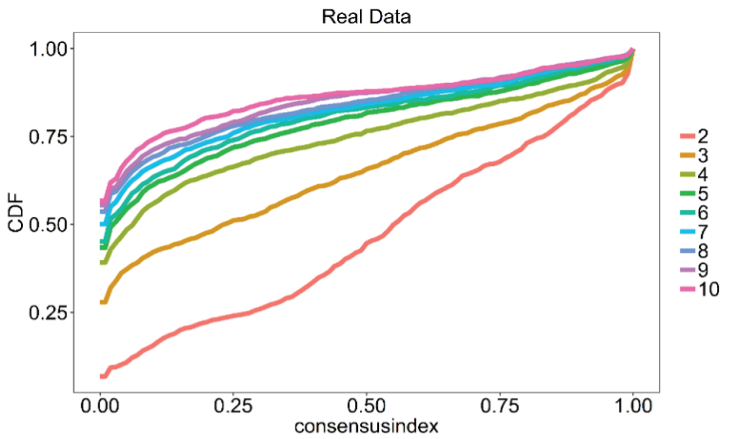


Figure 5.10: Consensus index.

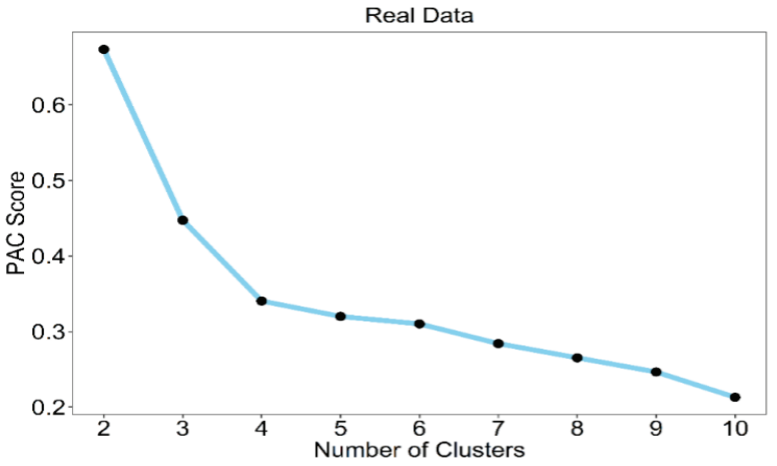


Figure 5.11: PAC score with number of clusters.

The next plot of **Figure 5.11** shows the PAC Score which measures the CDF plot flatness. With an increase of K, a more steady function is obtained.

Relative Cluster Stability Index (RCSI) graph 5.12 is created with the reference of PAC scores to decide optimal K. Maximum value of RCSI is at k = 4. RCSI values decide K and P values to reject the null.

Once we decide cluster values K = 4, then the **Figure 5.13** shows how four clusters are arranged:

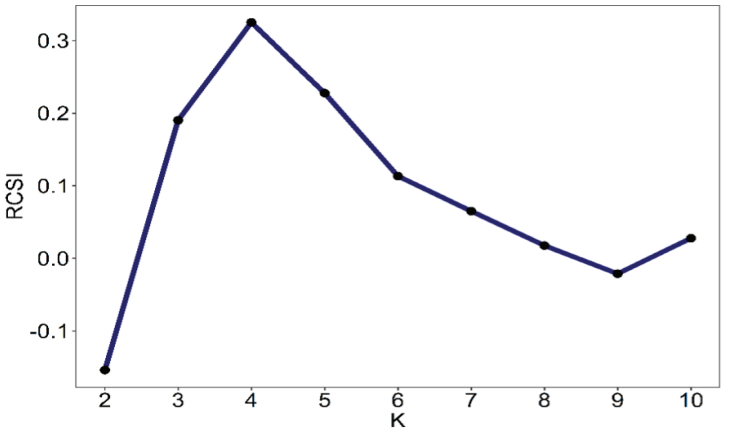


Figure 5.12: Relative cluster stability index (RCSI).

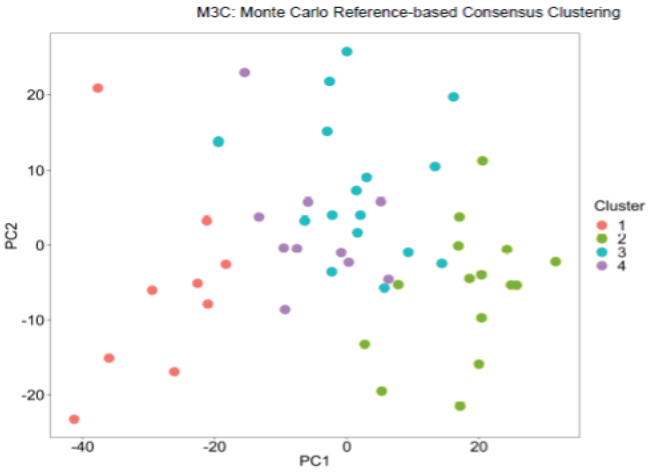


Figure 5.13: 4 Clusters are arranged.

Another plot 5.14 demonstrates the consensus matrix for our ideal clustering solution (where K = 4 in this instance). This should be quite crisp which reflects the significant stability of the results. The clusters do appear to be pretty un- ambiguous as supporting our position, as shown in the heatmap of the consensus matrix below.

Figure 5.15 shows the t-SNE of 4 cluster simulated datasets.

Figure 5.16 shows expression data and the ordered annotation data for 4 clusters.

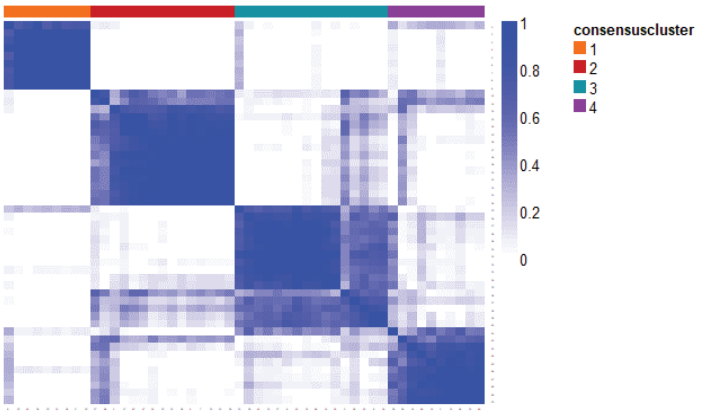


Figure 5.14: A heatmap of gbm consensus matrix.

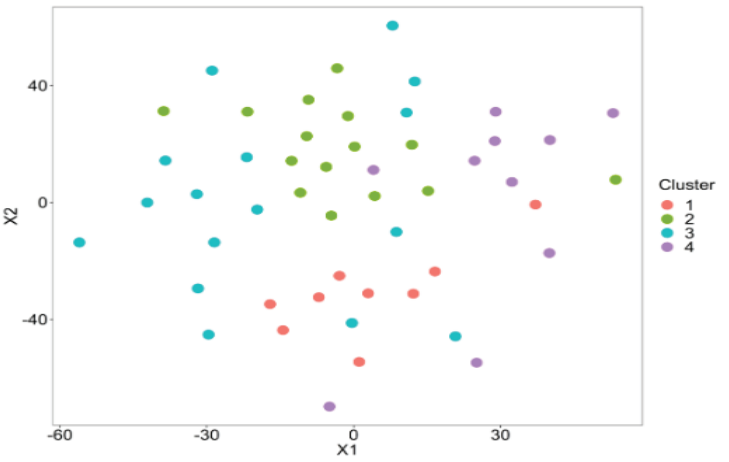


Figure 5.15: t-SNE of 4 cluster simulated dataset.

Gene Expression Clustering

The first step is to load the gene expression data into the frame. A vector of gene expression values is taken at different intervals like X = 100, 200, 300, 400 Time series vectors of **Figure 5.17** shows the number of rows and time points as number of columns.

The clustering object of time series data is created which has the basic attributes like estimated model parameters, clustering membership, posterior probabilities of the time series, mixing coefficients, or model likelihood. Further, clustering is applied with the number of clusters = 3 as shown in **Figures 5.18**, **5.19** and **5.20**.

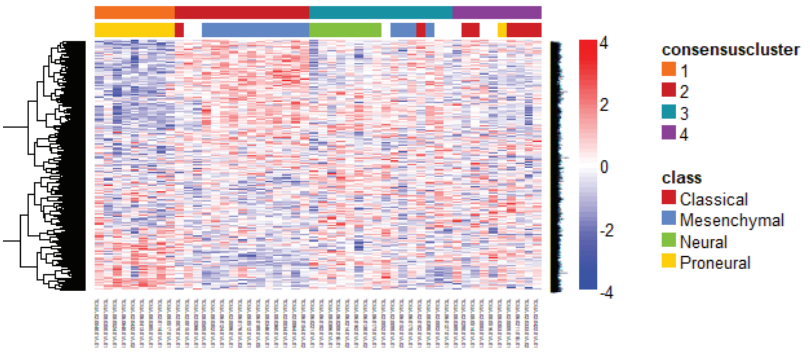


Figure 5.16: A heatmap of GBM consensus clusters with tumor classification.

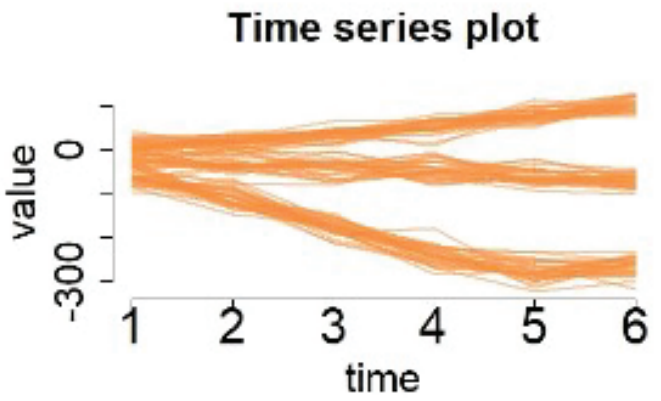


Figure 5.17: A time series plot with different intervals.

EM Algorithm Clustering is based on Expectation Maximized (EM) Algorithm:

- E-Step: Estimation of the lost variables in the dataset.
- M-Step: Maximize the parameter of the perfect in the occurrence of data.

Silhouette Analysis: The term ”silhouette” describes a technique for analyzing and verifying consistency inside data clusters. The method gives a brief graphic representation of how accurately each object has been identified.

- The range of the silhouette value, which measures how similar an item is to its cluster about other clusters, is from -1 to +1.

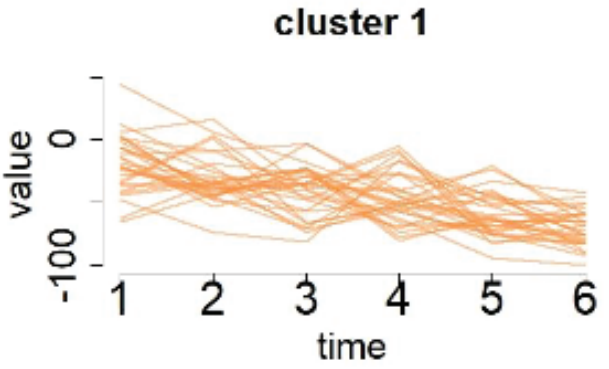


Figure 5.18: Cluster 1.

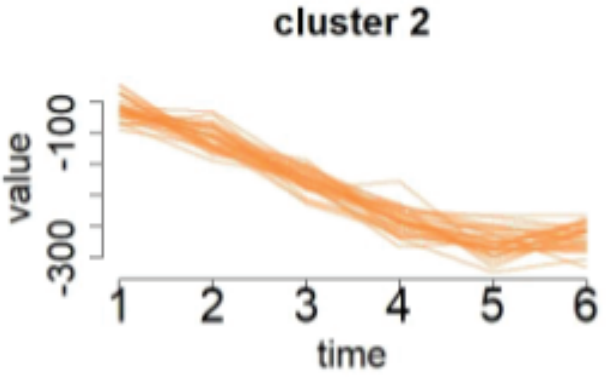


Figure 5.19: Cluster 2.

- 1 indicated poor matching and +1 indicates good matching.

Clustering consistency with silhouette Analysis. These plots 5.21 and 5.22 display the distribution of silhouette coefficients calculated with 3 and 4 clusters for each data point. The better a grouping is characterized, the wider the silhouette widths are and the more data points have silhouette widths above average.

By comparing the silhouette plots of Figures 5.21 and 5.22, the average silhouette width (black dotted line) for K=3 is better as compared to K =4. In this way, the user can use the silhouette plot to choose the best number of clusters corresponding to the data.

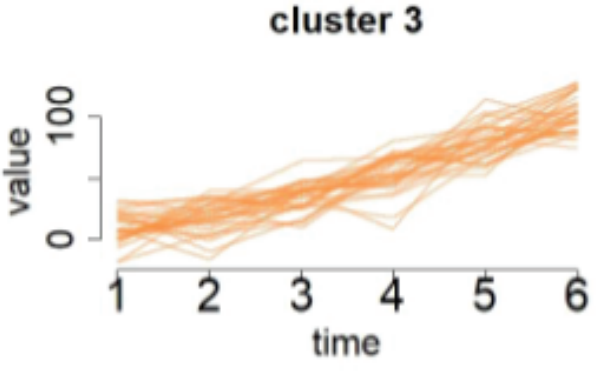


Figure 5.20: Cluster 3.

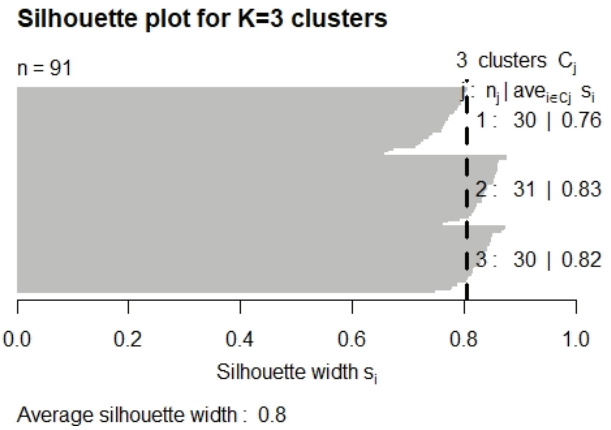


Figure 5.21: Silhouette plot for 3 clusters.

Overall contribution of work

- The whole work focuses on two types of data mining techniques: Supervised Learning and Unsupervised Learning.

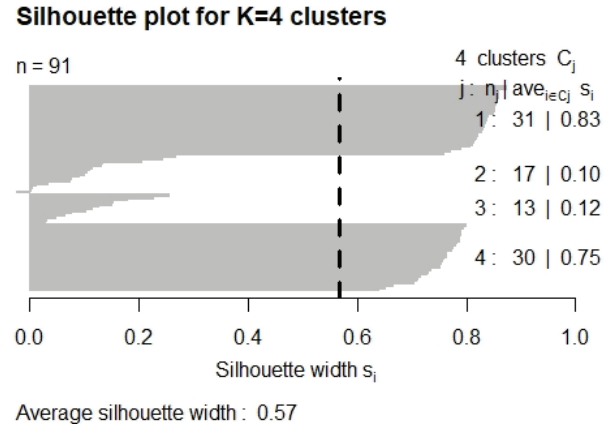


Figure 5.22: Silhouette plot for 4 clusters.

- Genes are made of DNA and different patterns of the DNA are called Gene Expression. The complete set of instructions is called Gnome. In this research work, we have tried to analyze and classify the cancer types, especially Leukemia from Genome. The input for the classification is the Gene Expression dataset which is available at (BioConductor 2018). The Expression data is in the form of a large matrix with thousands of parameters, so we have selected support Vector Classification to apply to it. (Meyer et al., 2003)

  The feature selection is based on Parametric Empirical Bays methods (Barrier et al., 2005) and Double-nested internal cross-validation. The output shows different types of leukemia.

- Based on the Gene selection procedure and classification applied to the Genome, the results show which genes are involved in the subtypes of Leukemia ALL, AML, CML, CLL, or NoL. Based on the genes category they are collected into a network.
- Another part of the work shows how to identify subtypes of cancer using unsupervised Learning Clustering. Different types of Clustering with their Analysis are done. The concept of Consensus Clustering is used and different fusion of clustering is performed.
- The resulting clusters' separation distance can be examined using a silhouette analysis. By showing a measure of how close each point in one cluster is to points in the adjoining clusters, the silhouette plot provides a visual method to assess characteristics such as the number of clusters.
- The silhouette plot will be in the opposite direction if the cancer types are not matched. Hence, samples with negative silhouette width are said to be dissimilar to one another.
- The next part of the research work is about Monte Carlo clustering, how it works, and how many optimum clusters can be selected. The concept is Silroulette analysis which is implemented and the results are shown in the graph.
- The last part of the research work is based on the study and analysis of Gene Expression clustering. The stability analysis is performed and to validate the clustering, the Silhouette analysis is performed.

Limitations

- One of the parts of the present study is focusing on Leukemia and its types to classify, whereas the other hemolytic disease is not taken into consideration which can be taken as another research topic.
- Different types of clustering are used to specify the categories and silhouette analysis is done to check how the samples are matched with the clusters. This study can not have specific results about the mortality ratios.

- Due to the increase in heterogeneous data usage, cost-effectiveness is also an important parameter to look into. This study has not shown any work related to cost.
- This study has not focused on the complexity part of algorithms.

Conclusions and Future Directions

CHAPTER SIX

Author
Dr. Tejal Upadhyay

*Corresponding author: Dr. Tejal Upadhyay, Assistant Professor, Department of Computer Science and Technology, Nirma University, S G Highway, Ahmedabad, Gujarat, India, Email: tejal.upadhyay@nirmauni.ac.in

This chapter discusses the contribution towards cancer identification using Genome and the future scopes of this research.

Conclusion

Data mining is the technique of finding knowledge from heterogeneous data. These data can be used as per the nature of the data, if it is data of a supermarket then the goal will be how to get more business from it. If it is data on what types of jobs are available in the market, then our focus will be on where we can start to fulfill the market requirements. If the data is of the human body which is a very complex structure, it can help to diagnose, treat and predict any disease.

In the thesis, the different techniques of data mining are used to improve the analysis and identification of the disease. A comprehensive review of methodologies implemented for pre-processing and reducing noise is presented. Supervised learning and unsupervised learning methodologies are implemented and result analysis is discussed.

Objectives and Outcome Mapping

Table 6.1 shows the mapping between the Objectives and Outcomes of the research work.

Sr.No	Objectives	Outcomes
1	item To enhance the link be- tween Computer Science and Medical Science to investigate the utility and effectiveness of Data Mining algorithms that will help society.	Medical data is taken as an input and Data Mining algorithms are applied to it, so the collaborative study allows Medical Science and Computer Science by referring to chapters 2, 3, 4, and 5.
2	To identify and predict the subtype of blood cancer from the Gene expression by choosing an appropriate classification technique which will help to detect early-stage of cancer.	Gene expression data of project MILE is taken as input with 60 samples and 20172 features to classify the subtypes of leukemia data sets.
3	To identify cancer subtypes, similarity index among clusters, and survival ratio of a cancer patient by applying Data Mining Algorithm.	For the similarity index, the silhouette analysis and for the death ratio, sur- vival analysis is derived.
4	To get the optimized number of clusters on large data sets, distributed across many data sets.	RCSI (Reference Clustering Stability Index) decides the optimal cluster count by applying Monte Carlo Simulation-based clustering which works on distributed data sets.
5	To get the optimized result of Sil- ilhouette analysis.	Performed the stability analysis and validate the clusters with silhouette analysis. Methods based on EM (Expectation Maximization) is used.

Table 6.1: Objectives and outcome mapping.

The Features and Discussion of Research Carried out in the Thesis included:

- Introduction and Overview of Genome, Gene Expression, and Cancer Types.
- How data mining techniques of preprocessing, classifications, clustering, etc can be applied to Gene Expression data?
- Expansion and Operation of Classifications and Clustering Methods for the Identification and Prediction of Cancer with the Help of R Programming.
- The result analysis and conclusion of identifying cancer from gene expression using classification and clustering techniques.
- Future scope to enhance the research area for diseases like Covid-19.

Future directions

- A genome is a life cycle of the complete organization of DNA-the particle that contains the genetic rules probable to make and manage the exercises of each creature. Different types of genomes and their characteristics studies of different animals is the new era of research.
- Each DNA molecule consists of two paired, twisted strands. Nucleotide bases are the four compound building blocks that make up each element. Adenine (A), thymine (T), guanine (G), and cytosine (C) are the bases. Direct bases on inverse strands couple; an A consistently binds with a T and a C binds with a G. The new future focus is to thoroughly examine A, T, G, and C.
- These base sets, which are found in the 23 sets of chromosomes at the center of each of our cells, are covered by around 300 crores of the human genome. Each DNA molecule has hundreds to thousands of characteristics that communicate the instructions for making proteins. Every single component of the human genome produces three proteins on average. The DNA characteristics offer new scientific insights as well.
- At the moment, the study of cancer genomes is moving quickly and with great excitement. Understanding of cancer processes is at an all-time high because of developments in sequencing technology, computational methodologygies, tumor models, and new approaches can be used to further this understanding.
- Understanding the importance of genomics data increasingly depends on multidisciplinary input and collaboration. The depth and breadth of sequencing technology have improved, and this has been crucial for better understanding cancer. Sequencing technology has the potential for further study.
- Given the enormous heterogeneity of cancers and the difficulty of effectively treating them, the ability to sequence cancer genomes has helped to quickly identify driver mutations and has assisted in figuring out the complex relationships between various cancer subclones over time and space. Another recent research area is various cancer subclones.

- It has become possible to detect previously unrecognized mutational pathways as sequencing tools have developed to the point where small amounts of cancer or individual cells may be sequenced. Finally, we can assert that people who are technologists or doctors who work in the field of medicine have a wealth of employment prospects.
- The study part can also include the complexity of any methods, which is used to decide the best methods to be used.
- The recent covid 19 is also a new era to work with, due to its severity of it and a lot of mutations allow working with.

References

CHAPTER SEVEN

1. Aibar, S., Fontanillo, C. & De Las Rivas, J. (2013), 'Genetclassifier', https://lirias.kuleuven.be/1925432?limo=0.
2. Antolin, A, A., Ameratunga, Malaka, Banerji, Udai, Clarke, A, P., Workman, Paul, Al-Lazikani & Bissan (2020), 'The kinase polypharmacology landscape of clinical parp inhibitors', Scientific reports 10(1), 1–14.
3. Baba, A. & Câtoi, C. (2007), 'Comparative oncology. Bucharest: The publishing house of the Romanian academy; 2007. Chapter 3, tumor cell morphology.
4. Barrier, Alain, Lemoine, Antoinette, Boelle, Pierre-Yves, Tse, Chantal, Brault, Didier, Chiappini, Franck, Breittschneider, Julia, Lacaine, Francois, Houry, Sidney, Huguier & Michel (2005), 'Colon cancer prognosis prediction by gene expression profiling', Oncogene 24(40), 6155.
5. Benko, Andrea, Wilson & Ben (2003), 'Online decision support gives plans an edge', Managed Healthcare 13(5), 20–20.
6. Bernhard, S., Smola, J, A., Bach & Francis (2002), Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT press.
7. Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G. & Milanesi, L. (2016), 'Methods for the integration of multi-omics data: mathematical aspects', BMC bioinformatics 17(S2), S15.
8. Biafore, S. (1999), 'Predictive solutions bring more power to decision makers.', Health Management Technology 20(10), 12–14.
9. Bioconductor (2018), 'An opensource software for bioinformatics', https://www.bioconductor.org.
10. Chen & Chien-Hsing (2014), 'A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection', Applied Soft Computing 20, 4–14.

11. Christy, T. (1997), 'Analytical tools help health firms fight fraud', Insurance & Technology 22(3), 22–26.

12. Chung, H. M. & Gray, P. (1999), 'Data mining', Journal of management information systems 16(1), 11–16.

13. Das, J., Gayvert, K. M. & Yu, H. (2014), 'Predicting cancer prognosis using functional genomics data sets', Cancer informatics 13, CIN–S14064.

14. Gene Pattern (2017), http://www.broadinstitute.org/cancer/software/genepattern/.

15. Genomic (2016), 'Basics of genomic', https://www.yourgenome.org/facts/ what-is-a-genome.

16. Gillespie, G. (2000), 'There's gold in them that database.', Health data management 8(11), 40–4.

17. Hayes, Neil, D., Monti, Stefano, Parmigiani, Giovanni, Gilks, Blake, C., Naoki, Kat- suhiko, Bhattacharjee, Arindam, Socinski, A, M., Perou, Charles, Meyerson & Matthew (2006), 'Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts', Journal of Clinical Oncology 24(31), 5079–5090.

18. Hoffman, Patrick, Grinstein, Georges, Marx, Kenneth, Grosse, Ivo, Stanley & Eugene (1997), DNA visual and analytic data mining, in 'Proceedings. Visualization'97 (Cat. No. 97CB36155)', IEEE, pp. 437–441.

19. Kauffman, T. L., Irving, S. A., Leo, M. C., Gilmore, M. J., Himes, P., McMullen, C. K., Morris, E., Schneider, J., Wilfond, B. S. & Goddard, K. A. (2017), 'The nextgen study: patient motivation for participation in genome sequencing for carrier status', Molecular Genetics & Genomic Medicine 5(5), 508–515.

20. Kendziorski, C., Newton, M., Lan, H. & Gould, M. (2003), 'On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles, Statistics in medicine 22(24), 3899–3914.

21. Kincade, K. (1998), 'Data mining: digging for healthcare gold', Insurance & Technology 23(2), 2–7.

22. Koh, H. C., Tan, G., et al. (2011), 'Data mining applications in healthcare', Journal of healthcare information management 19(2), 65.

23. Learning resource Platform (2019), https://www.himsslearn.org/about.

24. Magidson, J. (2013), 'Computer-implemented models predicting outcome variables and characterizing more fundamental underlying conditions. US Patent App. 13/520,407.

25. Medicine & Neuroscience (2013), 'Medicine and neuroscience', https://ijcsmc.com/ docs/papers/November2013/V2I11201348.pdf.

26. Meng, Tao, Soliman, T, A., Shyu, Mei-Ling, Yang, Yimin, Chen, Shu-Ching, Iyengar, SS, Yordy, S, J., Iyengar & Puneeth (2013), 'Wavelet analysis in current cancer genome research: a survey', IEEE/ACM transactions on computational biology and bioinformatics 10(6), 1442–14359.

27. Meyer, D., Leisch, F. & Hornik, K. (2003), 'The support vector machine under test, Neurocomputing 55(1-2), 169–186.

28. Meyer, P. E., Lafitte, F. & Bontempi, G. (2008), 'minet: Ar/Bioconductor package for inferring large transcriptional networks using mutual information', BMC bioinformatics 9(1), 461.

29. Milley, A. (2000), 'Healthcare and data mining.', Health Management Technology 21(8), 44–45.

30. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. (2003), 'Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data', Machine learning 52(1-2), 91–118.

31. Morris, C. N. (1983), 'Parametric empirical Bayes inference: theory and applications', Journal of the American Statistical Association 78(381), 47–55.

32. National Cancer Institute, U S Department of Health and Human Services (2018), https://www.cancer.gov/research/areas/genomic-sprogress.

33. Pati, S. K., Mallick, S., Chakraborty, A. & Das, A. (2019), Informative gene selection using clustering and gene ontology, in 'Emerging Technologies in Data Mining and Information Security', Springer, pp. 417–427.

34. Ritchie, D, M., Holzinger, R, E., Li, Ruowang, Pendergrass, A, S., Kim & Doky- oon (2015), 'Methods of integrating data to uncoverg enotype–phenotype interactions, Nature Reviews Genetics 16(2), 85–97.

35. Rousseeuw, P. J. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', Journal of Computational and applied mathematics 20, 53–65.

36. Shyamsundar, Radha, Kim, H, Y., Higgins, P, J., Montgomery, Kelli, Jorden, Michelle, Sethuraman, Anand, van de Rijn, Matt, Botstein, David, Brown, O, P.,

37. Pollack & R, J. (2005), 'A DNA microarray survey of gene expression in normal human tissues', Genome Biology 6(3), 1–9.

38. Silver, M., Sakata, T., Su, H.-C., Herman, C., Dolins, S. B., O Shea, M. J. et al. (2001), 'Case study: how to apply data mining techniques in a healthcare data warehouse', Journal of healthcare information management 15(2), 155–164.

39. Soh, K. P., Szczurek, E., Sakoparnig, T. & Beerenwinkel, N. (2017), 'Predicting cancer type from tumor DNA signatures', Genome medicine 9(1), 1–11.

40. Strehl, A. & Ghosh, J. (2002), 'Cluster ensembles—a knowledge reuse framework for combining multiple partitions', Journal of machine learning research 3(Dec), 583–617.

41. Su, C.-L., Deng, T.-R., Shang, Z. & Xiao, Y. (2015), 'Jarid2 inhibits leukemia cell proliferation by regulating ccnd1 expression', International Journal of Hematology 102(1), 76–85.

42. The National Center for Biotechnology Information (2020), http://www.ncbi.nlm.nih. gov.

43. Upton, G. & Cook, E. (2014), A Directory of Statistics, Oxford University Press.

44. Vardiman, J. W., Thiele, J., Arber, D. A., Brunning, R. D., Borowitz, M. J., Porwit, A., Harris, N. L., Le Beau, M. M., Hellström-Lindberg, E., Tefferi, A. et al. (2009), 'The 2008 revision of the world health organization (who) classification of myeloid neoplasms and acute leukemia: rationale and important changes', Blood 114(5), 937–951.

45. Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P. et al. (2010), 'Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1', Cancer cell 17(1), 98–110.

46. Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B. & Goldenberg, A. (2014), 'Similarity network fusion for aggregating data types on a genomic scale', Nature methods 11(3), 333.

47. Wang, Haixin, Glover & E, J. (2011), Noise analysis of time series data in gene regulatory networks, in '2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)', Vol. 4, IEEE, pp. 1848–1852.

48. WHO (2019), Global status report on Alcohol and Health 2018, World Health Organization.

49. Wilkerson, M. D. & Hayes, D. N. (2010), 'Consensusclusterplus: a class discovery tool with confidence assessments and item tracking', Bioinformatics 26(12), 1572–1573.

50. Winkler, U., Jensen, M., Manzke, O., Schulz, H., Diehl, V. & Engert, A. (1999), 'Cytokine-release syndrome in patients with b-cell chronic lymphocytic leukemia and high lymphocyte counts after treatment with an anti-cd20 monoclonal antibody (rituximab, idec-c2b8)', Blood, The Journal of the American Society of Hematology 94(7), 2217–2224.

List of Abbreviation

AI: Artificial Intelligence. 18

ALL: Acute Lymphoblastic Leukemia. 9, 29, 33–35, 63, 64, 83

AML: Acute Myeloid Leukemia. 9, 29, 33, 34, 36, 63, 64, 83

CC: Consensus Clustering. 48, 51, 52, 72

CCG: Centre for Cancer Genome. 22, 23

CGCI: Cancer Genome Characterization Initiative. 23

CLL: Chronic Lymphoblastic Leukemia. 9, 29, 33, 34, 36, 63, 64, 83

CML: Chronic Myeloid Leukemia. 9, 29, 33, 34, 37, 63, 64, 83

CNNF: Consensus Non-negative matrix factorization. 51, 52 CV Cross Validation. 31

DNA: DeoxyriboNucleic Acid. 2, 13, 14, 17, 18, 26

EM: Expectation-Maximization. 4, 25, 61, 80

HC: Hierarchical Clustering. 56

IC: Interactive Clustering. 51, 52

M3C: Monte Carlo Consensus-based Clustering. 3, 54, 56, 76

MAD: Median Absolute Deviation. 47

NCI: National Cancer Institute. 14, 22, 23

NoL: No Leukemia. 9, 33, 34, 37, 63, 64, 83

PCA: Principle Component Analysis. 56

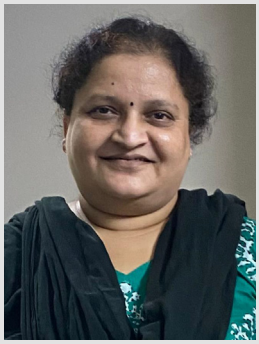
PEB: Parametric Empirical Bayes. 31

RCSI: Relative Cluster Stability Index. 24, 54, 76

RNA: Ribonucleic acid. 26

SNF: Similarity network fusion. 48, 52, 72

About the Authors



Dr. Tejal Upadhyay

My name is Tejal Upadhyay, and I have been dedicated to Nirma University for the past 23 years. I hold a BTech in Computer Engineering from L D College of Engineering, Ahmedabad, which I completed in 1996. In 2004, I pursued an M.Tech from Dharamsinh Desai University, Nadiad. Recently, in 2022, I completed my Ph.D. at Charusat University.

Throughout my extensive tenure, I have actively engaged with Nirma University, primarily focusing on teaching undergraduate and postgraduate students in the field of Computer Engineering. Furthermore, since 2005, I have held the position of student branch councilor for the Nirma University branch of the Computer Society of India (CSI), a distinguished National Organization.

ISBN 978-969-33-9255-5