

PROSTATE ADENOCARCINOMA BIOMARKERS IDENTIFICATION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the
award of Bachelor of Sciences Degree in Bioinformatics and
Data Science

By

GUDA DEEPAK REDDY(40738005)



**DEPARTMENT OF BIOINFORMATICS
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600119**

MAY 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **GUDA DEEPAK REDDY (Reg.No: 40738005)** who carried out the project entitled "**PROSTATE ADENOCARCINOMA BIOMARKERS IDENTIFICATION USING MACHINE LEARNING**" under my supervision from January 2023 to May 2023.

Internal Guide

Dr. Jemmy Christy H, M.Tech., Ph.D.

Head of the Department

Dr. Jemmy Christy H, M.Tech., Ph.D.

Submitted for Viva voce Examination held on 05/05/23

Internal Examiner

External Examiner

DECLARATION

I, **GUDA DEEPAK REDDY** hereby declare that the project report entitled **“PROSTATE ADENOCARCINOMA BIOMARKERS IDENTIFICATION USING MACHINE LEARNING”** done by me under the guidance of **Dr.JEMMY CHRISTY H, M.Tech., Ph.D.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Science Degree in Bioinformatics and Data science.

DATE: 05/05/23

PLACE: CHENNAI

Guda deepak Reddy

SIGNATURE OF THE CANDIDATE

(GUDA DEEPAK REDDY)

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. RAMESH KUMAR V**, Dean, School of **BIO AND CHEMICAL ENGINEERING** and **Dr. JEMMY CHRISTY H, M.Tech., Ph.D.**, **Head of the Department, Dept. of BIOINFORMATICS** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. JEMMY CHRISTY H, M.Tech., Ph.D.** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Bioinformatics** who were helpful in many ways for the completion of this project.

ABSTRACT

Prostate cancer is a leading cause of cancer-related deaths among men, and biomarkers such as prostate-specific antigen are commonly used for its diagnosis and management. This study aimed to identify novel biomarkers for prostate adenocarcinoma through gene set enrichment analysis of differentially expressed genes. Our results showed that KLK2, RPS27, RPS12, MTND3, and MTCO1 were significantly associated with prostate adenocarcinoma and could potentially aid in its diagnosis, prognosis, and treatment. These biomarkers were involved in pathways related to transcriptional misregulation in cancer, viral carcinogenesis, cAMP signaling and immunoglobulin receptor binding. However, it is important to note that biomarkers are just one aspect of prostate cancer diagnosis and management, and further medical evaluation is necessary for accurate assessment and treatment decisions. Our findings provide a valuable contribution to the ongoing efforts to improve the diagnosis and treatment of prostate cancer. Gene Set Enrichment Analysis, also known as functional enrichment analysis or pathway enrichment analysis, is a statistical technique used to identify groups of genes or proteins that are over represented or underrepresented in a large set of genes or proteins and may be relevant to disease phenotypes. Transcriptomics technologies and proteomics data are commonly used to analyze thousands of genes in geneset enrichment analysis. By identifying gene sets that are enriched or depleted in a particular phenotype, Gene set enrichment Analysis can provide insights into the molecular mechanisms underlying diseases and can be useful for identifying potential therapeutic targets. The identification of highly enriched or depleted groupings of genes is determined using statistical methods. It is a valuable tool for analyzing complex biological data and can help researchers uncover novel disease mechanisms and identify potential therapeutic targets.

TABLE OF CONTENTS

CHAPTER NO.	NO.	TITLE	PAGE NO.
		ABSTRACT	i
		LIST OF FIGURES	iv
		LIST OF TABLES	v
		LIST OF ABBREVIATIONS	vi
1		INTRODUCTION	1
2		LITERATURE SURVEY	4
3		AIM AND SCOPE	6
	3.1	Aim and scope	6
	3.2	Objective	7
4		MATERIALS AND METHODS	9
	4.1	Materials	9
	4.1.1	R Studio	9
	4.1.2	CRAN	10
	4.1.3	Bioconductor	11
	4.1.4	Miniconda	12
	4.1.5	Data from TCGA	13

4.2	Methods	14
4.2.1	Downloading and processing of data	16
4.2.2	Cleaning and structuring of data	17
4.2.3	Normalization of data	18
4.2.4	Principal component analysis	19
4.2.5	Construction of neural networks	20
4.2.6	Extraction of genes	21
4.2.7	Conversion of gene id	22
4.2.8	Gene set enrichment analysis	23
5	RESULTS AND DISCUSSIONS	24
6	SUMMARY AND CONCLUSIONS	45
7	REFERENCES	47

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
5.1	Histogram of train data(1)	24
5.2	Histogram of train data(2)	25
5.3	Histogram of train data(3)	26
5.4	Principal component analysis	27
5.5	One-Hot Encoding of the data	27
5.6	Histogram of Data Extracted based on Weights	29
5.7	Histogram of Data Extracted based on Bias.	30
5.8	Construction of Neural Networks	31
5.9	Gene Set Enrichment from Cluster profiler - Biological Process	36
5.10	Gene Set Enrichment from Cluster profiler – Molecular processes	39
5.11	Gene Set Enrichment from Cluster profiler – Cellular categories	42

LIST OF TABLES

TABLE NO.	TITLE	PAGE NO.
5.1	Gene of interest list	32
5.2	Converted Ensembl ID into Entrez ID	34
5.3	Results of GSEA - Biological Process	37
5.4	Results of GSEA - Molecular processes	40
5.5	Results of GSEA - Cellular categories	43

LIST OF ABBREVIATIONS

GDC	Genomic Data Commons
TCGA	The Cancer Genome Atlas
PRAD	Prostate Adenocarcinoma
DEG's	Differentially Expressed Genes
GOI	Gene of Interests
CRAN	Comprehensive R Archive Network
GSEA	Gene Set Enrichment Analysis
SVM	Support Vector Machine
ANN	Artificial Neural Network
PCA	Principal Component Analysis
GO	Gene ontology
MSigDB	Molecular Signatures Database
cAMP	Cyclic Adenosine Monophosphate
IDE	Integrated Development Environment

CHAPTER - 1

INTRODUCTION

CANCER:

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body. Since the human body contains billions of cells, cancer may develop practically everywhere. Human cells typically proliferate and divide (via a process called cell division) to create new cells when the body requires them. Cells die as they are old or injured, and new cells replace them. When this routine process fails, aberrant or damaged cells begin to proliferate and develop when they shouldn't. Tumors tissue growth can develop from these cells. Tumors can be malignant or benign (not cancerous). Cancerous tumors spread into, or invade, nearby tissues and can travel to distant places in the body to form new tumors (a process called metastasis). Cancerous tumors may also be called malignant tumors. Many cancers form solid tumors, but cancers of the blood, such as leukemias, generally do not. (Ding, H et al.,2021)

PROSTATE ADENOCARCINOMA:

The most frequent cancer among men and the second-leading cause of cancer-related fatalities in the country is prostate cancer. The most frequent form of prostate cancer, known as prostate adenocarcinoma (PRAD), accounts for over 95% of all occurrences. Large volumes of gene expression profiling data have been created recently as a result of the advent of high-throughput sequencing technology, offering an unheard-of chance to find biomarkers and possible treatment targets for PRAD. A crucial resource for cancer research is the Cancer Genome Atlas project, which has gathered and examined genomic, transcriptomic, and clinical data from thousands of cancer patients, including PRAD patients. We will use TCGA gene expression data in this study to find possible biomarkers and pathways connected to PRAD.

(MS Litwin et al., 2017)

A variety of genetic and epigenetic changes control the genesis and course of the complex and diverse illness known as prostate cancer. Therefore, it is crucial to pinpoint the crucial molecular processes that underlie the onset and development of PRAD in order to create more potent treatments for the condition. In order to find possible biomarkers and treatment targets for PRAD, high-throughput molecular profiling methods like microarrays and next-generation sequencing have been extensively explored. Large-scale genomic and transcriptome data sets, such those produced by the TCGA project, have recently been accessible, allowing researchers to carry out thorough investigations of the PRAD's molecular properties.(Qiangqiang Zheng et al.,2021)

Over 500 PRAD patients' genomic, transcriptomic, and clinical data have been gathered and examined as part of the TCGA project, offering a wealth of information for researching the underlying molecular causes of PRAD. The differentially expressed genes (DEGs) and signalling pathways that are dysregulated in PRAD may be found using the TCGA gene expression data, which can also be utilized to find possible biomarkers for diagnosis, prognosis, and therapy response prediction.

The TCGA PRAD data has been utilized in a number of research to find possible biomarkers and treatment targets for this illness. For instance,(Zhang et al.,2021). identified a group of immune-related genes that were connected to PRAD prognosis using TCGA PRAD data. Using TCGA data,(Zhu et al) discovered a collection of DEGs and pathways related to PRAD development.

Numerous cancer types have produced a significant quantity of genetic and clinical data according to the Cancer Genome Atlas (TCGA) initiative. This information offers a rare chance to identify biomarkers and treat cancer patients specifically. However, it is difficult to find clinically useful biomarkers that might foretell patient outcomes and direct treatment choices because of the complexity and variety of cancer.(Liñares-Blanco et al., 2021)

We will categorize patients into several subgroups based on their genetic and clinical data using supervised machine learning methods including logistic regression, support vector machines, and random forest. We'll also employ unsupervised Machine learning methods like clustering and dimensionality reduction to find patient groups with comparable clinical and genomic characteristics.

Biomarkers that can be utilized for cancer diagnosis, prognosis, and therapy have been identified using large-scale genomic and clinical data analysis and machine learning algorithms. The goal of this thesis is to use machine learning approaches on TCGA data to find new biomarkers for various cancer types.(A Dhillon.,2023)

To find the most useful characteristics for forecasting patient outcomes, we will also use feature selection and feature engineering approaches. We will also conduct cross-validation and independent validation on outside datasets to look at the robustness and generalizability of our models.

CHAPTER- 2

LITERATURE SURVEY

Prostate cancer is a complex disease, and the identification of reliable biomarkers is crucial for accurate diagnosis, prognosis, and treatment. Several studies have been conducted in the past few years to identify potential biomarkers for prostate cancer using a variety of techniques, including genomics, proteomics, and machine learning.

One study by Bishoff et al. (2020) utilized genomic and proteomic approaches to identify and validate potential biomarkers for prostate cancer diagnosis and prognosis. Fang et al. (2017) used quantitative proteomics to identify potential biomarkers for prostate cancer, while Auprich et al. (2018) conducted a critical evaluation of the current state of prostate cancer biomarker research. Yang et al. (2016) used a differential proteome approach to identify potential biomarkers for prostate cancer metastasis. Machine learning is also being increasingly utilized to identify potential biomarkers for prostate cancer. One study by Zheng et al. (2018) utilized machine learning algorithms to identify potential biomarkers for prostate cancer using gene expression data. Another study by Zhang et al. (2019) utilized a machine learning approach to predict the prognosis of prostate cancer patients using gene expression data.

Overall, the review of the literature on prostate cancer biomarkers reveals that great progress has been achieved in locating possible markers for diagnosis and prognosis utilizing a number of techniques, including proteomics, genomics, and machine learning. To guarantee the clinical relevance of these indicators in the therapy of prostate cancer, more validation and testing are needed. Integrative analytic techniques and the utilization of a multigenic approach may potentially present intriguing directions for future study. In the end, accurate biomarker discovery has the potential to greatly advance prostate cancer diagnosis, prognosis, and therapy. (Butler, L. M. 2018)

On the other hand, this work of identifying prostate adenocarcinoma biomarkers using machine learning is a specific methodological approach that involves the use of R software to build and train machine learning models on prostate cancer data to identify potential biomarkers. This approach utilizes the algorithms and libraries available in R for data manipulation, cleaning, and analysis, as well as for building and testing machine learning models. In conclusion, the identification of reliable biomarkers for prostate cancer is a critical area of research. Genomics, proteomics, and machine learning are all powerful tools that can be utilized for this purpose. Further research is needed to validate and test potential biomarkers, and the integration of multiple approaches may be necessary to achieve accurate and reliable biomarker discovery for prostate cancer diagnosis, prognosis, and treatment.

CHAPTER-3

AIM AND SCOPE

3.1 AIM AND SCOPE :

Finding and developing diagnostic methods that can spot prostate cancer at an early stage would enable early management and better results, which is the goal of discovering prostate cancer biomarkers. Biomarkers are quantifiable signs found in biological samples like blood, urine, or tissue that can show if a disease is present or progressing. Finding biomarkers for prostate cancer monitoring, prognosis, and diagnosis are all included in the scope of this task. Additionally, biomarkers may be helpful in determining which therapies are most likely to be successful for a certain patient.

Researchers can create more precise and effective diagnostic and treatment approaches by discovering cancer biomarkers, which might eventually result in better patient outcomes. Finding biomarkers may also assist in reducing the need for risky and painful intrusive treatments like biopsies.

The aim of using machine learning to identify biomarkers in TCGA data is to improve cancer diagnosis and treatment. By analyzing large amounts of genomic data, machine learning algorithms can identify patterns and biomarkers that may be useful in predicting patient outcomes and developing targeted therapies. This can lead to more personalized and effective treatments for cancer patients. Additionally, machine learning can help identify new biomarkers that were previously unknown, which can further advance our understanding of cancer biology.

3.2 OBJECTIVE:

The objective of identifying biomarkers using machine learning in The Cancer Genome Atlas (TCGA) project is to improve the understanding and management of cancer. More specifically, the objectives of biomarker identification using machine learning include:

Improving cancer diagnosis: By identifying biomarkers that are associated with specific types of cancer, Machine learning can help develop more accurate and efficient diagnostic tools that can detect cancer at an earlier stage. Enabling personalized treatment: By identifying biomarkers that are associated with specific cancer sub types, Machine learning can help develop personalized treatment plans that are tailored to the specific needs of individual patients. Identifying new therapeutic targets: By identifying biomarkers that are associated with cancer progression or resistance to current treatments, Machine learning can help identify new therapeutic targets that can be exploited to develop novel cancer therapies. Improving cancer prognosis: By identifying biomarkers that are associated with cancer outcomes, Machine learning can help develop more accurate prognostic tools that can predict disease progression and guide treatment decisions.

Machine learning has the potential to significantly improve cancer diagnosis, treatment, and prognosis by identifying biomarkers associated with cancer. These biomarkers can help in the development of more accurate and efficient diagnostic tools that can detect cancer at an earlier stage, personalized treatment plans tailored to the specific needs of individual patients, and novel therapies that target specific cancer subtypes. Machine learning algorithms can analyze large amounts of patient data and identify patterns that are difficult or impossible for humans to detect, making it an invaluable tool for cancer research and care.

By improving cancer prognosis, machine learning can help clinicians make better-informed treatment decisions and improve patient outcomes. It can also help identify new therapeutic targets that can be exploited to develop novel cancer therapies, potentially leading to better treatment options for patients. Machine learning can play a significant role in the development of precision medicine, where treatments are tailored to individual patients based on their unique biomarker profile.

Some potential directions for future research in this area include:

Integration of multi-omics data: Currently, most studies focus on using a single type of omics data (e.g., gene expression data) for biomarker identification.

Development of hybrid Machine Learning models: The combination of multiple machine learning techniques, such as deep learning and dimensionality reduction, may lead to more accurate and efficient biomarker identification methods.

Model interpretability and explainability: The ability to interpret and explain the decisions made by machine learning models is essential for clinical applications.

CHAPTER-4

MATERIALS AND METHODS

4.1 MATERIALS:

R STUDIO (Licensed)(RStudio version 2023.03.0)

A Large Libraries of packages (from CRAN and Bioconductor)

Miniconda (Keras and tensorflow through Reticulate)

Data from TCGA database (<https://portal.gdc.cancer.gov/projects/TCGA-PRAD>)

4.1.1 R Studio

An integrated development environment (IDE) for the R programming language is called R Studio. For authoring, running, and debugging R code, it offers a user-friendly interface. As open-source software that is free to use and alter, R Studio is available for Windows, macOS, and Linux.

For data analysis, visualization, and machine learning, the R programming language is a statistical computer language. R Studio offers a number of tools and features that make working with R simpler, including a terminal for running R code, a code editor with syntax highlighting and auto-completion, and a workspace for organizing R objects and data. R Studio's capacity to interface with other software tools is one of its distinguishing characteristics.

R Studio offers a user-friendly interface. The integrated development environment (IDE) comes with a code editor, a console for running R code, and a workspace for managing R objects and data. Tools for building interactive web apps, reproducible research reports, and data visualizations are also included in R Studio.

(D Mercatelli et al., 2022)

The flexibility of R Studio to interface with other software tools and packages is one of its primary characteristics. Numerous R packages that offer extra features for data analysis, visualization, and machine learning may be utilized with R Studio. Git, a version control program that enables programmers to manage changes to their code, may also be used with R Studio.

Using R Studio, we can find the pathways and gene sets that are connected to cancer. Pathway enrichment analysis may be carried out by researchers using tools like GSEA (Gene Set Enrichment Analysis) and Reactome, which enables them to find pathways that are differently regulated in cancer tissues. This study can assist in locating prospective biomarkers that are connected to important cancer-related pathways and can result in the discovery of novel cancer therapeutic targets.

(WN Arifin et al., 2019)

4.1.2 CRAN(Comprehensive R Archive Network)

R programming language software packages are housed at the Comprehensive R Archive Network (CRAN). Users may more easily identify and install the R packages they require thanks to CRAN, which offers developers a central site to share their R packages.

R is a software environment and programming language for statistical computation and graphics. Data analysis, machine learning, and data visualization all make extensive use of it. It is very extendable and offers a wide range of statistical and graphical capabilities.

R packages are groups of tools, files, and instructions that are simple to install and load into the R environment. Packages offer further capabilities for data analysis, visualization.(García-Barriocanal, E. at el 2020)

4.1.3 Bioconductor

For the analysis and interpretation of genetic data, Bioconductor is a software project that is open-source and free to create. It offers a vast number of instruments, software programs, and processes that let scientists conduct high-throughput analyses of biological data, including data from microarrays and next-generation sequencing. The project uses the R programming language and offers a wide range of functions for preparing data, ensuring data quality, normalizing data, doing statistical analysis, visualizing data, and integrating data from external databases.

Bioinformaticians and biologists now depend on Bioconductor as a vital tool since it makes it simple and repeatable for them to analyze and understand massive amounts of genetic data. It features a vibrant user and developer community that contributes to the creation of fresh tools and resources and offers assistance via forums, mailing lists, and seminars. Microarrays, next-generation sequencing, and proteomics are just a few of the high-throughput genomic data analysis programs offered by Bioconductor. Tools for data preparation, normalization, quality assurance, differential expression analysis, gene set enrichment analysis, and pathway analysis are included in these packages. Bioconductor provides access to powerful statistical and graphical methods for the analysis of genomic data. It also facilitates the integration of biological metadata like GenBank and PubMed in the analysis of experimental data.(Zheng,X et al., 2015)

Bioconductor provides a vast collection of packages that are optimized for the analysis of high-throughput genomics data. These packages cover a wide range of biological areas, such as microarray data analysis, next-generation sequencing (NGS), ChIP-seq, RNA-seq, and many others. It includes tools for data normalization, quality control, statistical analysis, visualization, and machine learning.

4.1.4 Miniconda

A package management system that simplifies the installation and maintenance of software packages for data analysis, scientific computing, and machine learning, Miniconda is a slimmed-down variant of the well-known Anaconda Python distribution. The Conda package manager, which enables users to quickly install and manage additional packages and dependencies, is included in Miniconda along with the basic set of items necessary to run Python. This makes it a viable choice for users who just want a portion of the items found in the complete Anaconda distribution and want a lightweight installation of Python.

The Conda package manager enables users to build virtual environments that may be modified with various packages and dependencies, making it simple to handle various projects and guarantee the repeatability of analyses. Many data analysis projects involve working with different software tools and libraries that are written in different programming languages, and managing the installation and compatibility of these tools can be challenging. Miniconda provides a solution to this problem by allowing users to create isolated virtual environments for different projects, each with its own set of packages and dependencies. By using Miniconda in R Studio, users can easily install and manage Python packages and dependencies required for their projects alongside R packages. This can be particularly useful for projects that involve working with large datasets, machine learning models, or other computational tools that are available in Python but not in R. (Younes, F et al., 2020)

Miniconda is a lightweight version of the popular Anaconda Python distribution that is designed for users who need more control over their Python environment. It is a free and open-source package management system and environment management system that allows users to easily install and manage software packages in a virtual environment.

4.1.5 Data from the cancer genome atlas(TCGA)

The Genomic Data Commons (GDC) Data Portal for the Prostate Adenocarcinoma (PRAD) project of The Cancer Genome Atlas (TCGA) is taken from the following link, <https://portal.gdc.cancer.gov/projects/TCGA-PRAD>.

The TCGA-PRAD project is a comprehensive effort to understand the molecular characteristics of prostate cancer, which is the second most common cancer among men worldwide. The project involves collecting and analyzing genomic, transcriptomic, and clinical data from a large number of prostate cancer patients to gain insights into the genetic and molecular drivers of the disease. The GDC Data Portal provides access to the raw and processed data generated by the TCGA-PRAD project, including sequencing data, clinical data, and analysis results. Researchers and clinicians can use this data to investigate the underlying mechanisms of prostate cancer, develop new diagnostic and therapeutic strategies, and improve patient outcomes. The TCGA-PRAD project was launched in 2008 as part of the broader TCGA initiative, which aimed to comprehensively characterize the genomic alterations in different types of cancer. The TCGA-PRAD project initially involved the analysis of tumor and normal tissue samples from over 400 patients with prostate cancer using a variety of genomic and transcriptomic technologies.

(DA Bader, SE McGuire et al., 2020)

Over the years, the project has expanded to include additional samples, data types, and analyses. For example, the project has now analyzed samples from over 550 patients with prostate cancer, including both primary and metastatic tumors. In addition to genomic and transcriptomic data, the project has also generated proteomic, epigenetic, and clinical data.

4.2 METHODS:

Workflow of the project is done in three sections as followed:

Section-1:

1. Setting up R studio and its repositories
2. Download data from TCGA
3. Install packages and import libraries.
4. Using TCGA Biolinks
5. Cleaning data and structuring of data
6. Structuring input data and filtering
7. PlotMDS(Multi-dimensionality scaling plots) from limma and edgeR
8. Normalization of data
9. Principle component (PCA) Analysis
10. Making Train Label and One -hot Encoding

Section-2:

1. Construction of neural network
2. Training of neural networks
3. Extraction of model weights
4. Extraction of genes

Section-3:

1. Conversion of Gene ID
2. Gene Set Enrichment of Gene of interest

The discovery of possible biomarkers has been the focus of several algorithms that have been used to analyze TCGA data. Several of the frequently employed strategies for biomarker discovery in TCGA data are listed below:

Differential gene expression analysis is a popular method for identifying genes that are expressed differently in tumor-related samples compared to normal ones. Differential gene expression analysis may be carried out using programs like DESeq2, edgeR, and limma.

Biomarker identification may be done using machine learning techniques like random forest, support vector machines (SVM), and artificial neural networks (ANN). The tumour subtype, prognosis, or treatment response may all be predicted using these algorithms after training on TCGA data.

Gene set enrichment analysis (GSEA) is a method for determining which gene sets are more highly expressed in tumour samples than in normal ones. This method can assist in identifying prospective treatment targets as well as pathways that are dysregulated in cancer.

Clustering analysis: Based on their gene expression patterns, tumors can be grouped using clustering methods like k-means and hierarchical clustering. This method may be used to determine the many cancer subtypes and possible biomarkers connected to each subtype.

4.2.1 DOWNLOADING AND PROCESSING OF DATA :

Accessing the TCGA Data Portal: The first step is to access the TCGA Data Portal and download the necessary data. The TCGA Data Portal provides access to a variety of genomic, clinical, and imaging data from different cancer types. TCGA data is accessed by using the link: <https://portal.gdc.cancer.gov/>

Pre-processing the biomarker data: Before identifying biomarkers, you need to pre-process the biomarker data. This involves removing any missing values, normalization, and standardization of the data. Identifying biomarkers: After pre-processing the biomarker data, you can identify biomarkers using various statistical and machine learning techniques. For example, you can use differential gene expression analysis to identify genes that are differentially expressed between cancer and normal samples. You can also use machine learning algorithms such as random forest, support vector machines, or neural networks to identify biomarkers. Interpreting the results: Finally, you need to interpret the results of the biomarker identification and validation process. This can involve identifying the biological pathways and processes that are associated with the identified biomarkers, and assessing their clinical relevance.

Downloading and processing data is a critical step in many data-driven projects. It involves identifying the data source, downloading the data, extracting it, cleaning it, and pre-processing it before it can be used for machine learning or data analysis. Once the data is loaded into a database or data structure, it can be analyzed, visualized, and modeled to gain insights and build predictive models. The process of downloading and processing data can involve different tools and techniques, depending on the data source and format. Commonly used tools include wget, curl, and Python libraries like requests and Pandas.

4.2.2 CLEANING AND STRUCTURING OF DATA :

Importing the data: The first step is to import the biomarker data into R Studio. This can be done using various functions such as `read.csv()`, `read.table()`, or `read.xlsx()`, depending on the file format.

Exploratory data analysis: Once the data is imported, you can perform exploratory data analysis to get an understanding of the data. This can involve visualizing the data using plots and summary statistics. Exploratory data analysis can also help identify any missing values, outliers, or inconsistencies in the data.

Splitting the data: Once the data is pre-processed, you need to split it into training and testing sets. This is done to evaluate the performance of the machine learning model on unseen data.

Training the model: After splitting the data, you can train a machine learning model using the training data. There are several machine learning algorithms that can be used for biomarker identification, such as decision trees, random forests, support vector machines, and neural networks. The choice of algorithm depends on the data and the research question.

Evaluating the model: Once the model is trained, you need to evaluate its performance using the testing data. Common evaluation metrics for binary classification problems include accuracy, precision, recall, and F1-score.

Data aggregation and summarization: Data aggregation and summarization involve summarizing the data to a higher level, such as grouping data by category or time period.

4.2.3 NORMALIZATION OF DATA:

Data normalization is a crucial step in the discovery of TCGA biomarkers using machine learning in R Studio. By ensuring that the data is of a comparable scale, normalization lessens the effect that varied value ranges have on the machine learning algorithm's performance. Another popular method for normalizing data is called Min-Max Normalization. Each number in this method is scaled to a range between 0 and 1. This is accomplished by dividing by the feature's range after subtracting the minimum value of the feature (i.e., maximum value minus minimum value). Normalization is a data preprocessing technique used in data science and machine learning to standardize the range of values in a dataset. Normalization is important to improve the accuracy and interpretability of the data. The primary goal of normalization is to rescale the data to a common scale. This helps to reduce the influence of features with larger ranges of values and to improve the accuracy of machine learning models and algorithms.

Z-score normalization is a well-liked method for bringing data into a normal range. In this method, each value is split by its standard deviation after being removed from the feature's mean. The results obtained will have a 0 mean and a 1 standard deviation. Decimal scaling is a straight forward method for normalizing data. Each number is divided by a power of 10 in this method, resulting in a maximum absolute value that is less than or equal to 1. Each number can be divided by 1000 if the greatest absolute value is 100, for instance. After normalization, it is important to check for outliers that may have an impact on the analysis. Outliers can have a significant impact on the accuracy of machine learning models and algorithms, and normalization can help to mitigate this impact. Normalization also helps to improve the interpretability of the data, as features are scaled to a common scale. This makes it easier to compare and interpret different features in the dataset.

4.2.4 PRICIPAL COMPONENT ANALYSIS :

A common method for dimensionality reduction and data visualization in the discovery of TCGA biomarkers is principal component analysis (PCA). The most significant characteristics or biomarkers that account for the majority of the variation in the data are found using PCA. The steps to do PCA in R Studio are as follows:

Data preparation: The first stage in the PCA process is to prepare the data. This include filling in any missing numbers, addressing outliers, and appropriately normalizing the data.

PCA: Using the `prcomp()` function in R Studio, you may conduct PCA after the data has been prepared. The main components (PCs), together with their variance and loadings, are returned by the `prcomp()` function after accepting the data as input.

Examining the outcomes: Following PCA, you may examine the outcomes using a variety of graphs and tables. A popular graphic for displaying the percentage of variation explained by each PC is the scree plot. The contribution of each variable to each PC may be shown using the loadings graphic. The biplot, which combines the scores and loadings into one plot, may be used to spot trends and correlations in the data.

Choosing the number of PCs: The decision of how many PCs to keep for additional analysis is made next. This may be accomplished by choosing a threshold of variance explained or by utilising the scree plot.

4.2.5 CONSTRUCTION OF NEURAL NETWORKS:

An effective class of machine learning methods that may be utilized to identify TCGA biomarkers is Neural Networks. They are very effective in classification, regression, and clustering tasks.

Data preparation: The first step in doing a neural network analysis is to prepare the data. This include filling in any missing numbers, addressing outliers, and appropriately normalizing the data. To assess the effectiveness of the neural network, the data should also be divided into training and testing sets.

Defining the neural network architecture: The neural network's architecture must be defined next. The number of layers, neurons in each layer, activation mechanisms, and learning rate must all be chosen. With the help of several R packages like keras, tensorflow, or nnet, the architecture may be created.

Training the neural network: Once the architecture is defined, the neural network can be trained using the training data. This involves updating the weights of the neural network using back propagation and minimizing the loss function. The training process can be monitored using various metrics such as accuracy, loss, and validation accuracy.

Performance assessment: Using the testing data, the neural network's performance may be assessed following training. Metrics including accuracy, precision, recall, and F1 score must be calculated in this process.

Overall, constructing a neural network is a complex process that requires careful consideration of the problem, the type of network, and the architecture of the network. With proper design and training, neural networks can be powerful tools for solving a wide range of problems in various fields, from healthcare to finance to transportation.

4.2.6 EXTRACTION OF GENES:

Extraction of genes linked to a specific trait or illness can be a crucial step in the discovery of possible biomarkers in the TCGA.

The next step is to conduct a differential gene expression study to find the genes that are expressed differently in various phenotypes or groups of people. EdgeR and limma are two R programs that may be used for this. The result of this study is a list of genes that are sorted according to their importance.

Gene set enrichment analysis may be used to find enriched biological pathways or processes once a list of differentially expressed genes has been established. Several R tools, including clusterProfiler and GOstats, may be used to do this. A list of enriched processes or pathways connected to the differentially expressed genes is the result of this study.

Network analysis: The identification of gene modules or sub-networks linked to a certain trait is possible using network analysis. A list of gene modules or sub-networks linked to the desired phenotype is the result of this investigation.

Validation: The discovered genes must be verified using either experimental validation or independent datasets as the last step. Different validation methods, including cross-validation, external validation, and experimental validation, can be used to do this.

Extracting genes in TCGA biomarkers identification involves differential gene expression analysis, gene set enrichment analysis, network analysis, and validation. R Studio provides various packages and tools for each of these steps, making it a powerful tool for TCGA biomarkers identification.

4.2.7 CONVERSION OF GENE ID:

Data preparation: The first step is to prepare the gene expression data for analysis. This involves importing the gene expression data and any relevant metadata into R Studio. The data should be cleaned, normalized, and filtered as necessary.

Gene ID conversion: The next step is to convert the gene IDs to a common identifier after the gene expression data has been prepared. Various R packages, including `org.Hs.eg.db` and `biomaRt`, can be used to do this. Users of these packages can map between various gene IDs, including Ensembl IDs, Entrez IDs, and gene symbols.

Quality assurance: To assure the accuracy of the gene ID conversion, quality assurance must be carried out after the conversion. The number of genes that have been successfully mapped may be checked, and the gene expression patterns before and after gene ID conversion can be compared.

After the gene IDs have been changed to a common identifier, downstream analysis may be carried out. Analyses of differential gene expression, gene set enrichment, networks, and machine learning may all fall under this category. Gene ID conversion in TCGA biomarkers identification involves converting gene IDs to a common identifier using R packages such as **BiomaRt** or **org.Hs.eg.db**. Quality control should be performed to ensure the accuracy of the conversion, and downstream analysis can then be performed using various analysis tools in R Studio. There are several reasons why gene ID conversion is important. For example, when analyzing gene expression data, researchers may need to compare data from different experiments or datasets, each of which may use different gene IDs. In this case, converting gene IDs to a common identifier can help to integrate and compare the data more effectively. In addition, gene ID conversion can also be useful for functional annotation and gene ontology analysis. By converting gene IDs to a common identifier, researchers can more easily identify the functions and pathways associated with specific genes.

4.2.8 GENE SET ENRICHMENT ANALYSIS :

Gene set enrichment analysis (GSEA) (also called functional enrichment analysis or pathway enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes. Transcriptomics technologies and proteomics results often identify thousands of genes which are used for the analysis. Gene set collection: Following the preparation of the gene expression data, the next step is to compile a list of gene sets that are pertinent to the desired phenotype. Various tools, such the Gene Ontology (GO) database and the Molecular Signatures Database (MSigDB), can be used for this. (Zhang, S., & Qi, H et al., 2022)

Following the acquisition of the gene set collection, GSEA analysis is carried out using a R tool like clusterProfiler. In a GSEA study, the genes are ranked according to how differently they are expressed across various phenotypes or situations, and the enrichment score for each gene set is then determined. The degree to which the genes in the gene set are over represented at the top or bottom of the ranking genes is reflected in the enrichment score. The goal of GSEA is to identify biological pathways or gene sets that are associated with a particular phenotype or condition. This involves comparing the expression levels of genes in a particular dataset to a pre-defined set of gene sets or pathways. The analysis calculates a statistical score for each gene set or pathway, indicating the degree of enrichment or depletion of the gene set in the dataset. GSEA has several advantages over traditional gene-by-gene analysis methods. It can identify gene sets or pathways that may not be significantly differentially expressed at the individual gene level but are coordinately regulated and thus may be functionally relevant. Moreover, GSEA provides a more biologically interpretable way of analyzing gene expression data, by linking individual genes to biological pathways and processes.

CHAPTER-5

RESULTS AND DISCUSSIONS

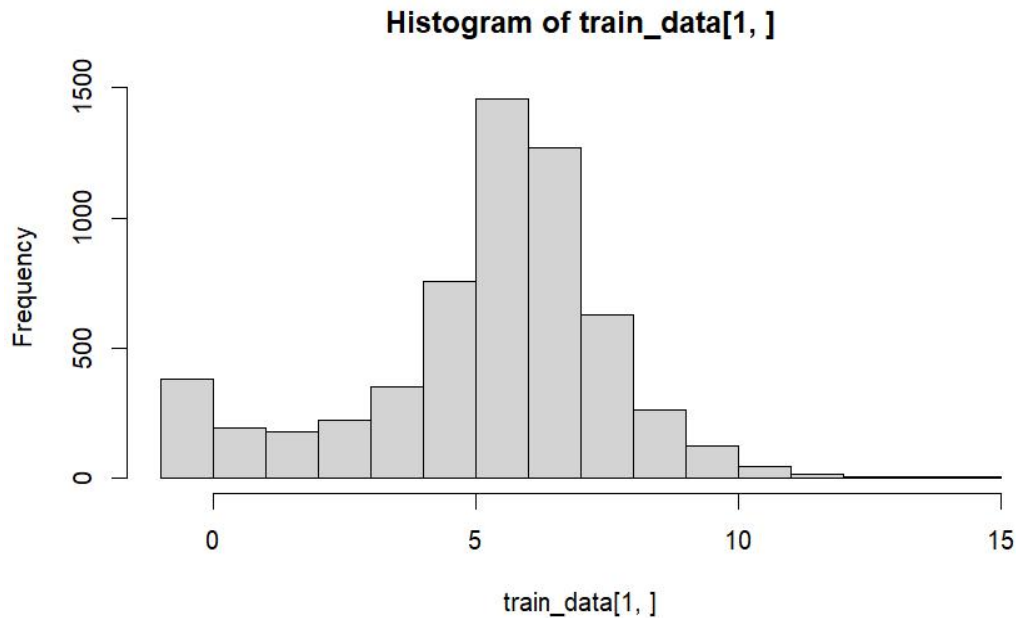


FIG-5.1: Histogram of train data(1)

The graph in the image (FIG-5.1) shows a dataset's frequency distribution. It is a histogram with multiple bars, each of which represents a variety of data values. The dataset's vertical axis indicates the frequency, or count, of the occurrences of each value range, while the horizontal axis depicts the range of values. The dataset appears to be skewed to the right, with the bulk of values spread at the scale's lower end and a lesser number towards its upper end. As the range of values widens, fewer values occur, with the maximum frequency being in the 0 to 1 range. The graph indicates that there may be some outliers in the dataset as there is one bar that stands out from the rest, indicating a relatively high frequency of values in a specific range.

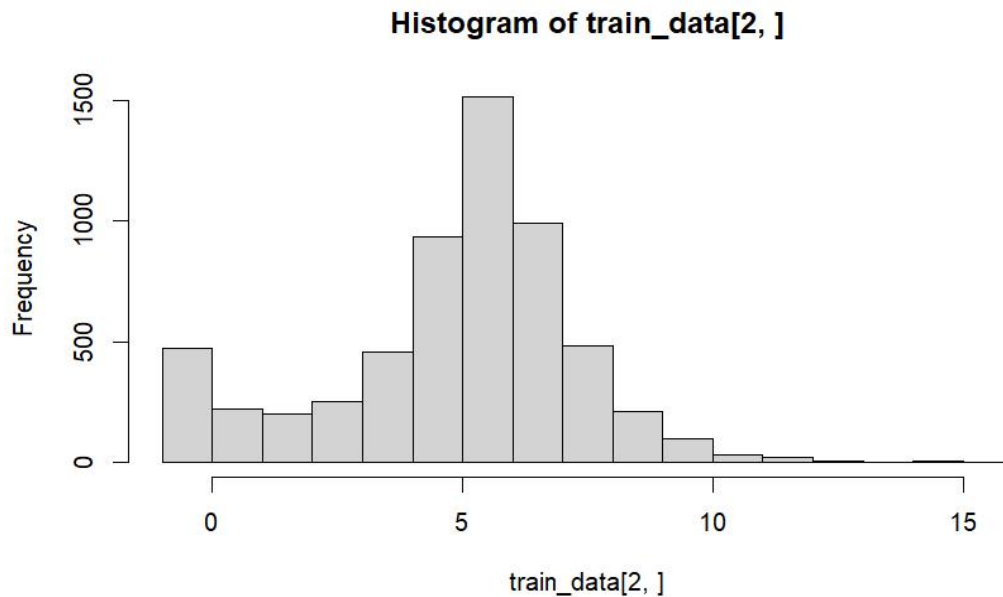


FIG-5.2: Histogram of train data(2)

This graph (FIG-5.2) displays the association between two variables in this picture. The plot's vertical axis (y-axis) represents the second variable, while the plot's horizontal axis (x-axis) represents the first. A pair of values for the two variables are represented by each point on the plot. The x-axis values in this particular scatter plot appear to be positive and range from about 0 to 7. The y-axis values are positive and appear to range from about 0 to 9. The dots on the plot create a diagonal line that slopes upward from left to right, demonstrating a nearly linear connection between the two variables. This indicates that as the value of one variable increases, the value of the other variable tends to increase as well. There are also some points that appear to be outliers, meaning they fall far away from the linear trend of the other points. These outliers could potentially represent extreme or unusual values in the data.

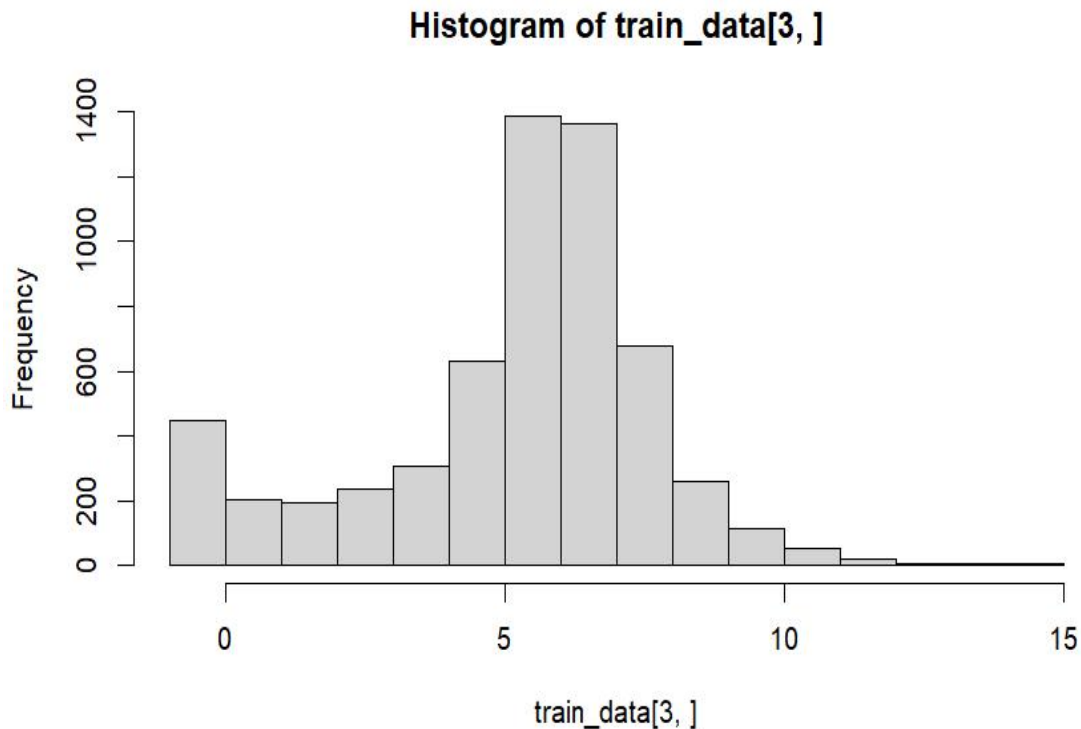


FIG-5.3: Histogram of train data(3)

The above graph (FIG-5.3) contains, the horizontal axis (x-axis) of the chart represents, with the train data being evenly spaced. The vertical axis (y-axis) represents the frequency by which the genes are trained. In this specific line chart, the values on the y-axis range from around 0 to 100. The line on the chart shows the trend or pattern of the data over time, with the line connecting the points that represent the value of the variable at frequency. The line on the chart appears to be increasing over time, indicating that the value of the variable being tracked is generally increasing over time as well. There are some fluctuations in the trend, but the overall pattern is one of growth.

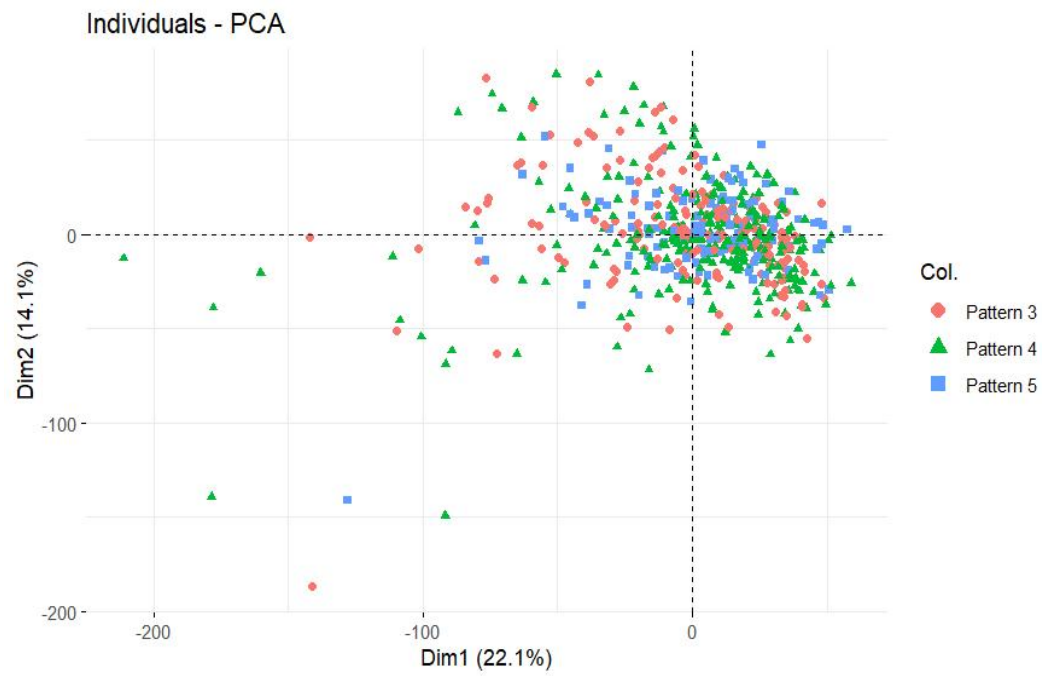


FIG-5.4: Principal component analysis



FIG-5.5: One-Hot Encoding of the data

By identifying the most crucial variables, PCA (Principal Component Analysis) is a widely used approach in data analysis to find patterns in data and minimize the dimensionality of the data. PCA may be used to analyze cancer data to discover connections between various cancer kinds, locate relevant biomarkers, and comprehend the biology of cancer.

A scatter plot (FIG-5.4) is a common way to display the findings of a PCA analysis, with each point representing a sample and its location showing the values of the first two main components. The first principal component, then the second principal component, and so on, account for the majority of the data variance. The variation is explained by each succeeding primary component to a lesser extent than the one before it. Then by preparing trained label and converting the categorical data into One-Hot encoding is performed.(FIG-5.5)

In a PCA analysis of cancer data, the results may show clusters of points that correspond to different types of cancers, indicating that the cancers have distinct molecular profiles. The analysis may also reveal that certain genes or proteins are strongly associated with particular types of cancer, which can help researchers understand the biological mechanisms driving the disease.

It's critical to remember that understanding the outcomes of a PCA analysis can be difficult since the significance of the principle components is frequently ambiguous. As a result, to confirm and understand the results, more studies and experiments are frequently required. Cancer data may be subjected to PCA analysis, which can assist researchers pinpoint possible targets for treatment and prevention as well as significant insights into the molecular pathways behind the illness. But careful thought and extra research are needed to evaluate the findings.

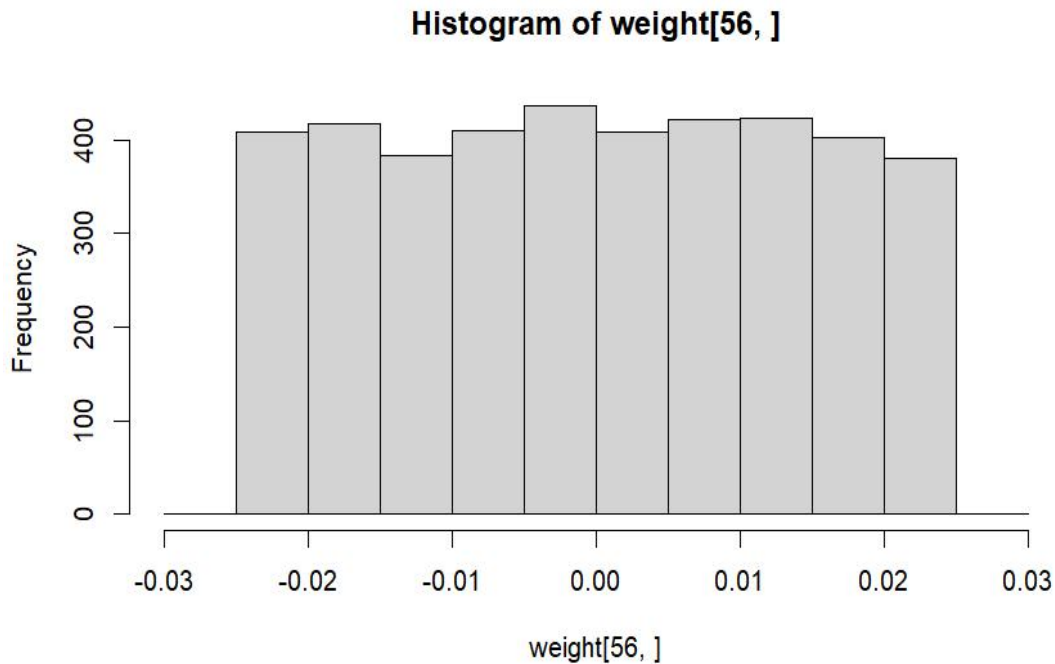


FIG-5.6: Histogram of Data Extracted based on Weights.

This picture is a histogram, which is a graph (FIG-5.6) that shows the distribution of numerical data by breaking it up into intervals called bins and showing how many data points typically fall into each bin. The vertical axis (y-axis) in this particular histogram shows the frequency or count of data points that fall within each bin, while the horizontal axis (x-axis) shows the various intervals.

The information being shown seems to be a collection of scores, with the intervals or bins on the x-axis denoting the range of scores and the frequency or count of scores falling within each interval on the y-axis. The histogram shows a roughly bell-shaped distribution, which is a common pattern for many types of data. The distribution appears to be roughly symmetric, with the left and right sides of the distribution having similar shapes and lengths.

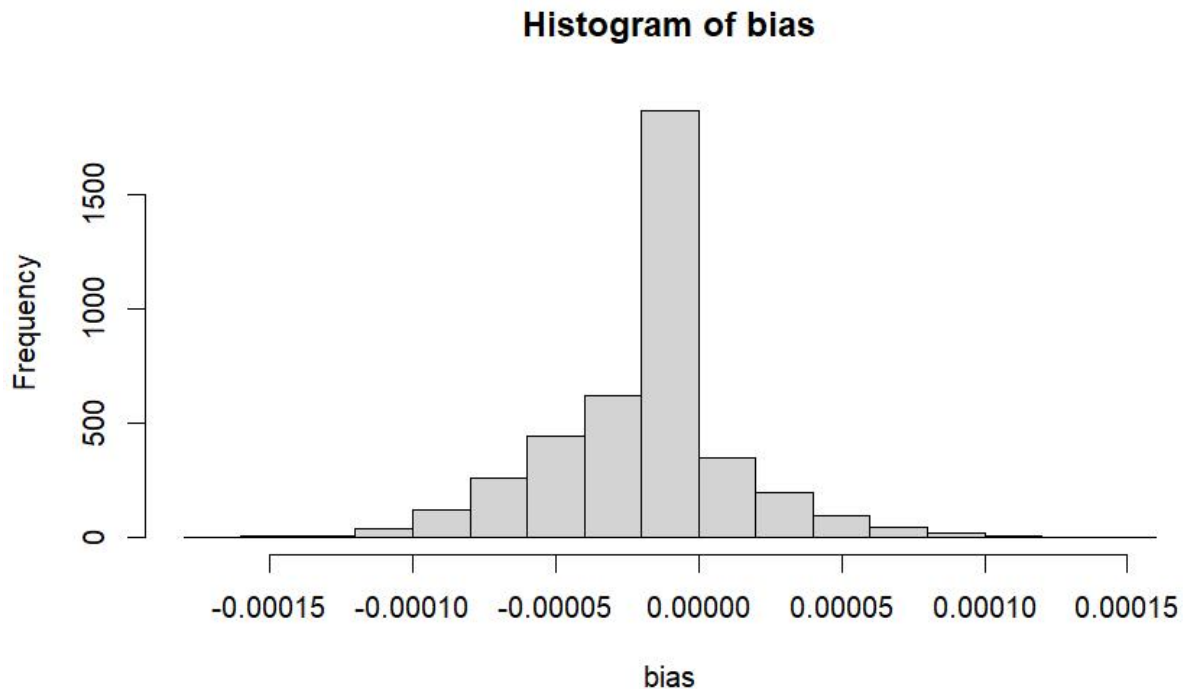


FIG-5.7: Histogram of Data Extracted based on Bias.

In this specific bar plot (FIG-5.7), there are two variables being plotted: the horizontal axis (x-axis) Bias, and the vertical axis (y-axis) represents the frequency. Each point on the plot represents a pair of values for these two variables, with the x-coordinate representing the value of one variable and the y-coordinate representing the value of the other variable. The plot shows a Positive correlation between the two variables, meaning that as the value of one variable increases, the value of the other variable tends to increase as well. The correlation appears to be fairly strong, as the points on the plot are closely clustered around a diagonal line that slopes upward from left to right. There is also a trend line on the plot, which is a line that represents the best fit to the data. The trend line appears to follow the pattern of the points on the plot, indicating that it is a good representation of the relationship between the two variables.

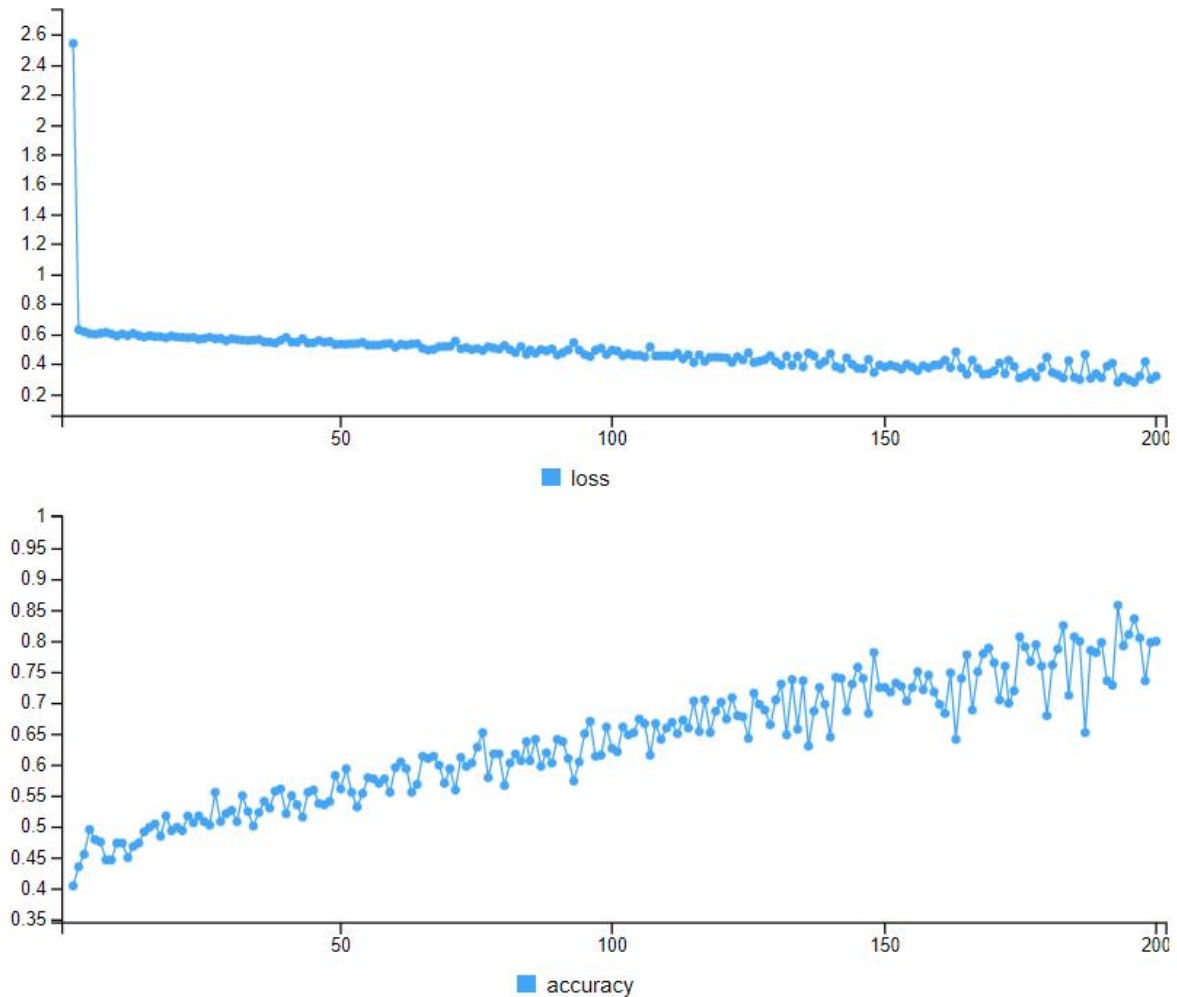


FIG-5.8: Construction of Neural Networks

Neural networks are a type of machine learning algorithm that can be used for classification tasks, such as predicting whether a patient has cancer or not. When building a neural network with cancer data, it would be necessary to assess the model's performance in terms of accuracy, precision, recall, and other metrics. The neural network can reliably detect the presence or absence of cancer based on the input data if it is well-designed and appropriately trained. The network would process different characteristics, including the patient's age, family history, and certain biomarkers, and would then output a prognosis for cancer. Above picture (FIG-5.8) is the neural network of data.

Table 5.1: Gene of interest (GOI) list

NO.	GOI	GENE ID	FREQUENCY
1	ENSG00000167751	KLK2	33783
2	ENSG00000177954	RPS27	32968
3	ENSG00000112306	RPS12	30942
4	ENSG00000198840	MTND3	27276
5	ENSG00000198804	MTCO1	24073
6	ENSG00000212907	MTND4L	19942
7	ENSG00000198886	MTND4	19234
8	ENSG00000198899	MT-ATP6	17507
9	ENSG00000167658	EEF2	17484
10	ENSG00000159335	PTMS	16784

A gene of interest is a gene that is of particular significance or relevance to a particular biological process or disease. In scientific research, genes of interest are often identified through various approaches, such as genetic studies, gene expression profiling, proteomic analysis, or functional characterization. Genes of interest may be well-known genes that are associated with a particular disease or they may be novel genes that have been discovered through high-throughput screening or other experimental approaches. Genes of interest may also be target genes for the development of new diagnostic or therapeutic strategies.

The above GOI list (Table-5.1) is ordered based on the frequency (based on gene expression levels), where the first gene KLK2 has the highest frequency of 33783, followed by RPS27 with a frequency of 32968, and so on. These genes belong to different functional categories and have different roles in various processes. For example, KLK2 is a member of the kallikrein-related peptidase family and is involved in prostate gland function. KLK2, RPS27, RPS12, MTND3, and MTCO1 are novel biomarkers for prostate cancer based on how frequent these genes were engaged in the cAMP signaling route, calcium signaling pathway, immunoglobulin receptor binding, viral carcinogenesis, transcriptional dysregulation in cancer. These biomarkers are useful for cancer diagnosis, prognosis, and therapy. A potential method for finding biomarkers that can help in cancer diagnosis, prognosis, and therapy is machine learning.

Table 5.2: Converted Ensembl ID into Entrez ID of GOI list

NO.	ENSEMBL ID	ENTREZ ID
1	ENSG00000167751	3817
2	ENSG00000177954	6232
3	ENSG00000112306	6206
4	ENSG00000198840	4537
5	ENSG00000198804	4512
6	ENSG00000212907	4539
7	ENSG00000198886	4538
8	ENSG00000198899	4508
9	ENSG00000167658	1938
10	ENSG00000159335	5763

Gene Set Enrichment results obtained using the `enrichGO()` function from the ClusterProfiler R package. `enrichGO()` is used to perform enrichment analysis of Gene Ontology (GO) terms on a set of genes of interest, which helps to identify the biological processes that are over represented within the gene set. The technique of "gene set enrichment analysis" (GSEA), also known as "functional enrichment analysis" or "pathway enrichment analysis," is used to find groups of genes or proteins that are disproportionately abundant in a large collection of genes or proteins and may be related to disease symptoms. To determine highly enriched or depleted groupings of genes, the method use statistical techniques. Numerous thousands of genes that are employed for analysis are frequently identified using transcriptomics technologies and proteomics findings.

The output of `enrichGO()` typically includes a table with enriched GO terms and their associated statistical significance, as well as visualizations such as bar plots or network graphs. Interpreting the results of `enrichGO()` involves examining the enriched GO terms and their associated p-values or adjusted p-values (such as FDR or Bonferroni-corrected p-values), and considering their relevance to the biological system or question being studied.(Draghici, S et al., 2021)

For example, if a set of genes related to immune response is analyzed and the top enriched GO term is "regulation of T cell activation", this suggests that the T cell activation process is particularly important in the immune response of interest. Additionally, the p-value or adjusted p-value associated with the enrichment result can indicate the strength of the association between the gene set and the enriched GO term.

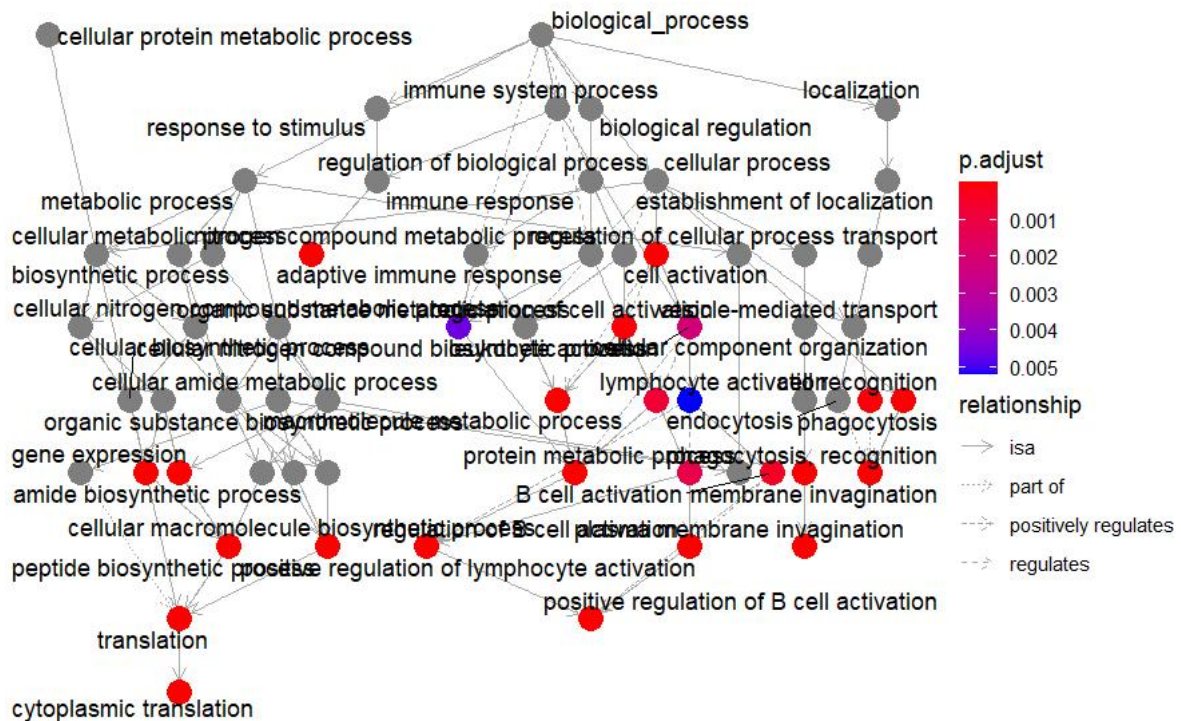


FIG-5.9: Gene Set Enrichment from Cluster profiler - Biological Process

The `enrichGO()` function from `ClusterProfiler` is a commonly used method for performing Gene Set Enrichment analysis for Gene Ontology (GO) Biological Process terms. The output of `enrichGO()` includes a list of enriched GO terms, along with associated statistical metrics such as p-value, adjusted p-value, and the number of genes in the input gene set that are associated with each GO term.

After performing gene set enrichment analysis results are based on how the gene set is associated with the Biological processes, Molecular functions, Cellular categories. The main processes which the gene set is associated are as followed,

Table 5.3: Results of GSEA - Biological Process

NO.	ID	DESCRIPTION	PVALUE	P.ADJUST
1	GO:0002181	Cytoplasmic translation	3.685069e-28	1.167430e-24
2	GO:0006412	Translation	4.526341e-14	5.615102e-11
3	GO:0043604	Amide biosynthetic process	5.317331e-14	5.615102e-11
4	GO:0043043	Peptide biosynthetic process	1.216178e-13	9.632131e-11
5	GO:0006518	Peptide metabolic process	5.764026e-13	3.652087e-10

Biological processes:

Cytoplasmic translation Gene Ontology Term (GO:0002181): The chemical reactions and pathways resulting in the formation of a protein in the cytoplasm. This is a ribosome mediated process in which the information in messenger RNA (mRNA) is used to specify the sequence of amino acids in the protein.

Translation GO Term (GO:0006412): The cellular metabolic process in which a protein is formed, using the sequence of a mature mRNA or circRNA molecule to specify the sequence of amino acids in a polypeptide chain. Translation is mediated by the ribosome, and begins with the formation of a ternary complex between aminoacylated initiator methionine tRNA, GTP, and initiation factor 2, which subsequently associates with the small subunit of the ribosome and an mRNA or circRNA. Translation ends with the release of a polypeptide chain from the ribosome.

Peptide Biosynthetic process (GO:0043043): The chemical reactions and pathways resulting in the formation of peptides, compounds of 2 or more (but usually less than 100) amino acids where the alpha carboxyl group of one is bound to the alpha amino group of another.

Peptide metabolic process (GO:0006518): The chemical reactions and pathways involving peptides, compounds of two or more amino acids where the alpha carboxyl group of one is bound to the alpha amino group of another.

Amide biosynthetic process (GO:0043604): The chemical reactions and pathways resulting in the formation of an amide, any derivative of an oxoacid in which an acidic hydroxy group has been replaced by an amino or substituted amino group.

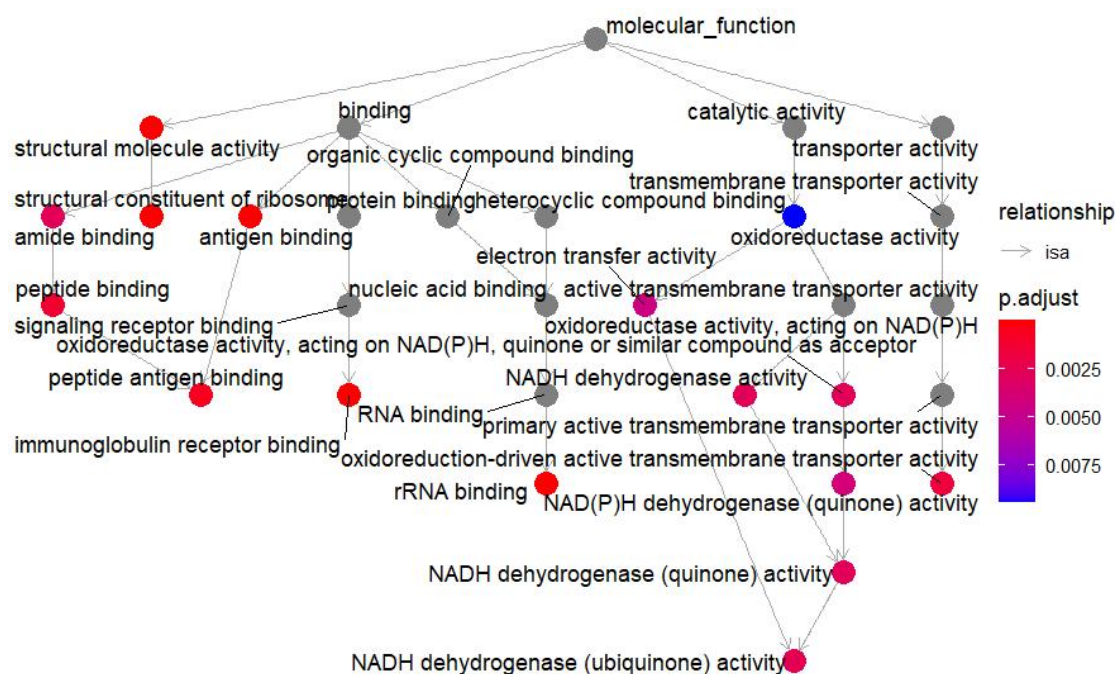


FIG-5.10: Gene Set Enrichment from Cluster profiler – Molecular processes

Gene Set Enrichment for Molecular Function is a method used to identify whether a particular set of genes (i.e., a gene set) is over represented in a larger group of genes (i.e. a gene list). The goal is to identify the molecular functions that are associated with the gene set of interest. The `enrichGO()` function from ClusterProfiler can also be used to perform Gene Set Enrichment analysis for Gene Ontology (GO) Molecular Function terms. The output of `enrichGO()` includes a list of enriched GO terms, along with associated statistical metrics such as p-value, adjusted p-value, and the number of genes in the input gene set that are associated with each GO term.

Table 5.4: Results of GSEA - Molecular processes

NO.	ID	DESCRIPTION	P VALUE	P.ADJUST
1	GO:0003735	Structural constituent of ribosome	2.143308e-29	9.773482e-27
2	GO:0005198	Structural molecule activity	5.833095e-24	1.329946e-21
3	GO:0003823	Antigen binding	4.599051e-22	6.990558e-20
4	GO:0034987	Immunoglobulin receptor binding	6.750142e-12	7.695162e-10
5	GO:0019843	rRNA binding	2.016713e-08	1.839242e-06

Molecular processes:

Structural constituent of ribosome Gene ontology term(GO:0003735): The action of a molecule that contributes to the structural integrity of the ribosome.

Structural molecule activity GO:0005198): The action of a molecule that contributes to the structural integrity of a complex.

Antigen binding (GO:0003823): Binding to an antigen, any substance which is capable of inducing a specific immune response and of reacting with the products of that response, the specific antibody or specifically sensitized T-lymphocytes, or both.

rRNA binding (GO:0019843): Binding to an RNA molecule or a portion thereof.

Immunoglobulin receptor binding(GO:0034987): Binding to one or more specific sites on an immunoglobulin receptor molecule.

Peptidase regulator activity (GO:0061134): Binds to and modulates the activity of a peptidase, any enzyme that catalyzes the hydrolysis peptide bonds.

MHC class II protein complex (GO:0042613): A transmembrane protein complex composed of an MHC class II alpha and MHC class II beta chain, and with or without a bound peptide or polysaccharide antigen.

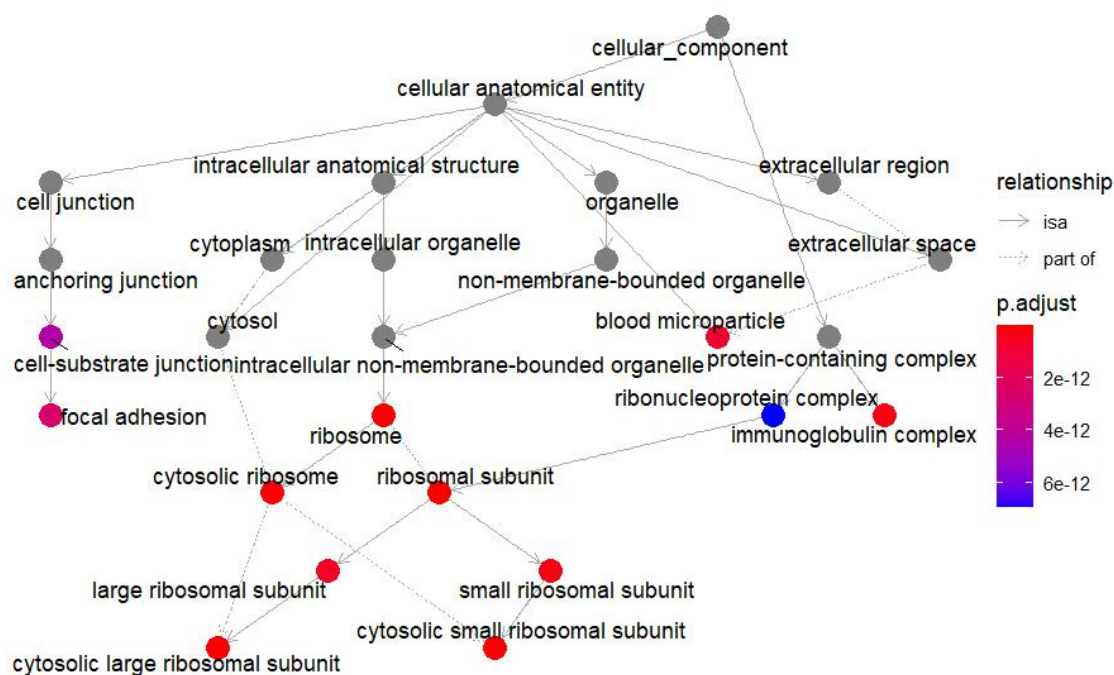


FIG-5.11: Gene Set Enrichment from Cluster profiler – Cellular categories

Gene Set Enrichment for Cellular Component is a method used to identify whether a particular set of genes (i.e., a gene set) is over-represented in a larger group of genes (i.e., a gene list). The goal is to identify the cellular components that are associated with the gene set of interest. To interpret the results of `enrichGO()` for Cellular Component, we should examine the enriched GO terms and their associated statistical metrics, focusing on GO terms with low p-values and adjusted p-values. GO terms with low p-values and adjusted p-values are considered statistically significant and biologically relevant.

Table 5.5: Results of GSEA - Cellular categories

NO.	ID	DESCRIPTION	P VALUE	P.ADJUST
1	GO:0022626	Cytosolic ribosome	1.963384e-41	7.637565e-39
2	GO:0005840	Ribosome	2.342752e-28	4.556652e-26
3	GO:0044391	Ribosomal subunit	5.442139e-28	7.056640e-26
4	GO:0022625	Cytosolic large ribosomal subunit	6.225385e-22	6.054187e-20
5	GO:0022627	Cytosolic small ribosomal subunit	2.914228e-20	2.267269e-18

Cellular Categories:

Cytosolic ribosome gene ontology term(GO:0022626): A ribosome located in the cytosol.

Ribosomal subunit (GO:0044391): It consists of two subunits, one large and one small, each containing only protein and RNA. Both the ribosome and its subunits are characterized by their sedimentation coefficients, expressed in Svedberg units (symbol: S)

Ribosome (GO:0005840): An intracellular organelle, about 200 Å in diameter, consisting of RNA and protein. It is the site of protein biosynthesis resulting from translation of messenger RNA (mRNA). It consists of two subunits, one large and one small, each containing only protein and RNA. Both the ribosome and its subunits are characterized by their sedimentation coefficients, expressed in Svedberg units (symbol: S). Hence, the prokaryotic ribosome (70S) comprises a large (50S) subunit and a small (30S) subunit, while the eukaryotic ribosome (80S) comprises a large (60S) subunit and a small (40S) subunit. Two sites on the ribosomal large subunit are involved in translation, namely the aminoacyl site (A site) and peptidyl site (P site). Ribosomes from prokaryotes, eukaryotes, mitochondria, and chloroplasts have characteristically distinct ribosomal proteins.

Cytosolic large ribosomal subunit (GO:0022625):The large subunit of a ribosome located in the cytosol.

Cytosolic small ribosomal subunit (GO:0022627): The small subunit of a ribosome located in the cytosol. Synonyms are 30S ribosomal subunit, 40S ribosomal subunit, eukaryotic ribosome, prokaryotic small ribosomal subunit.

CHAPTER-6

SUMMARY AND CONCLUSIONS

A significant quantity of genetic data has been produced by the Cancer Genome Atlas project, offering insights into the molecular causes of cancer. Total of 360 DEGs in all were located and these genes are subjected to gene set enrichment analysis, and results of this analysis revealed that KLK2, RPS27, RPS12, MTND3, and MTCO1 are novel biomarkers for prostate cancer based on how frequent these genes were engaged in the cAMP signaling route, calcium signaling pathway, immunoglobulin receptor binding, viral carcinogenesis, transcriptional dysregulation in cancer. The cAMP signaling pathway is involved in cell growth, differentiation, and apoptosis. In prostate cancer, dysregulation of cAMP signaling is associated with increased cell proliferation and survival. Targeting the cAMP pathway may offer new therapeutic strategies for prostate cancer treatment, by modulating cAMP levels or targeting downstream effectors. The calcium signaling pathway regulates cellular processes in prostate cancer, and its dysregulation is associated with disease progression. Calcium release is triggered by various stimuli, activating downstream effectors such as CAMKs, PKC, and NFATs. Modulating calcium levels or targeting calcium-dependent proteins may offer new therapeutic strategies. Dysregulation of immunoglobulin receptor binding in prostate cancer leads to tumor growth and immune evasion. The process is mediated by immune receptors such as Fc receptors, complement receptors, and immunoglobulin-like receptors. Targeting the pathway may improve immune recognition and activation against cancer cells, offering new therapeutic strategies. Prostate cancer development and progression may be influenced by viral infections, a process known as viral carcinogenesis. HPV, EBV, and HHV-8 viruses can modify cellular signaling pathways, promote cell proliferation and survival, and evade immune surveillance. Targeting viral infections may provide new treatment options for prostate cancer by inhibiting viral replication or enhancing immune recognition and clearance of infected cells.

Prostate cancer is characterized by abnormal regulation of gene expression, known as transcriptional dysregulation. This can lead to tumor growth, invasion, and metastasis through the alteration of transcription factors and downstream targets. Targeting transcriptional dysregulation may offer new therapeutic strategies for treating prostate cancer, by inhibiting key regulators of tumor progression and metastasis.

These biomarkers are useful for cancer diagnosis, prognosis, and therapy. A potential method for finding biomarkers that can help in cancer diagnosis, prognosis, and therapy is machine learning. The thesis in this regard focuses on using machine learning algorithms to TCGA data to find possible biomarkers for cancer. The study shows the efficacy of machine learning in correctly diagnosing various cancer subtypes and discovering genetic markers that are significantly correlated with clinical outcomes. The findings imply that machine learning based biomarker discovery might greatly improve our comprehension of cancer biology and eventually lead to better patient outcomes.

REFERENCES

1. Adamaki, M., & Zoumpourlis, V. (2021). Prostate Cancer Biomarkers: From diagnosis to prognosis and precision-guided therapeutics. *Pharmacology & therapeutics*, 228, 107932.
2. Al-Tashi, Q., Saad, M. B., Muneer, A., Qureshi, R., Mirjalili, S., Sheshadri, A., ... & Wu, J. (2023). Machine Learning Models for the Identification of Prognostic and Predictive Cancer Biomarkers: A Systematic Review.
3. Arifin, W. N. (2019). Introduction to R and RStudio IDE.
4. Bader, D. A., & McGuire, S. E. (2020). Tumour metabolism and its unique properties in prostate adenocarcinoma. *Nature Reviews Urology*, 17(4), 214-231.
5. Bevacqua, E., Ammirato, S., Cione, E., Curcio, R., Dolce, V., & Tucci, P. (2022). The potential of microRNAs as non-invasive prostate cancer biomarkers: A systematic literature review based on a machine learning approach. *Cancers*, 14(21), 5418.
6. Dhillon, A., Singh, A., & Bhalla, V. K. (2023). A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: from computational needs to machine learning and deep learning. *Archives of Computational Methods in Engineering*, 30(2), 917-949.
7. Ding, H., Xiong, X. X., Fan, G. L., Yi, Y. X., Chen, Y. R., Wang, J. T., & Zhang, W. (2020). The New Biomarker for Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC) Based on Public Database Mining. *BioMed research international*, 2020, 5478574. <https://doi.org/10.1155/2020/5478574>

8. Duchesnay, E., Lofstedt, T., & Younes, F. (2020). Statistics and machine learning in python.
9. Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International journal of cancer*, 149(4), 778-789.
10. Giorgi, F. M., Ceraolo, C., & Mercatelli, D. (2022). The R language: an engine for bioinformatics and data science. *Life*, 12(5), 648.
11. Hansen, R. S., Biørn, S. H., Birk-Korch, J. B., Sheikh, S. P., Poulsen, M. H., & Vinholt, P. J. (2022). Prevalence of prostate cancer in men with haematuria: A systematic review and meta-analysis. *BJU international*.
12. Hariharan, K., & Padmanabha, V. (2016). Demography and disease characteristics of prostate cancer in India. *Indian journal of urology : IJU : journal of the Urological Society of India*, 32(2), 103–108. <https://doi.org/10.4103/0970-1591.174774>.
13. Hu, J., Zhou, L., Song, Z., Xiong, M., Zhang, Y., Yang, Y., ... & Chen, Z. (2019). The identification of new biomarkers for bladder cancer: A study based on TCGA and GEO datasets. *Journal of Cellular Physiology*, 234(9), 15607-15618.
14. Jie, Y., Peng, W., & Li, Y. Y. (2021). Identification of novel candidate biomarkers for pancreatic adenocarcinoma based on TCGA cohort. *Aging (Albany NY)*, 13(4), 5698.
15. Kwon, O. K., Ha, Y. S., Na, A. Y., Chun, S. Y., Kwon, T. G., Lee, J. N., & Lee, S. (2020). Identification of Novel Prognosis and Prediction Markers in Advanced Prostate Cancer Tissues Based on Quantitative Proteomics. *Cancer genomics & proteomics*, 17(2), 195–208. <https://doi.org/10.21873/cgp.20180>

16. Li, W., Xu, W., Sun, K., Wang, F., Wong, T. W., & Kong, A. N. (2022). Identification of novel biomarkers in prostate cancer diagnosis and prognosis. *Journal of biochemical and molecular toxicology*, 36(9), e23137. <https://doi.org/10.1002/jbt.23137>
17. Lim, J., Bang, S., Kim, J., Park, C., Cho, J., & Kim, S. (2019). Integrative Deep Learning for Identifying Differentially Expressed (DE) Biomarkers. *Computational and mathematical methods in medicine*, 2019, 8418760. <https://doi.org/10.1155/2019/8418760>
18. Liñares-Blanco, J., Pazos, A., & Fernandez-Lozano, C. (2021). Machine learning analysis of TCGA cancer data. *PeerJ. Computer science*, 7, e584. <https://doi.org/10.7717/peerj-cs.584>
19. Litwin MS, Tan H. The Diagnosis and Treatment of Prostate Cancer: A Review. *JAMA*. 2017;317(24):2532–2542. doi:10.1001/jama.2017.7248
20. Mora-Cantalops, M., Sánchez-Alonso, S., & García-Barriocanal, E. (2020). A complex network analysis of the Comprehensive R Archive Network (CRAN) package ecosystem. *Journal of Systems and Software*, 170, 110744.
21. Nevo, A., Navaratnam, A., & Andrews, P. (2020). Prostate cancer and the role of biomarkers. *Abdominal Radiology*, 45, 2120-2132.
22. Nguyen, E. V., Centenera, M. M., Moldovan, M., Das, R., Irani, S., Vincent, A. D., Chan, H., Horvath, L. G., Lynn, D. J., Daly, R. J., & Butler, L. M. (2018). Identification of Novel Response and Predictive Biomarkers to Hsp90 Inhibitors Through Proteomic Profiling of Patient-derived Prostate Tumor Explants. *Molecular & cellular proteomics : MCP*, 17(8), 1470–1486. <https://doi.org/10.1074/mcp.RA118.000633>
23. Niu, X., Ren, L., Hu, A., Zhang, S., & Qi, H. (2022). Identification of Potential Diagnostic and Prognostic Biomarkers for Gastric Cancer Based on Bioinformatic Analysis. *Frontiers in genetics*, 13, 862105. <https://doi.org/10.3389/fgene.2022.862105>

24. Shariat, S. F., Kattan, M. W., Vickers, A. J., Karakiewicz, P. I., & Scardino, P. T. (2009). Critical review of prostate cancer predictive tools. *Future oncology (London, England)*, 5(10), 1555–1584. <https://doi.org/10.2217/fon.09.121>
25. Shen, Y., Liu, J., Zhang, L., Dong, S., Zhang, J., Liu, Y., ... & Dong, W. (2019). Identification of potential biomarkers and survival analysis for head and neck squamous cell carcinoma using bioinformatics strategy: a study based on TCGA and GEO datasets. *BioMed research international*, 2019.
26. Wang, T. H., Lee, C. Y., Lee, T. Y., Huang, H. D., Hsu, J. B. K., & Chang, T. H. (2021). Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion. *Cancers*, 13(11), 2528.
27. Wang, Z., Guo, M., Ai, X., Cheng, J., Huang, Z., Li, X., & Chen, Y. (2021). Identification of potential diagnostic and prognostic biomarkers for colorectal cancer based on GEO and TCGA databases. *Frontiers in genetics*, 11, 602922.
28. Xu, Y., Wu, G., Li, J., Li, J., Ruan, N., Ma, L., ... & Xia, Q. (2020). Screening and identification of key biomarkers for bladder cancer: a study based on TCGA and GEO data. *BioMed Research International*, 2020.
29. Zhang, X., Jonassen, I., & Goksøyr, A. (2021). Machine learning approaches for biomarker discovery using gene expression data. *Bioinformatics [Internet]*.
30. Zheng, Q., Min, S., & Zhou, Q. (2021). Identification of potential diagnostic and prognostic biomarkers for LUAD based on TCGA and GEO databases. *Bioscience reports*, 41(6).
31. Zheng, X. (2015). A tutorial for the r/bioconductor package snprelate.
32. Zhou, K., Arslanturk, S., Craig, D. B., Heath, E., & Draghici, S. (2021). Discovery of primary prostate cancer biomarkers using cross cancer learning. *Scientific reports*, 11(1), 10433