

Machine Learning Based Cancer Type Classification using NGS Data

Enrolment Number: 181505

Name of Student: Anas Malik

Name of Supervisor: Dr. Tiratha Raj Singh



SESSION- 2018-2022

*Dissertation Submitted in partial fulfilment of the requirement for the
degree of*

BACHELOR OF TECHNOLOGY

DEPARTMENT OF BIOINFORMATICS AND BIOTECHNOLOGY

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled “**Machine Learning Based Cancer type classification using NGS Data**” in the partial fulfilment of the requirements for the award of the Degree of B.Tech in Bioinformatics and submitted in Jaypee University of Information Technology, Solan, is an authentic record of my own work carried out during the period from session 2021 to 2022 under the supervision of Personal Investigator **Dr. Tiratha Raj Singh**, Assistant Professor, Jaypee University of Information Technology

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

ANAS MALIK

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:14-05-2022

Signature of Supervisor

CERTIFICATE

This is to affirm that the project report entitled: “*Machine Learning based Cancer type Classification using NGS Data*” put together by Anas Malik is in its partial satisfaction for the award of level of Bachelor of Technology in Bioinformatics to Jaypee University of Information Technology, Wagnaghat Solan, has been done under my watch

This work has not been submitted incomplete or completely to some other college or institution so as to accomplish any award or some other degree.

Signature.....

Supervisor Name: **Dr. Tiratha Raj Singh**

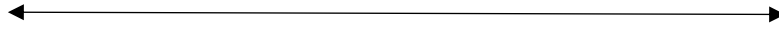
Designation: **Associate professor**

Jaypee University of Information Technology

Wagnaghat, Solan

Himachal Pradesh (173221)

ACKNOWLEDGMENT



I would like to express our deep and sincere sense of gratitude to our supervisor **Dr. Tiratha Raj Singh** for his unwavering support and guidance throughout the course of the project. Without his expertise in the subject, we would not have been able to complete the project credibly and on time. We would also like to acknowledge with thanks the kind of patronage, motivation, and constructive criticism we received from **Mr. Rohit Shukla** as these contributed immensely to the evolution of our ideas on the project.

.....

Name: ANAS MALIK

Enrolment no.: 181505

B.Tech Bioinformatics

2018-2022

List of Abbreviations

1. NGS – Next generation Sequencing
2. RNA – Ribonucleic Acid
3. SEQ - Sequence
4. mRNA – Messenger RNA
5. snRNA – Small Nuclear RNA
6. PCA – Principal Component Analysis
7. L-R – Logistic regression
8. KNN- K-Nearest Neighbour
9. SVM – Support Vector Machine
10. MLP – Multi layer perceptron
11. NCI - National Cancer Institute
12. GDC – Genomics Data Commons
13. TCGA – The cancer Genome Atlas
14. FPKM – Fragments per Kilobase of transcript per Million mapped reads
15. MIN - Minimum
16. MAX - Maximum
17. EXP - Expression
18. MT - Mitochondrial

ABSTRACT

Cancer is the diseases where therapy sensitivity differs greatly between patients living at the different regions [1]. However, thanks to tremendous developments in genetics and technological advancements, our understanding of cancer biology has improved considerably over the years. Genome structure and functions, as well as the processes that regulate gene expression, have been widely researched in the domain of oncology since the completion of the Human Genome Project and the advancement of next-generation strategies (NGS) [2,3].

The classification of various genes using the gene expression data i.e., the transcriptomics data can help to finding the relatable Biomarkers which can be helpful in the cancer diagnosis and also by knowing the various targeted genes causing the various kinds of cancers can be helpful in the drug discovery for the targeted gene.

RNA Seq data analysis can lead to differentiate and the analysis of the different types of tumours causing by the different types of cancer-causing genes and also find the different types of allele specific expression . As the use of various machine learning and deep learning approaches in the transcriptomic data for the analysis of genes and the prediction of gene can enhance the way for identification of the highly related biomarker and also help to compare the different types of tumours caused by the same or different types of genes in different types of cancer.

Keywords: NGS, Gene Expression, Cancer, Machine learning, prediction, Biomarkers, RNA sequence, Transcriptomics

TABLE OF CONTENT

Chapters	Title	Page No.
	Acknowledgement	4
	List of Abbreviations	5
	Abstract	6
	Table of content	7
	List of figures	8
Chapter -1	Introduction	9-12
1.1	Background	10
1.2	Transcriptomic data	11
1.3	Predicting health outcomes using gene expression data	12
Chapter -2	Review of Literature	13-18
2.1	Overview	14
2.1.1	Unsupervised Learning	15
2.1.2	Principle component analysis (PCA)	15
2.1.3	Auto encoder	15
2.1.4	Supervised Learning	15
2.1.4.1	Logistic Regression	16
2.1.4.2	Random Forest	16
2.1.4.3	k-nearest Neighbour	17
2.1.4.4	MLPClassifier	17
2.1.4.5	Support Vector Machine	17
2.1.5	Semi- Supervised Learning	18
Chapter -3	Methodology	19-31
3.1	NIH: GDC Portal	20
3.1.1	The national Cancer Institute	21
3.1.2	GDC Data Portal	21
3.2	Dataset preparation	22
3.3	Methodology	22-30
3.4	Data analysis	31
Chapter -4	Results and Discussion	32-39
4.1	Results	33

4.2 Discussion		37
-----	Appendix-1	40-42
-----	Appendix-2	43-45
-----	Appendix-3 (GUI)	46
REFERENCES		47-49

LIST OF FIGURES

Figure No.	Title	Page No.
1	Data driven detection of cancer and prediction of genes	9
2	Principle component analysis (PCA)	15
3	Logistic Regression	16
4	Random forest	16
5	K-Nearest neighbour	17
6	Multi-layer perceptron classifier (MLP)	17
7	Support Vector Machine (SVM)	18
8-11	GDC Data Portal (Data files, Data files cart, Data Transfer tool)	20,24,25
12	FKMP file in .txt Format	25
13-17	Data set after Various data cleaning processes	26-29
18-19	Dataset using Pandas data frame	30
15-20	Plot of Gene Expression Values	31
21	Graph plot of Gene expression value of biomarker gene	34
22	ROC Curve of SVM and Logistic Regression	35
23-26	Graphic user Interfaces of the prediction tool	38-39

Chapter-1: Introduction

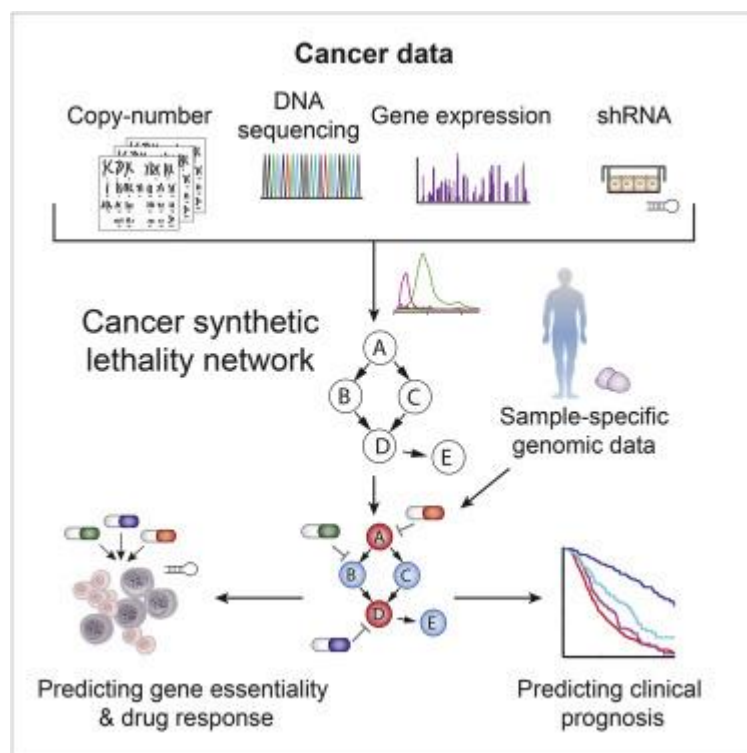


Figure 1: Data driven detection of cancer and prediction of genes source: Predicting cancer specific vulnerability via data driven of synthetic lethality, <https://www.sciencedirect.com/science/article/pii/S009286741400>

1.1 Background:

Cancer is the diseases where therapy sensitivity differs greatly between patients living at the different regions [1]. However, thanks to tremendous developments in genetics and technological advancements, our understanding of cancer biology has improved considerably over the years. Genome structure and functions, as well as the processes that regulate gene expression, have been widely researched in the domain of oncology since the completion of the Human Genome Project and the advancement of next-generation strategies (NGS) [2,3].

Transcriptomics is a branch of genetic mutation that has shown remarkable growth in recent years [2,4]. A transcriptome is a collection of all RNA molecules transcript from a specific genome in a specific cell, at a specified stage of development, and under specific physiological or pathological settings [5,6]. The transcriptome profile can be thought of as a summary of the cellular structure, so analysing it only yields a lot of information on gene function [7]. As a result, transcriptome analysis is thought to be an effective tool for studying cancer cells that are continually changing at the cellular level.

Here, we have dealt with transcriptomics data and created a dataset containing gene expression data associated with different types of cancers which can be further used to create machine learning models that can help in predicting health outcomes [25].

Earlier the cancer prediction and the classification of cancer into different segment was very clinical based and morphological with many limitations in finding the biomarker gene.

In order to get the best way to find the solution of a problems of cancer type classification we need a systematic approach based on the analysis of gene expression data [8]. And the accurate predictions of different type of cancer by using various deep learning and machine learning techniques has the very great value in providing the best target finding for the drug discovery and the better treatment of the cancer at the early states.

Also, the presence of the cancer-causing genes can be identified by using the RNA Seq analysis and can predict the type of cancer and the severity of the cancer by various machine learning approaches and various statistical approach

1.2 Transcriptomic data:

Transcriptomics is a term used for the information from the mRNA transcript within the biological sample when all the large-size samples are analysed and the dataset of the gene is known as transcriptomics data. It refers to a cell's entire set of transcriptomes at a certain genotypic or biological condition. The characterization of all transcriptional activity (i.e., coding and non-coding) or a selection of RNA transcripts within a particular sample is called transcriptomic data analysis. Transcriptome analysis enables the identification of potential genes and expression markers linked to desired characteristics. The various analysis cases are the:

1. RNA-Seq sequencing data analysis
2. mRNA data analysis
3. snRNA re-sequencing data analysis

Transcriptomics data provides the identification of the expressed genes within the genes and also the identification of splicing genes variants, relative expression of relative genes, expression value of genes, differential gene analysis, and target gene prediction can be analysed by the transcriptomics data of RNA. In research, the transcriptome is also used to gain insights into various carcinogenesis and transcription regulation and help in the generation of phylogenetic relationships. Transcriptomics-based data benefits from the proper data pre-processing techniques and regularized regression methods.

The analysis of transcriptomics data is having been used to understand the tumour, heterogeneity of tumours, and also helps to classify the tumour into various molecular subtypes. As transcriptomics data also consist of gene expression data which can be used to establishing signatures and also to predict the response and patient response to a particular cancer.

By doing the analysis of the transcriptomics data by using the various machine learning and statistical approach can be used to determine the gene response to the various tissues of the organism.

1.3 Predicting Health Outcomes from Gene Expression:

Characterizing and the classification of the genes into the category of normal cell and the abnormal cell based on the gene expression patterns within the cell under various conditions are very important [9] for the prediction of health outcomes from the gene expression. Gene expression profiling is a proven Biological technology approach for capturing the response of cells to a disease or drug treatments [10] as well as predicting the genes responsible for causing the disease.

The gene expression value is the prediction value of various genes that express themselves for particular cancer or disease [11]

Some genes are closely linked, and some gene co-expression approaches have been explored and applied to evaluate the correlation among their gene expression values. [12].

Therefore, gene expression values are significant and can help us to identify certain unique patterns among different types of cancers which can be further used in disease prediction.

mRNA transcriptomics Gene Expression data consist of data that contain the expression value of different genes which are responsible in causing the cancer, as multiple genes can be responsible for causing a single type of cancer or tumour also same gene can also be responsible for causing the different types of cancer or different types of tumours based on their expression values.

Gene expression profiling is a very important parameter which can be used to capture the cell response to a particular drug.

In various studies it comes the outcome that the gene expression data is highly correlated and about ~1000 genes can cover around nearly 80% of gene expression profile [13] and these genes are referred to as landmark genes, while the other genes are referred to as target genes. [14]. And the cost for the analysing the only 1000 landmark genes are much lower than the complete genome.

Chapter-2: Review of Literature

2.1 Overview:

Genes are found in isolation rarely, so it is expected the combination of genes may act more efficiently as than the individual gene in predictive phenotypes of gene.

Gene expression is a way for genes/ genes by which they are transcribed and work as productive gene products like proteins or functional RNA like rRNA, tRNA or sRNA. Regulation and alteration in gene expression can be done at various levels such as at transcription initiation, splicing or alternate splicing, post translational mechanisms etc. Gene expression is an integral part in developing the cells and in their differentiation. These alterations (like managing the expression level or location or time of gene transcription) in cells let them adjust to various critical conditions around them. Analysing the gene expression open the door for detecting and quantifying the gene transcription of all the genes of genome. This leads to the discovery of various genes involved in molecular pathways or diseases.

In the RNA- Seq data by analysing the protein coding regions we can get the insights of the proteins which contain the frequent chances of getting mutations that may lead to the conversion of the normal cell into tumour cells [21].

It has been seen that there is the specific pattern of the of the expression of the gene occur at the various stages of the cell development, physiological function of the cell, and the conversion of the cell into the abnormal cell [17]. Sometimes due to the change in the gene expression value of the certain genes the cell or the Normal cell get converted into the tumour cell (i.e., Malignant tumour or Benign Tumour) through the series of mutations in the gene [18]. Micro array and the serial analysis of the gene are the 2 methods which can be helpful in measuring the genome wide expression values [19]

Various classification methods of the machine learning and statistics has been extensively studies by the various researchers, and many new algorithms are developed which are extensively helpful in the prediction and the classification of the cancer genome data

As the combination of gene can enhance the working of model. some feature could show some biological importance to a particular disease and the combination of genes can show the interaction graph or ontological activities of a gene [16]

Some genes can be existed in different form and responsible for the causing different types of cancers which can be classified into benign and malignant.

To get the better insights the of the cancer classification, there are various machine learning approaches that can be utilized for analysing gene expression data.

2.1.1 Unsupervised Learning

Unsupervised machine learning is a type of machine learning in which the algorithm is not given any labels or training data points earlier. As a result, unsupervised learning algorithms should be able to discover any patterns in the training data that emerge naturally. Typical examples include aggregation, in which the algorithm automatically groups its training examples into categories with similar characteristics, key component analysis, in which the algorithm identifies which features are most useful for distinguishing between different training models, and remainder. This is in contrast to supervised learning, which uses pre-allocated class labels in the training data.

2.1.2 Principal component analysis (PCA).

PCA is a mathematical procedure that converts a group of connected variables into a set of unrelated variables by using orthogonal conversion. PCA is a popular method for analysing test data and incorporating machine learning into hypothetical models.

In addition, PCA is an unregulated mathematical method used to test the correlation between a set of variables. It is also known as the standard factor analysis when the regression determines the line of equity best.

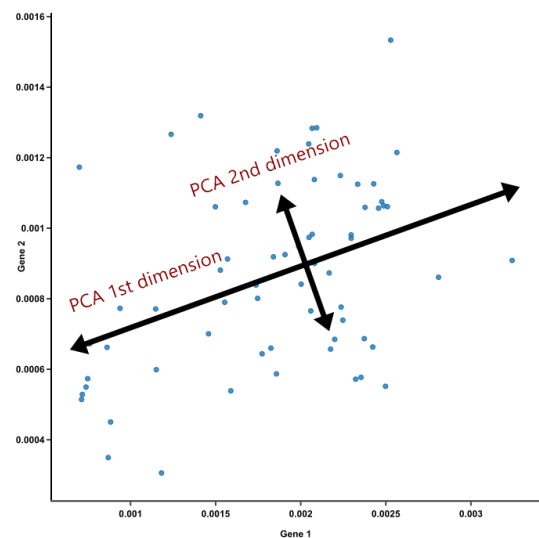


Figure 2: PCA, Source: Bio-Turing's blog

2.1.3 Autoencoder

An autoencoder is a type of artificial neural network that is used to read active code-free data codes (unregulated reading). The code is verified and refined in an attempt to reproduce the input from encoding. The autoencoder learns the representation (codec) of a data set, usually for size reduction, by training the network to ignore unimportant data ("noise").

2.1.4 Supervised Learning

Supervised learning is a form of learning algorithms of machine learning in which machines are trained using "well-labelled" training data and then predict output based on that data. Labelled data indicates that some input data has already been assigned to the appropriate output.

The training data provided by the computers functions as a supervisor in supervised learning, teaching the machines to anticipate output efficiently. When a student is learning under the supervision of a teacher, the same idea applies.

The technique of supplying input and output data directly to a machine learning model is known as supervised learning. A supervised learning algorithm's goal is to create a map function that maps input variables (x) to output variables (y).

2.1.4.1 Logistic Regression

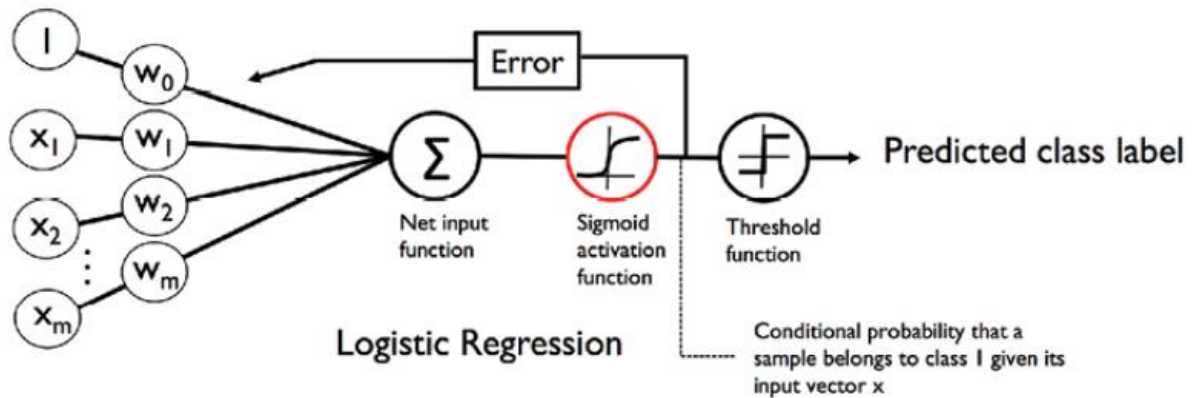


Figure 3: Logistic Regression, Source: Data Analytics

Logistic regression is an algorithm of a supervised learning phase used to predict the probability of targeted variability. The nature of the target or dependent variable is dichotomous, which means that there will be only two possible phases.

In basic terms, the dependent variation is either 1 (representing success / yes) or 0 (representing failure / no)

The asset regression model predicts $P(Y = 1)$ as the X function statistically. It's one of the most basic machine learning algorithms, and it can be used to solve a variety of classification problems like spam identification, diabetes prognosis, cancer detection, and so on [20].

2.1.4.2 Random Forest

Random forests, also known as randomly forested forests, are an integrated learning strategy for segregation, retrieval, and other activities that involves the training of a large number of deciduous trees. Random forest clearing is a class picked by many trees during segregation processes. The average rate or projection for each tree is reversed with retrospective procedures. In their training set, random decision-making forests correct the practise of overcrowded deciduous trees. 587–588 Although random forests frequently make outstanding decisions, their accuracy lags behind that of advanced trees. However, data characteristics can have an impact on their performance.

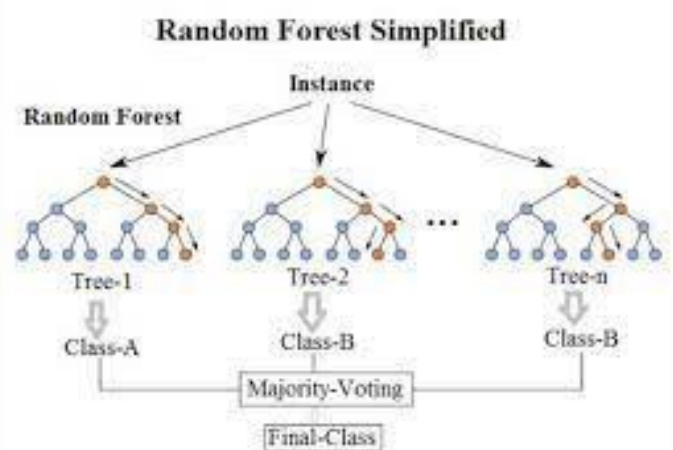


Figure 4: Random Forest, Source: Wikipedia

2.1.4.3 K-Nearest neighbour

The KNN algorithm is a sort of supervised machine learning technique that can be utilised for both editing and retrospective tasks. However, it is mostly utilised in the industry to discern between anticipated problems. The K-nearest neighbours (KNN) algorithm predicts new data point values based on 'feature similarity,' which implies that a new data point will be assigned a value based on how closely the points in the training set are related.

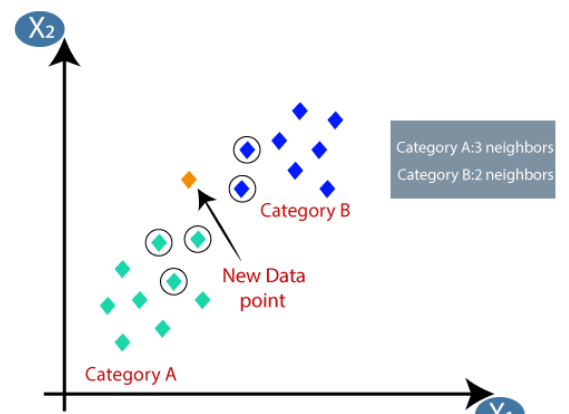


Figure 5: KNN, Source: Wikipedia

2.1.4.4 MLP Classifier

A feedforward Artificial Neural Network of the Multilayer Perceptron (MLP) class (ANN). MLP is a word that is sometimes loosely used to any ANN feedforward, and other times exclusively applied to networks made up of multiple layers of perceptron (with threshold activation) When there is only one hidden layer in a multilayer perceptron, it is frequently referred to as a "vanilla" sensory network. There are at least three layers in the MLP: i.e., 1. Input layer, 2. Hidden, and 3. Output layer. Each node is a neuron that uses the indirect opening function in addition to the input nodes..

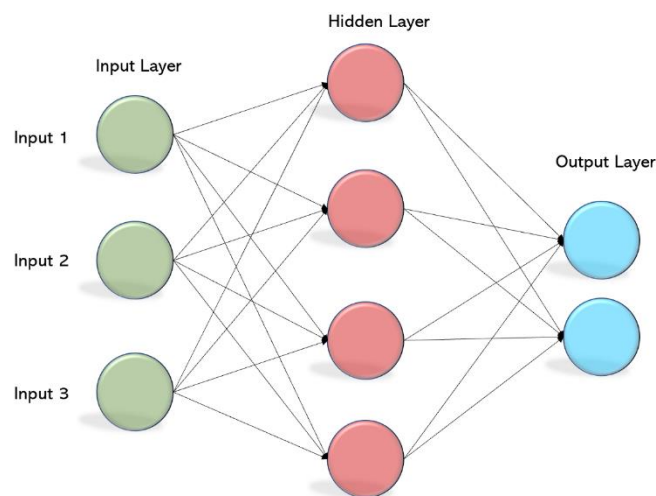


Figure 6: MLP, Source: becominghuman.ai

2.1.4.5 SVM

SVM, or Support Vector Machine, is a common and popular Supervised Learning technique for classification and regression problems. However, in Machine Learning, it is mostly utilised to solve classification difficulties.

The SVM algorithm's purpose is to produce a better line or decision line that can divide n-dimensional space into classes so that a data point can be conveniently placed in the proper category in the future. The hyperplane is the optimal decision-making limit.

SVM chooses the points that help in the formation of the hyperplane. Supporting vectors are the extreme instances, which is why the technique is termed as the Vector Support Machine. Take a look at the diagram below, which shows two distinct categories separated by resolution or hyperplane.

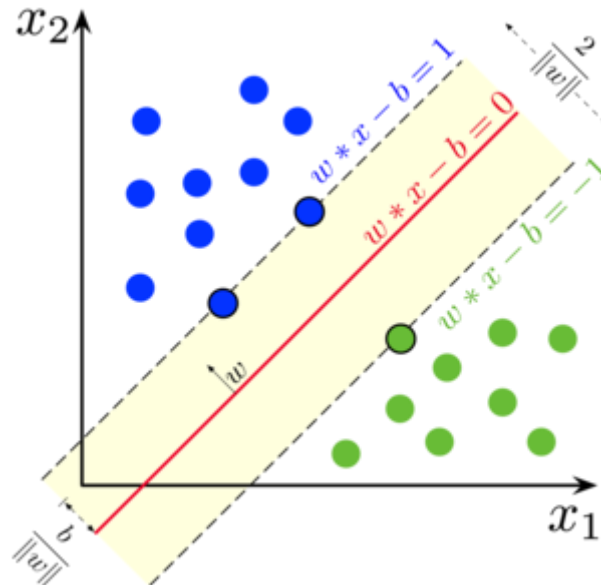


Figure 7: SVM, Source: Wikipedia

2.1.5 Semi-Supervised Learning:

Supervised Learning is a machine learning method that combines the labelled data during training. Slight supervised learning falls between supervised Learning (without labelled data, training data) and supervised reading (with labelled training data only). It is a special occasion for weak surveillance.

When data without label is used with a small amount of labelled data, it can produce significant improvements in reading accuracy. Obtaining data with a reading problem label usually requires a competent agent (e.g., to record audio part) or physical examination (e.g., to specify a 3D protein structure or to determine if there is oil somewhere). The costs associated with the labelling process may make large, well-labelled training sets impossible, while unregistered data acquisition costs more. In such cases, reading with little supervision can be very helpful.

While classifying the normal and cancerous some tuples get misclassified and those errors grouped into misclassification. And the Misclassification rate can be classified into the false positive rate and the false negative rate [18]. With varied asymmetric weights for misclassification mistakes, a ROC curve (Receiver Operating characteristics curve) can be used to estimate the "power" of a classification technique.

Chapter-3: Methodology

3.1 NIH: GDC Portal:

3.1.1 The National Cancer Institute:

The National Cancer Institute (NCI) is a prominent cancer research organisation dedicated to accelerating scientific discovery and lowering cancer's global impact in the United States and around the world. The National Cancer Institute is the primary federal institution in charge of cancer research and education. Its research focuses on a wide range of topics, from working on the basic science to epidemiology to the clinical trials. The National Cancer Institute's budget is utilised to support research at universities by supporting the student in the research field and research centres across the country.

The NCI supports a network of 71 NCI-designated Cancer Centres with a dedicated focus on cancer research and treatment and is also responsible for maintaining the National Clinical Trials Network.

3.1.2 GDC (Genomic Data Commons) Data Portal:

The GDC Data Portal is a sophisticated data-driven platform that enables to use modern web technologies to search and download cancer disease (benign and malignant tumour) data for research..

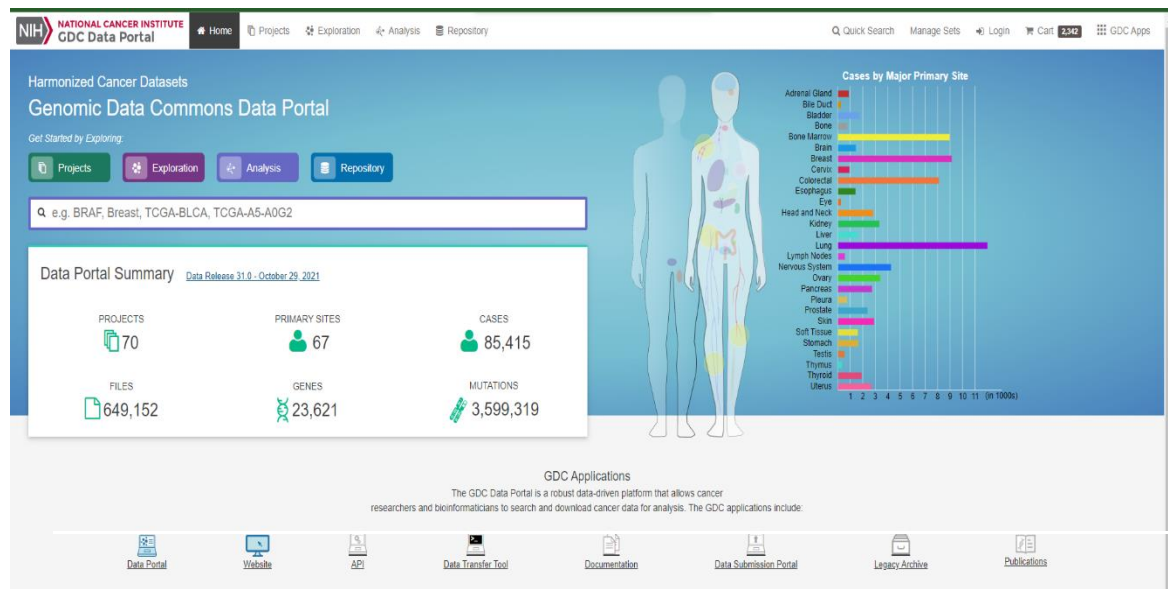


Figure 8: GDC Data Portal, Source: GDC Data Portal

Key features of the GDC Data Portal are as follows.

2. Browse harmonized data

The GDC portal allows users to gather details regarding specific aspects of the GDC harmonized data like related projects, files, annotations, participants, etc. It enables users to navigate between results and pages in an efficient manner by allowing them to add more elements to the personalized cart while browsing through the portal.

3. Access legacy archives

The GDC Data Portal provides access to harmonised GDC data as well as a historical data repository from TCGA and other NCI studies. It enables users to access the Legacy Archive Portal, which allows them to search for and retrieve legacy files.

4. Visually identify data

The visualization feature on the GDC data portal allows users to find the data that they are looking for to narrow down their research. The visualizations in GDC Data Portal include various types of tables, charts, and plots.

5. Search data with filters

The users can easily look for specific data from the huge pool of datasets in the GDC dataset portal. In order to enhance their search, GDC Portal uses predetermined filters called facets and smart search technology.

6. Create personalized cart

GDC allows users to add selected files to their personalized cart. The cart also provides detailed statistics about the included files for a better understanding of the data in the cart.

7. Download data quickly and securely

GDC data Portal enable users with 2 options to download the data, i.e. either directly from the browser or from the GDC Data Transfer tool.[15]

3.2 Dataset Preparation:

The preliminary step in the **KDD** (Knowledge Discovery of Data) Process is to obtain the relevant data in order to identify and extract critical **patterns** in the huge dataset. The data to be processed should be in a consistent state because the dataset is the evidence base for building data mining models. If some significant attributes are missing, then the entire study may be unsuccessful.

So, it is important to convert the “raw” data into a clean and efficient dataset using the techniques of **Pre-processing and cleansing**. Here, **data reliability** is improved which includes handling missing quantities and removing noise and outliers. In the next stage, the appropriate dataset is created for data mining, this is referred to as **Data Transformation**. Now, the data is completely transformed from a raw state to a processed dataset ready for application of data mining and machine learning algorithms.

Steps for Dataset preparation

Data Collection – The very first step in dataset preparation is identifying the relevant data for the project and then collecting and preparing the sample data.

Data Pre-processing – After preparing the sample data, it is exposed to various data pre-processing techniques to convert it into an efficient dataset that could be fed to an ML model. The data pre-processing techniques are:

- **Data cleaning** – Dealing with noisy, inconsistent, and missing values.
- **Data integration** – The data is structured in order to eliminate data redundancy and increase data integrity; the data is organised according to a sequence of normal forms.
- **Data reduction** – This allows us to have a condensed representation of the data set that is less in size but yet maintains its integrity.
- **Low variance** Attributes having variance less than the threshold is removed.
- **Data transformation** – This transforms the data into an appropriate form for Data Modelling.

Aggregation – Aggregation operations are applied on the dataset to come up with new attributes.

Normalization – The data is normalised i.e. scaled between a smaller range.

Discretization – In this, raw values are replaced by discrete intervals.

3.3 Methodology:

- ❖ First, we explored the GDC Data Portal to find RNA and gene expression data pertaining to different types of cancers.
- ❖ Using the smart search technology of the GDC Data Portal, we were able to retrieve the relevant data by applying certain filters in the data repository.
- ❖ The filters that we applied are as follows,
 - **Data Category – Transcriptome profiling**

It's a tool for identifying genes that have comparable expression patterns. It's utilised to figure out how a cell type's genetic control works. A transcriptome is a collection of messenger ribonucleic acid (RNA) molecules that an organism expresses. It refers to the small percentage of genetic material that is transcribed into RNA molecules, and it provides details on the important biological activities that keep the cell functioning.
 - **Data Type – Gene Expression Quantification**

Gene expression can be defined as the conversion of genetic information into RNA or protein and it helps us to use the information encoded in the genes. By linking the expression of certain genes to biological processes helps in understanding the gene function.

Gene Expression Value gives us the RNA-Seq Count Value, i.e. number of RNA transcripts that make the corresponding protein.
 - **Experimental Strategy – RNA-Seq**

It is a sequencing method that is used to determine the gene expression levels and further generate a gene expression profile for tumor samples of different cancer types to decide which gene expression levels are responsible for tumor development.
 - **Workflow Type – HTSeq FPKM**

It's a straightforward way for normalising expression levels. The read count is normalised depending on the length of the gene and the total number of mapped reads.

The following is the formula for calculating FPKM values:

$$FPKM = [RMg * 10^9] / [RMt * L]$$

- RMg: The number of reads mapped to the gene
- RMt: The total number of reads mapped to protein-coding sequences in the alignment.
- L: The length of the gene in base pairs [13] L: The length of the gene in base pairs [13]

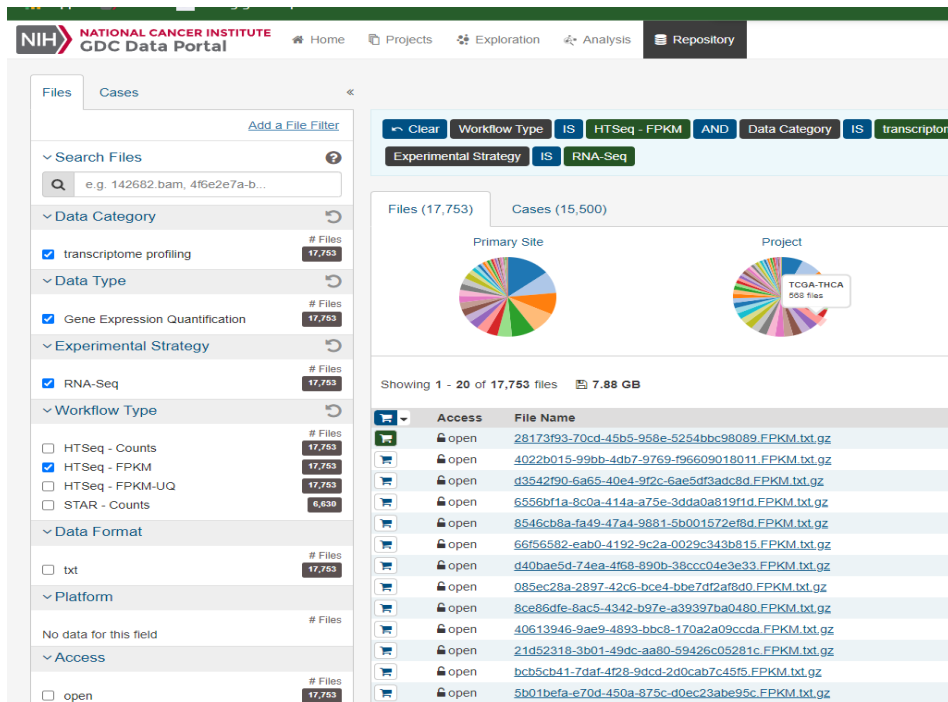


Figure 9: Data Filters, Source: GDC Data Portal

- ❖ After applying the relevant filters and selecting the primary site of cancer, we added the files to the cart and downloaded the ‘Manifest’ for the same.

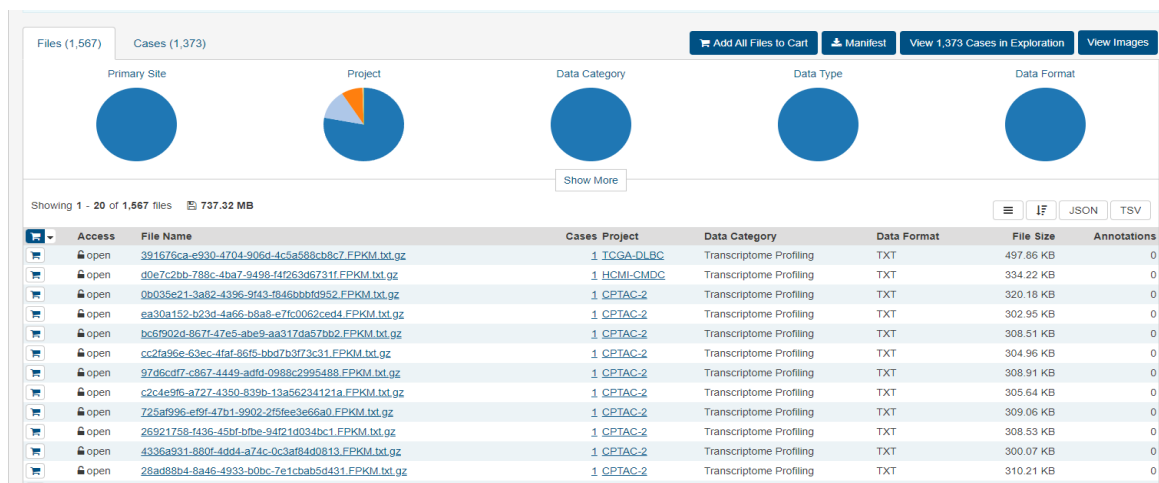


Figure 10: Data Cart, Source: GDC Data Portal

- ❖ The manifest file was then uploaded on the GDC Data Transfer Tool which we had downloaded from the GDC site.

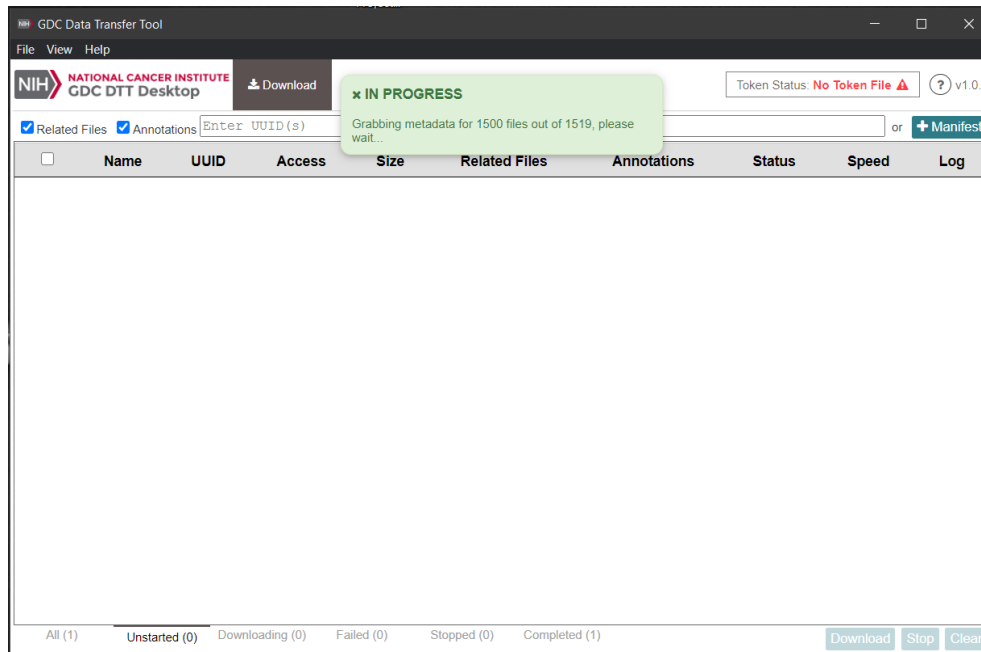


Figure 11: GDC Data Transfer Tool

- ❖ After downloading we got FPKM files as tab-delimited files with Ensemble gene IDs in the first column and expression values in the second column after downloading the files via the GDC Data Transfer Tool.
- ❖ The same procedure was repeated for downloading the FPKM files for all the 5 cancer types (Brain, Kidney, Breast, Liver, and Stomach)
- ❖ The FPKM files for the 5 cancer types were then later converted to .csv format on Microsoft Excel for data manipulation and cleaning.

```
File Edit Format View Help
ENSG00000242268.2      1.52310159826
ENSG00000270112.3      0.426897070983
ENSG00000167578.15     3.76307067468
ENSG00000273842.1      0.332739118388
ENSG00000078237.5      5.34440883008
ENSG00000146083.10     16.5343711559
ENSG00000225275.4      0.0
ENSG00000158486.12     0.00867434867456
ENSG00000198242.12     144.411605192
ENSG00000259883.1      0.168311616305
ENSG00000231981.3      0.0
ENSG00000269475.2      0.0
ENSG00000201788.1      0.0
ENSG00000134108.11     33.0408269698
ENSG00000263089.1      0.0151563018187
ENSG00000172137.17     1.54939589478
ENSG00000167700.7      7.78277345478
ENSG00000234943.2      0.0
ENSG00000240423.1      0.153997139444
ENSG00000060642.9      2.46209238251
ENSG00000271616.1      0.0
ENSG00000234881.1      0.0
ENSG00000236040.1      0.0
ENSG00000231105.1      0.0523470174148
ENSG00000243044.1      0.0
ENSG00000182141.8      3.91015461157
ENSG00000269416.4      1.29925741919
ENSG00000264981.1      0.0
ENSG00000272065.1      0.0056602105205
```

Figure 12: FPKM file in .txt format

Figure 13: Dataset before cleaning and integration

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	BRAIN			KIDNEY			BREAST			liver and intrahepatic bile ducts			Stomach	
2	ENSG00000242268.2	1.523101598		ENSG00000000003.13	19.8286		ENSG00000242268.2	0		ENSG00000242268.2	0.061253732		ENSG00000242268.2	0
3	ENSG00000270112.3	0.426897071		ENSG00000000005.5	0.1751		ENSG00000270112.3	0.007055003		ENSG00000270112.3	0		ENSG00000270112.3	0.00817977
4	ENSG00000167578.15	3.763070675		ENSG00000000419.11	33.6306		ENSG00000167578.15	2.778368413		ENSG00000167578.15	3.262405588		ENSG00000167578.15	6.203430605
5	ENSG00000273842.1	0.332739118		ENSG00000000457.12	3.5494		ENSG00000273842.1	0		ENSG00000273842.1	0		ENSG00000273842.1	0
6	ENSG00000078237.5	5.34440883		ENSG00000000460.15	1.7322		ENSG00000078237.5	4.167733097		ENSG00000078237.5	1.227641224		ENSG00000078237.5	4.216048833
7	ENSG00000146083.10	16.53437116		ENSG00000000938.11	3.3539		ENSG00000146083.10	14.29348218		ENSG00000146083.10	2.763596657		ENSG00000146083.10	19.46372395
8	ENSG00000225275.4	0		ENSG00000000971.14	56.9317		ENSG00000225275.4	0.064007891		ENSG00000225275.4	0		ENSG00000225275.4	0
9	ENSG00000158486.12	0.008674349		ENSG00000001036.12	23.0282		ENSG00000158486.12	0.016891922		ENSG00000158486.12	0		ENSG00000158486.12	0.076940957
10	ENSG00000198242.12	144.4116052		ENSG00000001084.9	5.6486		ENSG00000198242.12	123.76932		ENSG00000198242.12	113.0220279		ENSG00000198242.12	353.4035466
11	ENSG00000259883.1	0.168311616		ENSG00000001167.13	5.5291		ENSG00000259883.1	0		ENSG00000259883.1	0.112814914		ENSG00000259883.1	0.135719457
12	ENSG00000231981.3	0		ENSG00000001460.16	1.9928		ENSG00000231981.3	0.049520508		ENSG00000231981.3	0		ENSG00000231981.3	0.172246422
13	ENSG00000269475.2	0		ENSG00000001461.15	29.0712		ENSG00000269475.2	0		ENSG00000269475.2	0		ENSG00000269475.2	0
14	ENSG00000201788.1	0		ENSG00000001497.15	3.9256		ENSG00000201788.1	0		ENSG00000201788.1	0		ENSG00000201788.1	0
15	ENSG00000134108.11	33.04082697		ENSG00000001561.6	12.3742		ENSG00000134108.11	38.92205988		ENSG00000134108.11	12.68868081		ENSG00000134108.11	15.17509763
16	ENSG00000263089.1	0.015156302		ENSG00000001617.10	3.9914		ENSG00000263089.1	0.042163572		ENSG00000263089.1	0		ENSG00000263089.1	0.136879778
17	ENSG00000172137.17	1.549395895		ENSG00000001626.13	0.2262		ENSG00000172137.17	0.794930644		ENSG00000172137.17	0.008512458		ENSG00000172137.17	0.946242748
18	ENSG00000167700.7	7.782773455		ENSG00000001629.8	11.6663		ENSG00000167700.7	7.841897048		ENSG00000167700.7	21.38325322		ENSG00000167700.7	20.27358339
19	ENSG00000234943.2	0		ENSG00000001630.14	0.5478		ENSG00000234943.2	0		ENSG00000234943.2	0.076905658		ENSG00000234943.2	0.277558889
20	ENSG00000240423.1	0.153997139		ENSG00000001631.13	7.9401		ENSG00000240423.1	0		ENSG00000240423.1	0		ENSG00000240423.1	0.016556915
21	ENSG00000060642.9	2.462092383		ENSG00000002016.15	4.6645		ENSG00000060642.9	4.456516262		ENSG00000060642.9	7.852745296		ENSG00000060642.9	4.145984618
22	ENSG00000271616.1	0		ENSG00000002079.11	0.0259		ENSG00000271616.1	0		ENSG00000271616.1	0		ENSG00000271616.1	0
23	ENSG00000234881.1	0		ENSG00000002330.12	16.6279		ENSG00000234881.1	0		ENSG00000234881.1	0		ENSG00000234881.1	0
24	ENSG00000236040.1	0		ENSG00000002549.11	39.2113		ENSG00000236040.1	0		ENSG00000236040.1	0		ENSG00000236040.1	0.021033106
25	ENSG00000231105.1	0.052347017		ENSG00000002586.16	24.3206		ENSG00000231105.1	0.072812526		ENSG00000231105.1	0		ENSG00000231105.1	0.514967469
26	ENSG00000243044.1	0		ENSG00000002587.8	0.8026		ENSG00000243044.1	0		ENSG00000243044.1	0		ENSG00000243044.1	0
27	ENSG00000182141.8	3.910154612		ENSG00000002726.18	88.6557		ENSG00000182141.8	3.274243827		ENSG00000182141.8	0.204755882		ENSG00000182141.8	1.203483248
28	ENSG00000269416.4	1.299257419		ENSG00000002745.11	0.2089		ENSG00000269416.4	0.043811226		ENSG00000269416.4	0.030787929		ENSG00000269416.4	1.866751953
29	ENSG00000264981.1	0		ENSG00000002746.13	0.4542		ENSG00000264981.1	0		ENSG00000264981.1	0		ENSG00000264981.1	0

- ❖ There were certain Ensemble IDs in the dataset which had gene expression values equal to zero, so we removed the rows from each cancer type having zero gene expression values so that there is no redundancy in the final dataset.
- ❖ As the number of observations (rows) for each cancer type was not the same, we performed data reduction and selected 2500 instances of each cancer so that there is uniformity in the data for each and every cancer type.

	A	B	C	D	E	F	G	H	I	J	K				
	BRAIN			KIDNEY			BREAST			Liver and intrahepatic bile ducts			Stomach		
1	Gene name	Gene Classification	Valu	Gene name	Gene Classification	Valu	Gene name	Gene Classification	Valu	Gene name	Gene Classification	Valu	Gene name	Gene Classification	Valu
2	ENSG00000242268.2	1.523101598	ENSG000000000003.1	19.8286	ENSG00000270112.3	0.007055003	ENSG00000242268.2	0.061253732	ENSG00000270112.3	0.008179777					
3	ENSG00000270112.3	0.426897071	ENSG000000000005.5	0.1751	ENSG00000167578.1	2.778368413	ENSG00000167578.1	3.262405588	ENSG00000167578.1	6.203430605					
4	ENSG00000167578.1	3.763070675	ENSG0000000000419.1	33.6306	ENSG00000078237.5	4.167733097	ENSG00000078237.5	1.227641224	ENSG00000078237.5	4.216048833					
5	ENSG00000273842.1	0.332739118	ENSG0000000000457.1	3.5494	ENSG00000146083.1	14.29348218	ENSG00000146083.1	2.763596657	ENSG00000146083.1	19.46372395					
6	ENSG00000078237.5	5.34440883	ENSG0000000000460.1	1.7322	ENSG00000225275.4	0.064007891	ENSG00000198242.1	113.0220279	ENSG00000158486.1	0.076940957					
7	ENSG00000146083.1	16.53437116	ENSG0000000000938.1	3.3539	ENSG00000158486.1	0.016891922	ENSG00000259883.1	0.112814914	ENSG00000198242.1	353.4035466					
8	ENSG00000158486.1	0.008674349	ENSG0000000000971.1	56.9317	ENSG00000198242.1	123.76932	ENSG00000134108.1	12.68868081	ENSG00000259883.1	0.135719457					
9	ENSG00000198242.1	144.4116052	ENSG000000001036.1	23.0282	ENSG00000231981.3	0.049520508	ENSG00000172137.1	0.008512458	ENSG00000231981.3	0.172246422					
10	ENSG00000259883.1	0.168311616	ENSG000000001084.9	5.6486	ENSG00000134108.1	38.92205988	ENSG00000167700.7	21.38325322	ENSG00000134108.1	15.17509763					
11	ENSG00000134108.1	33.04082697	ENSG000000001167.1	5.5291	ENSG00000263089.1	0.042163572	ENSG00000234943.2	0.076905658	ENSG00000263089.1	0.136879778					
12	ENSG00000263089.1	0.015156302	ENSG000000001460.1	1.9928	ENSG00000172137.1	0.794930644	ENSG00000060642.9	7.852745296	ENSG00000172137.1	0.946242748					
13	ENSG00000172137.1	1.549395895	ENSG000000001461.1	29.0712	ENSG00000167700.7	7.841897048	ENSG00000182141.8	0.204755882	ENSG00000167700.7	20.27358339					
14	ENSG00000167700.7	7.782773455	ENSG000000001497.1	3.9256	ENSG00000060642.9	4.456516262	ENSG00000269416.4	0.030787929	ENSG00000234943.2	0.277558889					
15	ENSG00000240423.1	0.153997139	ENSG000000001561.6	12.3742	ENSG00000231105.1	0.072812526	ENSG00000185105.4	0.005897769	ENSG00000240423.1	0.016556915					
16	ENSG00000060642.9	2.462092383	ENSG000000001617.1	3.9914	ENSG00000182141.8	3.274243827	ENSG00000102174.8	0.131705577	ENSG00000060642.9	4.145984618					
17	ENSG00000231105.1	0.052347017	ENSG000000001626.1	0.2262	ENSG00000269416.4	0.043811226	ENSG00000166391.1	6.896360171	ENSG00000236040.1	0.021033106					
18	ENSG00000182141.8	3.910154612	ENSG000000001639.8	11.6663	ENSG00000275265.1	0.039670825	ENSG00000070087.1	1.698307192	ENSG00000231105.1	0.514967469					
19	ENSG00000269416.4	1.299257419	ENSG000000001630.1	0.5478	ENSG00000185105.4	0.029373841	ENSG00000153561.1	8.320058473	ENSG00000182141.8	1.203483248					
20	ENSG00000275265.1	0.09506832	ENSG000000001631.1	7.9401	ENSG00000233540.1	0.091162753	ENSG00000269416.4	0.056924279	ENSG00000269416.4	1.866751953					
21	ENSG00000185105.4	0.149583696	ENSG000000002016.1	4.6645	ENSG00000102174.8	0.161162725	ENSG00000179262.8	29.98789942	ENSG00000185105.4	0.919535211					
22	ENSG00000102174.8	1.206075063	ENSG000000002079.1	0.0259	ENSG00000166391.1	0.715205829	ENSG00000214198.6	0.050311999	ENSG00000102174.8	0.173743865					
23	ENSG00000271647.1	0.015470703	ENSG000000002330.1	16.6279	ENSG00000270469.1	0.044513502	ENSG00000258630.1	0.034089868	ENSG00000166391.1	13.02497563					
24	ENSG00000166391.1	0.005225427	ENSG000000002549.1	39.2113	ENSG00000070087.1	0.918794758	ENSG00000127511.8	1.603772553	ENSG00000270469.1	0.025805105					
25	ENSG00000270469.1	0.133341817	ENSG000000002586.1	24.3206	ENSG00000266261.1	0.063223205	ENSG00000095587.8	0.01244964	ENSG00000070087.1	5.359742842					
26	ENSG00000070087.1	181.7635797	ENSG000000002587.8	0.8026	ENSG00000153561.1	15.64671192	ENSG00000064601.15	75.03247842	ENSG00000266261.1	0.036651384					
27	ENSG00000280038.1	0.006456132	ENSG000000002726.1	88.6557	ENSG00000269148.1	0.118129746	ENSG0000027766.1	11.95650634	ENSG00000153561.1	5.398079831					
28	ENSG00000266261.1	0.719672174	ENSG000000002745.1	0.2089	ENSG00000273406.1	0.037920641	ENSG0000008517.15	319.2575414	ENSG00000269148.1	0.41088888					
29	ENSG00000153561.1	9.437633563	ENSG000000002746.1	0.4542	ENSG00000179262.8	34.30206118	ENSG00000215246.5	0.420496371	ENSG00000179262.8	40.19805268					
30	ENSG00000269148.1	0.452943303	ENSG000000002822.1	2.687	ENSG00000214198.6	0.182239219	ENSG00000236893.1	0.094595214	ENSG00000214198.6	0.033014577					
31	ENSG00000179262.8	42.51424971	ENSG000000002834.1	19.1559	ENSG00000258630.1	0.04244615	ENSG00000088179.7	0.469059223	ENSG00000278099.1	0.126836002					
32	ENSG00000214198.6	0.081885633	ENSG000000002919.1	8.2342	ENSG00000127511.8	5.015893715	ENSG00000070081.14	7.35714405	ENSG00000166368.2	0.0377008					
33	ENSG00000234900.1	0.029668097	ENSG000000002933.6	42.0323	ENSG00000225269.2	0.020687937	ENSG00000275479.1	0.252849706	ENSG00000234900.1	0.086123211					
34	ENSG00000127511.8	11.27358953	ENSG000000003056.6	43.8346	ENSG00000214062.5	0.02493125	ENSG00000261789.1	0.008722453	ENSG00000258630.1	0.012303316					
35	ENSG00000225269.2	0.012394294	ENSG000000003096.1	1.5306	ENSG00000095587.8	0.057576476	ENSG00000161558.9	1.077167531	ENSG00000127511.8	5.016792335					
36	ENSG00000214062.5	0.17937923	ENSG000000003137.7	0.4637	ENSG00000274219.1	0.019879103	ENSG00000276644.3	0.010760227	ENSG00000095587.8	0.06162073					
37	ENSG00000095587.8	0.275955888	ENSG000000003147.1	3.1363	ENSG00000064601.1	15.77545292	ENSG00000196167.8	0.069719695	ENSG00000064601.1	55.07164609					
38	ENSG00000064601.1	22.23743939	ENSG000000003249.1	1.417	ENSG0000027766.1	3.326476538	ENSG00000179833.3	2.798138932	ENSG0000027766.1	4.93165964					
39	ENSG0000027766.1	2.330514168	ENSG000000003393.1	3.1603	ENSG00000008517.1	2.266824785	ENSG00000064225.11	2.517559541	ENSG00000008517.1	68.59460135					
40	ENSG00000008517.1	3.201408626	ENSG000000003400.1	3.4163	ENSG00000215246.5	0.020735468	ENSG00000057294.12	0.953000104	ENSG00000215246.5	0.282485121					
41															

Figure 14: Dataset after data cleaning and integration

- ❖ For this dataset to be used for creating a Machine Learning model, it was imperative to represent different cancer types in encoded form. So, each cancer type was binary encoded.
- ❖ The **binary encoding** of various cancer types is as follows,
 - Brain – [1,0,0,0,0]
 - Kidney – [0,1,0,0,0]
 - Breast – [0,0,1,0,0]
 - Liver – [0,0,0,1,0]
 - Stomach – [0,0,0,0,1]

	A	B	C	D	E	F	G	H
1	Ensemble_ID	Gene_exp_val	Brain	Kidney	Breast	Liver	Stomach	
2	ENSG00000242268.2	1.523101598	1	0	0	0	0	
3	ENSG00000270112.3	0.426897071	1	0	0	0	0	
4	ENSG00000167578.15	3.763070675	1	0	0	0	0	
5	ENSG00000273842.1	0.332739118	1	0	0	0	0	
6	ENSG00000078237.5	5.34440883	1	0	0	0	0	
7	ENSG00000146083.10	16.53437116	1	0	0	0	0	
8	ENSG00000158486.12	0.008674349	1	0	0	0	0	
9	ENSG00000198242.12	144.4116052	1	0	0	0	0	
10	ENSG00000259883.1	0.168311616	1	0	0	0	0	
11	ENSG00000134108.11	33.04082697	1	0	0	0	0	
12	ENSG00000263089.1	0.015156302	1	0	0	0	0	
13	ENSG00000172137.17	1.549395895	1	0	0	0	0	
14	ENSG00000167700.7	7.782773455	1	0	0	0	0	
15	ENSG00000240423.1	0.153997139	1	0	0	0	0	
16	ENSG00000060642.9	2.462092383	1	0	0	0	0	
17	ENSG00000231105.1	0.052347017	1	0	0	0	0	
18	ENSG00000182141.8	3.910154612	1	0	0	0	0	
19	ENSG00000269416.4	1.299257419	1	0	0	0	0	
20	ENSG00000275265.1	0.09506832	1	0	0	0	0	
21	ENSG00000185105.4	0.149583696	1	0	0	0	0	
22	ENSG00000102174.8	1.206075063	1	0	0	0	0	
23	ENSG00000271647.1	0.015470703	1	0	0	0	0	
24	ENSG00000166391.13	0.005225427	1	0	0	0	0	
25	ENSG00000270469.1	0.133341817	1	0	0	0	0	
26	ENSG00000070087.12	181.7635797	1	0	0	0	0	
27	ENSG00000280038.1	0.006456132	1	0	0	0	0	
28	ENSG00000266261.1	0.719672174	1	0	0	0	0	
29	ENSG00000153561.11	9.437633563	1	0	0	0	0	

Figure 15: Dataset with binary encoded values associated with different cancers

❖ Data Normalization

It was found that the attribute values for 'Gen_exp_val' were on different scales, so we applied **Min-Max normalization** on the 'Gen_exp_val' column in order to create a better data model using the following formula,

$$\mathbf{X(normalized)} = \mathbf{(X - X(min)) / (X(max) - X(min))}$$

	A	B	C	D	E	F	G	H	I	J	K
1	Ensemble_ID	Gene_exp_val	Brain	Kidney	Breast	Liver	Stomach	Norml_gene_exp_val	Max	Min	
2	ENSG00000242268.2	1.523101598	1	0	0	0	0	0.00019908	7647.17	0.000705	
3	ENSG00000270112.3	0.426897071	1	0	0	0	0	5.5732E-05			
4	ENSG00000167578.15	3.763070675	1	0	0	0	0	0.000491995			
5	ENSG00000273842.1	0.332739118	1	0	0	0	0	4.34192E-05			
6	ENSG00000078237.5	5.34440883	1	0	0	0	0	0.000698782			
7	ENSG00000146083.10	16.53437116	1	0	0	0	0	0.002162064			
8	ENSG00000158486.12	0.008674349	1	0	0	0	0	1.04215E-06			
9	ENSG00000198242.12	144.4116052	1	0	0	0	0	0.018884229			
10	ENSG00000259883.1	0.168311616	1	0	0	0	0	2.19175E-05			
11	ENSG00000134108.11	33.04082697	1	0	0	0	0	0.004320569			
12	ENSG00000263089.1	0.015156302	1	0	0	0	0	1.88978E-06			
13	ENSG00000172137.17	1.549395895	1	0	0	0	0	0.000202518			
14	ENSG00000167700.7	7.782773455	1	0	0	0	0	0.00101764			
15	ENSG00000240423.1	0.153997139	1	0	0	0	0	2.00456E-05			
16	ENSG00000060642.9	2.462092383	1	0	0	0	0	0.000321869			
17	ENSG00000231105.1	0.052347017	1	0	0	0	0	6.75311E-06			
18	ENSG00000182141.8	3.910154612	1	0	0	0	0	0.000511228			
19	ENSG00000269416.4	1.299257419	1	0	0	0	0	0.000169808			
20	ENSG00000275265.1	0.09506832	1	0	0	0	0	1.23397E-05			
21	ENSG00000185105.4	0.149583696	1	0	0	0	0	1.94685E-05			
22	ENSG00000102174.8	1.206075063	1	0	0	0	0	0.000157623			
23	ENSG00000271647.1	0.015470703	1	0	0	0	0	1.93089E-06			
24	ENSG00000166391.13	0.005225427	1	0	0	0	0	5.91143E-07			
25	ENSG00000270469.1	0.133341817	1	0	0	0	0	1.73446E-05			
26	ENSG00000070087.12	181.7635797	1	0	0	0	0	0.023768648			
27	ENSG00000280038.1	0.006456132	1	0	0	0	0	7.52079E-07			
28	ENSG00000266261.1	0.719672174	1	0	0	0	0	9.40174E-05			
29	ENSG00000153561.11	9.437633563	1	0	0	0	0	0.001234042			

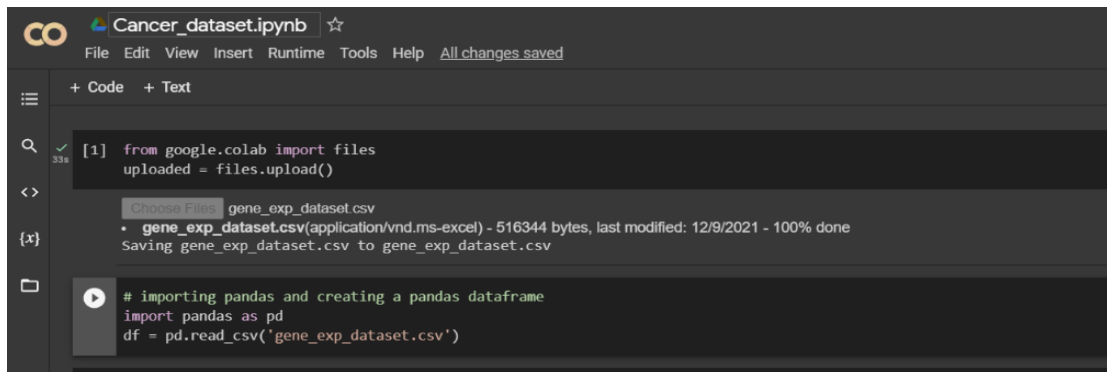
Figure 16: Data Normalization

	A	B	C	D	E	F	G
1	Ensemble_ID	Gene_exp_val	Brain	Kidney	Breast	Liver	Stomach
2	ENSG00000242268.2	0.00019908	1	0	0	0	0
3	ENSG00000270112.3	5.5732E-05	1	0	0	0	0
4	ENSG00000167578.15	0.000491995	1	0	0	0	0
5	ENSG00000273842.1	4.34192E-05	1	0	0	0	0
6	ENSG00000078237.5	0.000698782	1	0	0	0	0
7	ENSG00000146083.10	0.002162064	1	0	0	0	0
8	ENSG00000158486.12	1.04215E-06	1	0	0	0	0
9	ENSG00000198242.12	0.018884229	1	0	0	0	0
10	ENSG00000259883.1	2.19175E-05	1	0	0	0	0
11	ENSG00000134108.11	0.004320569	1	0	0	0	0
12	ENSG00000263089.1	1.88978E-06	1	0	0	0	0
13	ENSG00000172137.17	0.000202518	1	0	0	0	0
14	ENSG00000167700.7	0.00101764	1	0	0	0	0
15	ENSG00000240423.1	2.00456E-05	1	0	0	0	0
16	ENSG00000060642.9	0.000321869	1	0	0	0	0
17	ENSG00000231105.1	6.75311E-06	1	0	0	0	0
18	ENSG00000182141.8	0.000511228	1	0	0	0	0
19	ENSG00000269416.4	0.000169808	1	0	0	0	0
20	ENSG00000275265.1	1.23397E-05	1	0	0	0	0
21	ENSG00000185105.4	1.94685E-05	1	0	0	0	0
22	ENSG00000102174.8	0.000157623	1	0	0	0	0
23	ENSG00000271647.1	1.93089E-06	1	0	0	0	0
24	ENSG00000166391.13	5.91143E-07	1	0	0	0	0
25	ENSG00000270469.1	1.73446E-05	1	0	0	0	0
26	ENSG00000070087.12	0.023768648	1	0	0	0	0
27	ENSG00000280038.1	7.52079E-07	1	0	0	0	0
28	ENSG00000266261.1	9.40174E-05	1	0	0	0	0
29	ENSG00000153561.11	0.001234042	1	0	0	0	0

Figure 17: Dataset with normalized Gene expression values

❖ Creating a pandas data frame

We converted the dataset which was in .csv format into a panda's data frame, which is a 2-D heterogeneous tabular data structure. Using the panda's library of the python programming language, we can perform data manipulation and analysis on any dataset.



```

Cancer_dataset.ipynb
File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

[1] from google.colab import files
    uploaded = files.upload()

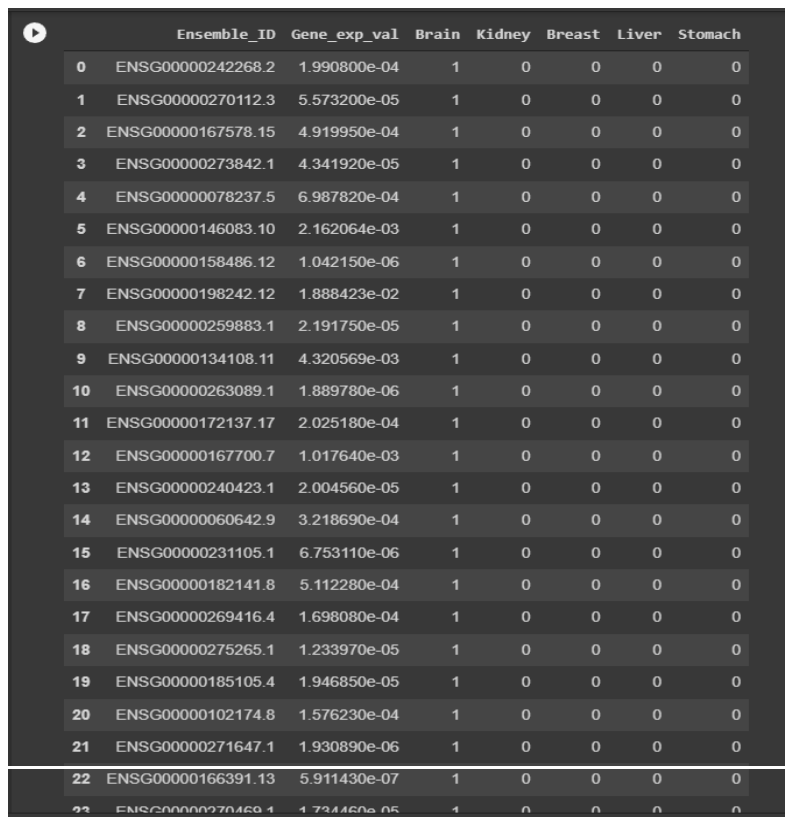
Choose Files: gene_exp_dataset.csv
• gene_exp_dataset.csv(application/vnd.ms-excel) - 516344 bytes, last modified: 12/9/2021 - 100% done
Saving gene_exp_dataset.csv to gene_exp_dataset.csv

# importing pandas and creating a pandas dataframe
import pandas as pd
df = pd.read_csv('gene_exp_dataset.csv')

```

Figure 18: Python script to create a pandas data frame

Output



	Ensemble_ID	Gene_exp_val	Brain	Kidney	Breast	Liver	Stomach
0	ENSG00000242268.2	1.990800e-04	1	0	0	0	0
1	ENSG00000270112.3	5.573200e-05	1	0	0	0	0
2	ENSG00000167578.15	4.919950e-04	1	0	0	0	0
3	ENSG00000273842.1	4.341920e-05	1	0	0	0	0
4	ENSG00000078237.5	6.987820e-04	1	0	0	0	0
5	ENSG00000146083.10	2.162064e-03	1	0	0	0	0
6	ENSG00000158486.12	1.042150e-06	1	0	0	0	0
7	ENSG00000198242.12	1.888423e-02	1	0	0	0	0
8	ENSG00000259883.1	2.191750e-05	1	0	0	0	0
9	ENSG00000134108.11	4.320569e-03	1	0	0	0	0
10	ENSG00000263089.1	1.889780e-06	1	0	0	0	0
11	ENSG00000172137.17	2.025180e-04	1	0	0	0	0
12	ENSG00000167700.7	1.017640e-03	1	0	0	0	0
13	ENSG00000240423.1	2.004560e-05	1	0	0	0	0
14	ENSG00000060642.9	3.218690e-04	1	0	0	0	0
15	ENSG00000231105.1	6.753110e-06	1	0	0	0	0
16	ENSG00000182141.8	5.112280e-04	1	0	0	0	0
17	ENSG00000269416.4	1.698080e-04	1	0	0	0	0
18	ENSG00000275265.1	1.233970e-05	1	0	0	0	0
19	ENSG00000185105.4	1.946850e-05	1	0	0	0	0
20	ENSG00000102174.8	1.576230e-04	1	0	0	0	0
21	ENSG00000271647.1	1.930890e-06	1	0	0	0	0
22	ENSG00000166391.13	5.911430e-07	1	0	0	0	0
23	ENSG00000270469.1	1.734460e-05	1	0	0	0	0

Figure 19: Pandas data frame

3.4 Data Analysis:

- ❖ After visualising the data, we can plot the scatter plot of the gene expression data and can check the data

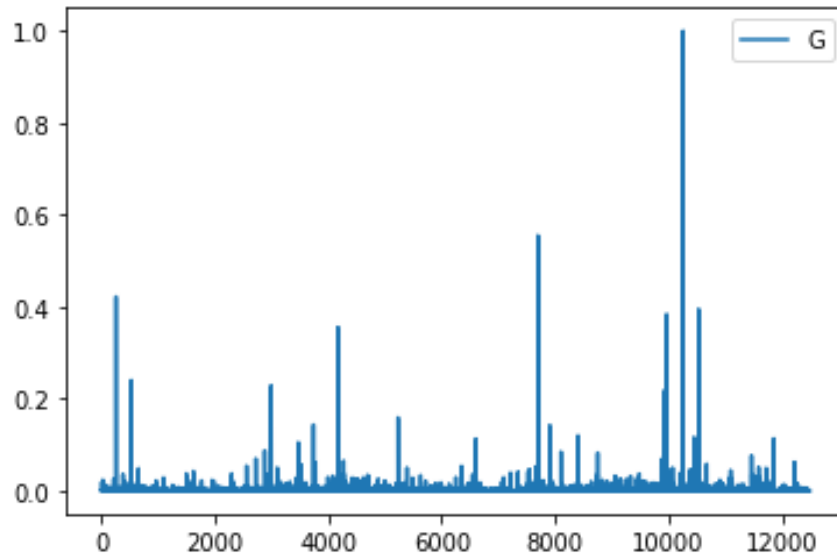


Figure 20 plot of gene expression values

As we can see there are some genes whose expression values are more than the avg. expression values as these can be consider as the important biomarker

- ❖ Now after the analysis of the gene expression data we can find that there are many genes which are responsible for the causing cancer all the five types cancer we took in our project. So, after selecting those genes which are present in all five and four types of cancer

```

ENSG0000000762.1/ 5
ENSG00000044524.9 5
ENSG00000091732.14 5
ENSG00000049239.11 5
ENSG00000096717.10 5
ENSG00000088356.5 5
ENSG00000081853.14 5
ENSG00000079435.8 5
ENSG00000095203.13 5
ENSG00000099849.13 5
ENSG00000069849.9 5
ENSG00000009724.15 5
ENSG00000092529.21 5
ENSG00000074755.13 5
ENSG00000071967.10 5
ENSG00000100628.10 5
ENSG00000079819.15 5
ENSG00000091592.14 5
ENSG00000062524.14 5
ENSG00000034053.13 5
ENSG00000085760.13 5
ENSG00000100354.19 5
ENSG00000146535.12 4
ENSG00000260992.1 4
ENSG00000136021.16 4
ENSG00000142552.6 4
ENSG00000185614.4 4
ENSG00000181649.5 4
ENSG00000132953.15 4
ENSG00000184271.14 4
ENSG00000124596.15 4
ENSG00000144821.8 4
ENSG00000139641.11 4
ENSG00000006377.10 4
ENSG00000252643.1 4
ENSG00000224510.0 4

```

Chapter-4:

Results and Conclusion

In cancer the genes expression data and the clinical data are very useful for the researchers and the scientist to analysis the type of the cancer and the biological function of the genes and also can predict the chance of getting cancer (Malignant and Benign).

After analysing the gene expression, we get the various genes (biomarkers) which are majorly responsible for causing the cancer in Brain, Kidney, Stomach, Breast, Liver

As the gene expression values of the genes is greater than the 0.05 which is much higher than the gene expression values of the other genes, and can be consider as the majorly responsible in causing the cancer.

The genes we get whose value is more than the 0.05 are as follow:

ENSEMBLE_ID	GENE_EXP_VALUES	BRAIN	KIDNEY	BREAST	LIVER	STOMACH
['ENSG00000198888.2', '0.421934507', '1', '0', '0', '0', '0'],						
['ENSG00000211459.2', '0.240198852', '1', '0', '0', '0', '0'],						
['ENSG00000074800.12', '0.143524518', '0', '1', '0', '0', '0'],						
['ENSG00000198888.2', '0.159039148', '0', '0', '1', '0', '0'],						
['ENSG00000149925.15', '0.050142541', '0', '0', '1', '0', '0'],						
['ENSG00000198034.9', '0.053678367', '0', '0', '1', '0', '0'],						
['ENSG00000197956.8', '0.112828986', '0', '0', '1', '0', '0'],						
['ENSG00000227097.5', '0.052429472', '0', '0', '0', '1', '0'],						
['ENSG00000198888.2', '0.555072301', '0', '0', '0', '1', '0'],						
['ENSG00000211459.2', '0.142393782', '0', '0', '0', '1', '0'],						
['ENSG00000134240.10', '0.082487431', '0', '0', '0', '1', '0'],						
['ENSG00000225630.1', '0.050067724', '0', '0', '0', '0', '1'],						
['ENSG00000198888.2', '1', '0', '0', '0', '0', '1'],						
['ENSG00000211679.2', '0.116433674', '0', '0', '0', '0', '1'],						
['ENSG00000211459.2', '0.394978443', '0', '0', '0', '0', '1'],						
['ENSG00000090382.5', '0.058317492', '0', '0', '0', '0', '1'],						
['ENSG00000198034.9', '0.076560264', '0', '0', '0', '0', '1'],						
['ENSG00000134240.10', '0.051111419', '0', '0', '0', '0', '1'],						
['ENSG00000239951.1', '0.113469962', '0', '0', '0', '0', '1'],						
['ENSG00000144713.11', '0.062445755', '0', '0', '0', '0', '1']]						

Now there are many genes are present in multiple cancers so by checking this:

We get the following data:

ENSG00000198888.2	4
ENSG00000211459.2	3
ENSG00000198034.9	2
ENSG00000134240.10	2
ENSG00000090382.5	1
ENSG00000149925.15	1
ENSG00000197956.8	1
ENSG00000225630.1	1
ENSG00000211679.2	1
ENSG00000074800.12	1
ENSG00000227097.5	1
ENSG00000144713.11	1
ENSG00000239951.1	1

Our main biomarker gene is the **ENSG00000198888.2** which is present in Brain, Breast, Liver and stomach

- Gene name of the **ENSG00000198888.2: MT-ND1**

It is the Mitochondrially encoded NADH: ubiquinone oxidoreductase core [Source: HGNC Symbol; Acc: HGNC:7455].

It provides the synthesis for making a protein called NADH dehydrogenase and present in the active mitochondria

Uniport id: P03886

```
[ 'ENSG00000198888.2', '0.421934507', '1', '0', '0', '0', '0' ]
[ 'ENSG00000198888.2', '0.159039148', '0', '0', '1', '0', '0' ]
[ 'ENSG00000198888.2', '0.555072301', '0', '0', '0', '1', '0' ]
[ 'ENSG00000198888.2', '1', '0', '0', '0', '0', '1' ]
```

- Gene name of the **ENSG00000211459.2: MT-RNR1**

It is a Mitochondrially encoded 12s rRNA. [Source: HGNC Symbol; Acc: HGNC:7470]

12s rRNA occupied the 1/16 of the entire mitochondrial genome

We found this gene in Brain, Liver and stomach

```
[ 'ENSG00000211459.2', '0.240198852', '1', '0', '0', '0', '0' ]
[ 'ENSG00000211459.2', '0.142393782', '0', '0', '0', '1', '0' ]
[ 'ENSG00000211459.2', '0.394978443', '0', '0', '0', '0', '1' ]
```

- Gene name of the **ENSG00000198034.9: RPS4X**

It is a ribosomal protein S4 x-linked , it is a protein coding gene

Cytoplasmic ribosomes catalyse the protein synthesis and consist of small 40s and large 60s subunits.

Gene ontology of this gene includes RNA Binding and Structural constituent of ribosomes.

```
[ 'ENSG00000198034.9', '0.053678367', '0', '0', '1', '0', '0' ]
[ 'ENSG00000198034.9', '0.076560264', '0', '0', '0', '0', '1' ]
```

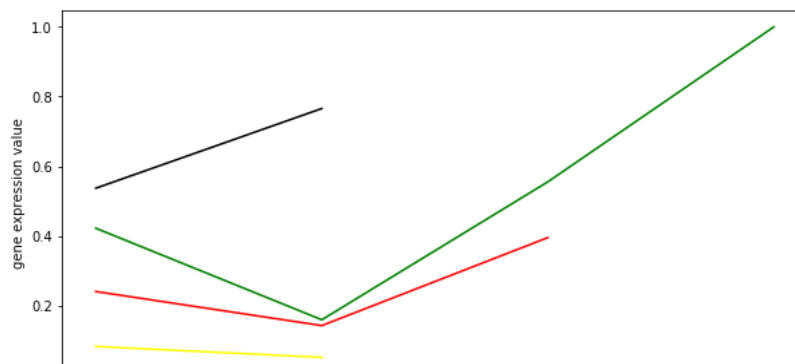


Figure 21: plot of the gene expression values of the biomarker genes

We took another data base based on the data from (source: <https://www.kaggle.com/datasets?search=cancer>) based on the image on the grey scale which consist of radius, texture, perimeter, area, smoothness, compactness, concavity, concave point, symmetry, fractal dimension.

also these all are divided into the mean, standard error and worst

As the data set distribution consist of the 357 benign and 212 malignant

As we applied the various machine learning algorithm on that database and classified the data into the malignant and benign

We applied

- Random Forest
- MLP Classifier
- K Neighbours Classifier
- Support vector Machine
- Logistic Regression

Random Forest: Random Forest is the multiple number of decision trees, as in our project while using the random forest algorithm the no. of tree we used is 100 i.e. (n_estimators=100).

After using the random forest algorithm, we get the accuracy as follow:

ACCURACY OF THE TRAINING MODEL: 1.0
ACCURACY OF THE TEST MODEL: 0.9649122807017544

MLP Classifier: Multilayer perceptron classifier is a deep learning algorithm which connects the neural network. While working we used the backpropagation method of the MLP Classifier and perform the task of classification

After using the MLP classifier algorithm we get the accuracy as follow:

ACCURACY OF THE TRAINING MODEL: 0.9274725274725275
ACCURACY OF THE TEST MODEL: 0.9649122807017544

K-Neighbour classifier:

After using the K-Neighbour classifier algorithm we get the following accuracy:

ACCURACY OF THE test MODEL: 0.9385964912280702
ACCURACY OF THE train MODEL: 0.9428571428571428

Support Vector Machine (SVM): As it is mainly used for the classification. In our data and basically it creates the hyperplane between 2 classes

While working on our database we get the accuracy as follow:

ACCURACY OF THE TRAINING MODEL: 1.0

ACCURACY OF THE TEST MODEL: 0.6228070175438597

Logistic Regression: It is also used for the classification problem, in our database we use d it for the classification of the of malignant and the benign tumour.

After the prediction using the logistic regression algorithm, we got the accuracy of the model as follows:

ACCURACY OF THE TEST MODEL: 0.956140350877193

ACCURACY OF THE TRAINING MODEL: 0.9582417582417583

Now to compare the 2-classification problem i.e., support vector machine and logistic regression

It is better to create the ROC-curve

so the ROC curve of the following above classification model is:

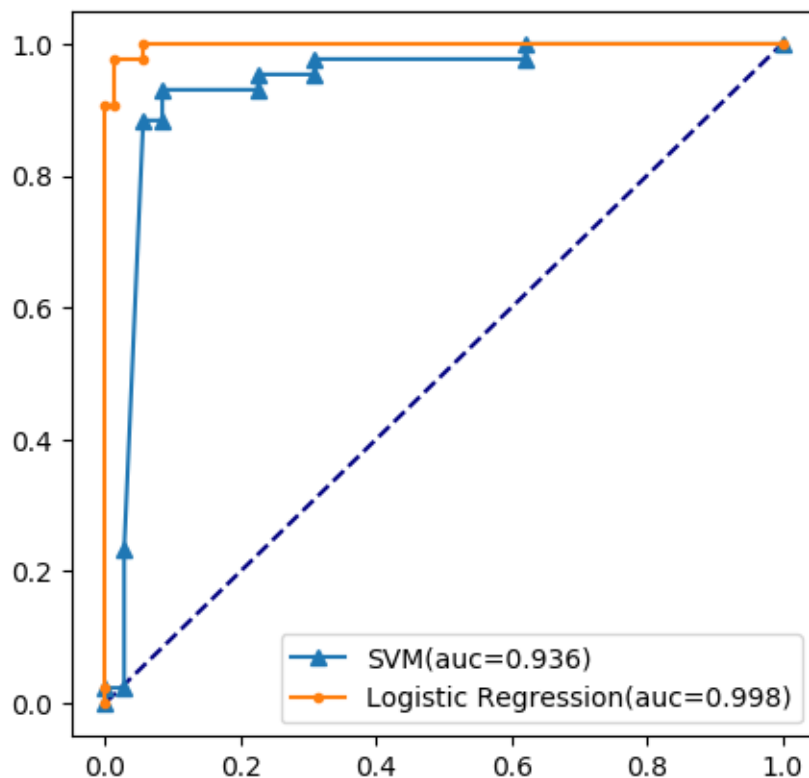


Figure 22: ROC-Curve of the SVM and Logistic Regression

Discussion:

Different studies had been done on the mitochondrial gene (MT gene) and it is found that the expression of the of these genes in progressive stages of various kinds of cancers. In our study it is found that the MT genes are responsible for the gene expression to cause various types of cancer for example in Breast, Brain, Liver Stomach, Kidney. These genes can be helpful in early risk assessment and the target for the Chemoprevention of the disease

MT genes are considered to be the more adenocarcinomas and it is found that the gene expression value of the MT genes is relatively higher in the carcinomas that the normal adjacent tissues,

Also, when it comes to the breast cancer the gene expression value of the of the studied MT genes is much lower in tubular than the adenocarcinomas

For example: In our study genes expression value of MT genes in various types of cancer:

Brain: '0.421934507', '0.240198852'

Breast: '0.159039148'

Liver: '0.555072301', '0.142393782'

Stomach: '0.394978443', '1.000000'

Although in various studies it comes that the expression value of the MT genes in the late stages is much higher (3rd and 4th stages) in adenocarcinomas.

To analyse the expression of the genes western blot analysis can be used.

The alteration of the gens could be result in the formation of mtDNA mutation and post transcriptional modification(mutation) and that can result in the influence of progressive states of various tumours.

Mitochondrial-encoded genes

Primer Sequence

MT-RNR1:

Forward: 5'-TAGAGGAGCCTGTTCTGTAATCGAT

Reverse: 5'-CGACCCTTAAGTTTCATAAGGGCTA

MT-ND1:

Forward: 5'-CCACCTCTAGCCTAGCCGTTTA

Reverse: 5'-GGGTCATGATGGCAGGAGTAAT

GUI Output:

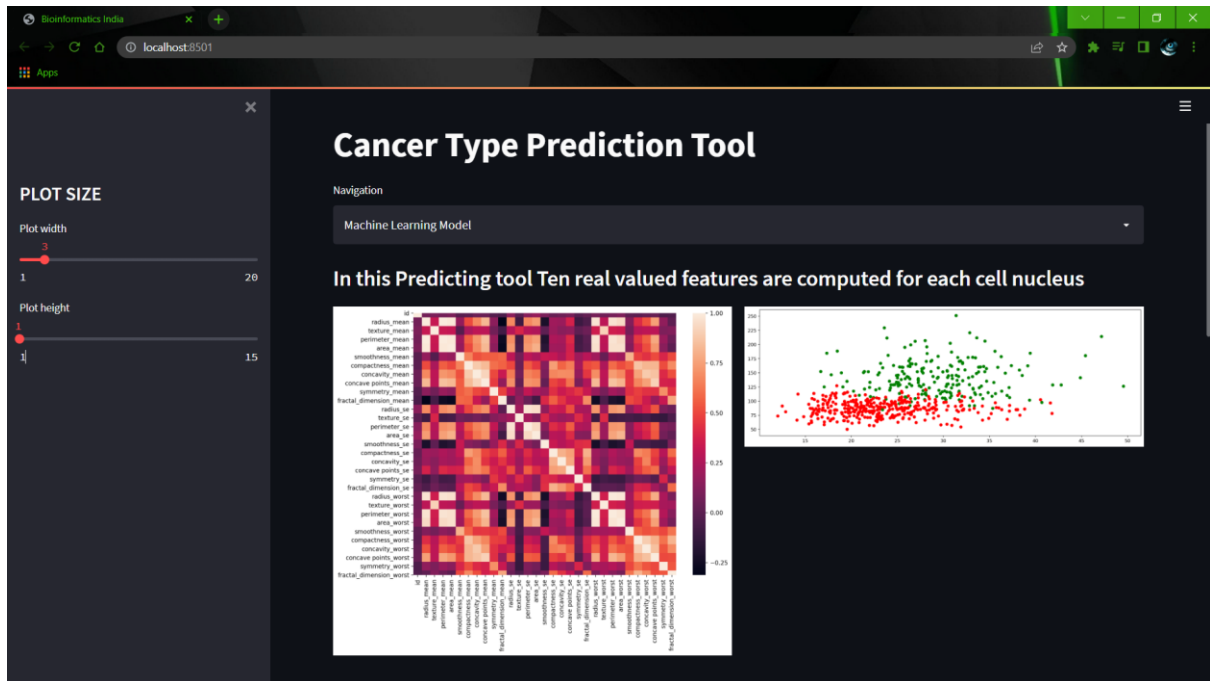


Figure: 23 Graphic user interfaces of the prediction tool

Model data:

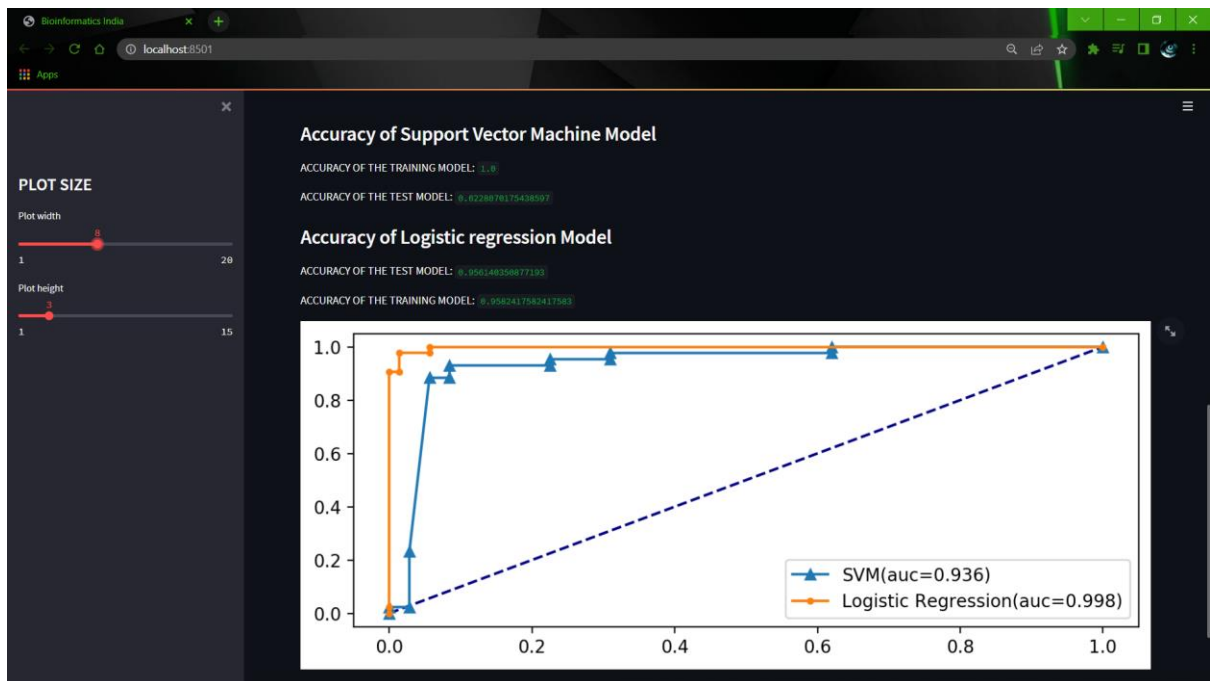


Figure: 24 Graphic user interfaces of the prediction tool

Cancer Prediction

The screenshot shows a web browser window with the URL localhost:8501. The page title is "Information (i)". Below the title, there are several descriptive text blocks: "Radius(the mean distance from the center to the points on the perimeter)", "Texture(standard deviation from grey scale values)", "Smoothness(local variation in radius length)", "Concavity(severity of concave points)", "No. of concave portions", and "Fractal Dimension (coastline approximation -1)".

Below these descriptions is a form with several input fields, each with a minus and plus button for adjustment:

- Radius: 0.00
- Texture: 0.00
- Smoothness: 0.00
- Concavity: 0.00
- Concave Points: 0.00
- Symmetry: 0.00
- Fractal_Dimension: 0.00

Below the input fields is a section titled "Which ML/Deep-Learning Model Do you want to use?" with five radio button options:

- Random Forest
- MLPClassifier
- Support Vector Machine
- Logistic Regression
- KNeighborsClassifier

A "Submit" button is located at the bottom left of the form area.

Figure: 25 Graphic user interfaces of the prediction tool

Biomarker identifier and predictor

The screenshot shows a web browser window with the URL localhost:8501. The page title is "Cancer causing genes(Biomarkers)". Below the title, there are several paragraphs of text:

"In cancer the genes expression data and the clinical data are very useful for the researchers and the scientist to analysis the type of the cancer and the biological function of the genes and also can predict the chance of getting cancer (Malignant and Benign)."

"After analysing the gene expression, we get the various genes (biomarkers) which are majorly responsible for causing the cancer in Brain, Kidney, Stomach, Breast, Liver"

"As the gene expression values of the genes is greater than the 0.05 which is much higher than the gene expression values of the other genes, and can be consider as the majority responsible in causing the cancer"

Biomarkers

Gene name of the ENSG00000198888.2: MT-ND1

It is the Mitochondrially encoded NADH: ubiquinone oxidoreductase core

Gene Expression Values:

Brain: 0.421934507

Breast: 0.159039148

Liver: 0.555072301

Stomach: 1.000000000

Biomarker Prediction

Information: Upload the csv file only, file should contain 1st column: Gene name or ID or Ensemble ID and 2nd column with gene expression values

Upload CSV File

Drag and drop file here
Limit: 200MB per file • CSV

Browse files

Figure: 26 Graphic user interfaces of the prediction tool

Appendix-1

To visualise the entire data set in pandas' data frame:

```
import matplotlib.pyplot as plt

import numpy as np
import pandas as pd
data=pd.read_csv('gene_exp_dataset.csv')
data.head(10)
# To check the data information
data.info()

# To describe the data value
data.describe()

#To check the shape of data
data.shape

#To check the count of each ensemble id in the dataset
pd.set_option('display.max_rows', None)
data['Ensemble_ID'].value_counts()

# To plot the graph of the gene expression values
plt.plot(data['Gene_exp_val'])
plt.legend('Gene Expression Values')

# To append the ensemble id into a list
file2=open('gene_exp_dataset.csv')
data2=csv.reader(file2)
sample=[]
for i in data2:
    for j in rows:
        if (i[0]==j):
            sample.append(i)
len(sample)

# To append the data in a list
import csv
file3=open('gene_exp_dataset.csv')
data3=csv.reader(file3)
rows1 = []
for row in data3:
    rows1.append(row)
rows1.pop(0)
len(rows1)
```



```
#Data filtration (select only the particular ensemble id which are available
in all 5 or 4 types of cancers)
i=0
j=0
s=[]
while j<len(rows1):
    for i in rows:
        a=i[0]
        if a == rows1[j][0]:
            s.append(rows1[j])
    j=j+1
print(len(s))
print(s)

# To check and filter the data whose gene expression value is more than 0.05
check=[]
for l in s:
    h=float(l[1])
    if h>=0.05:
        check.append(l)
check

#Append data into a csv file
df = pd.DataFrame(check)

# Append data frame to CSV file
df.to_csv('biomarkers.csv', mode='a', index=False, header=False)

# To visualise the data from biomarker file
data4=pd.read_csv('biomarkers.csv')
data4.head(25)
data4['ENSEMBLE_ID'].value_counts()

# To find the values of selected ensemble id
for i in range(0,len(check)):
    if check[i][0]=='ENSG00000198888.2':
        print(check[i])
for i in range(0,len(check)):
    if check[i][0]=='ENSG00000211459.2':
        print(check[i])
for i in range(0,len(check)):
    if check[i][0]=='ENSG00000198034.9':
        print(check[i])
for i in range(0,len(check)):
    if check[i][0]=='ENSG00000134240.10':
        print(check[i])
```

```
#Plot the graph of biomarkers and its expression value graph
plt.figure(figsize=(10,5))
sb=plt.subplot()
x1=[1,2,3,4]
sb.plot(x1,[0.421934507,0.159039148,0.555072301,1],c='green')
sb.plot([1,2,3],[0.240198852,0.142393782,0.394978443],c='red')
sb.plot([1,2],[0.53678367,0.76560264],c='black')
sb.plot([1,2],[0.082487431,0.051111419],c='yellow')
plt.xlabel("number of cancer")
plt.ylabel("gene expression value")
```

Appendix-2

```

#To visualise the database
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
data=pd.read_csv('cancer.csv')
data.head(10)

#drop the unused column
data=data.drop(['Unnamed: 32','Unnamed: 33'],axis=1)
#check the no. of features in the database
data.columns

#Check the no. of count of benign and malignamt tumour types
cancertype=data['diagnosis'].value_counts()
cancertype

# check the correaltion between the various featus of the database
#and create the heatmap of the corelation
datacorr=data.corr()
datacorr
plt.figure(figsize=(10,8))
sns.heatmap(datacorr)

#also create the scatter plot
plt.figure(figsize=(15,5))
sb=plt.subplot()
sb.scatter(data[data['diagnosis']=='M']['texture_worst'],data[data['diagnosis']
]=='M']['perimeter_worst'],c='green')
sb.scatter(data[data['diagnosis']=='B']['texture_worst'],data[data['diagnosis']
]=='B']['perimeter_worst'],c='red')

#division of data into training and test
x=data.drop(['diagnosis','id','smoothness_se','compactness_se'],axis=1)
y=data['diagnosis']
x,y
cancertype={'B':0,'M':1}
y=y.map(cancertype)
y
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2,
random_state=42)

#apply the random forest
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(n_estimators = 100)

```

```
# Training the model on the training dataset
# fit function is used to train the model using the training sets as
parameters
clf.fit(x_train, y_train)

# performing predictions on the test dataset
y_pred = clf.predict(x_test)
y_pred_train=clf.predict(x_train)

# metrics are used to find accuracy or error
from sklearn import metrics
print()

# using metrics module for accuracy calculation
print("ACCURACY OF THE TRAINING MODEL: ", metrics.accuracy_score(y_train,
y_pred_train))
print("ACCURACY OF THE TEST MODEL: ", metrics.accuracy_score(y_test, y_pred))

#apply the MLPClassifier
from sklearn.neural_network import MLPClassifier
mlp = MLPClassifier(hidden_layer_sizes=(8,8,8), activation='relu',
solver='adam', max_iter=500)
mlp.fit(x_train,y_train)

predict_train = mlp.predict(x_train)
predict_test = mlp.predict(x_test)

from sklearn.metrics import classification_report,confusion_matrix
print(confusion_matrix(y_train,predict_train))

#apply the KNeighbour classifier
from sklearn.neighbors import KNeighborsClassifier
model = KNeighborsClassifier(n_neighbors=2)

# Train the model using the training sets
model.fit(x_train,y_train)

#Predict Output
predicted= model.predict(x_test)
predicted2=model.predict(x_train)
print(predicted)
print("ACCURACY OF THE test MODEL: ", metrics.accuracy_score(y_test,
predicted))
print("ACCURACY OF THE train MODEL: ", metrics.accuracy_score(y_train,
predicted2))

#apply Support vector machine:
```

```
from sklearn.svm import SVC
svc=SVC(kernel='rbf',random_state=4)
svc.fit(x_train,y_train)

svm_predicted= svc.predict(x_test)
svm_predicted2=svc.predict(x_train)

print("ACCURACY OF THE TRAINING MODEL: ", metrics.accuracy_score(y_train,
svm_predicted2))
print("ACCURACY OF THE TEST MODEL: ", metrics.accuracy_score(y_test,
svm_predicted))

y_svc_pred=svc.decision_function(x_test)
y_svc_pred

#apply the Logistic regression:
from sklearn.linear_model import LogisticRegression
lr_model=LogisticRegression()
lr_model.fit(x_train,y_train)

lr_predicted= lr_model.predict(x_test)
lr_predicted2=lr_model.predict(x_train)
print("ACCURACY OF THE TEST MODEL: ", metrics.accuracy_score(y_test,
lr_predicted))
print("ACCURACY OF THE TRAINING MODEL: ", metrics.accuracy_score(y_train,
lr_predicted2))
y_lr_pred=lr_model.decision_function(x_test)

y_lr_pred

#plot the ROC curve between the SVM and Logistic Regression
from sklearn.metrics import roc_curve,auc
lr_fpr,lr_tpr,threshold1=roc_curve(y_test,y_lr_pred)
auc_lr=auc(lr_fpr,lr_tpr)
svm_fpr,svm_tpr,threshold2=roc_curve(y_test,y_svc_pred)
auc_svm=auc(svm_fpr,svm_tpr)

plt.figure(figsize=(5,5),dpi=100)
plt.plot([0, 1], [0, 1], color="navy", linestyle="--")
plt.plot(svm_fpr,svm_tpr,linestyle='-',
',marker='^',label='SVM(auc=%0.3f)'%auc_svm)
plt.plot(lr_fpr,lr_tpr,marker='.',label='Logistic
Regression(auc=%0.3f)'%auc_lr)
plt.legend()
plt.show()
```

Appendix-3 (GUI)

```
from multiapp import MultiApp
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import streamlit as st
import seaborn as sns

from apps import model, prediction, gene_exp

st.title('Cancer Type Prediction Tool')
app = MultiApp()

app.add_app("Machine Learning Model", model.app)
app.add_app("Prediction", prediction.app)
app.add_app("Gene Expression", gene_exp.app)

app.run()
```

References

1. Unger F.T., Witte I., David K.A. Prediction of individual response to anticancer therapy: Historical and future perspectives. *Cell. Mol. Life Sci.* 2015;72:729–757. doi: 10.1007/s00018-014-1772-3. [[PMC free article](#)] [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
2. Casamassimi A., Federico A., Rienzo M., Esposito S., Ciccodicola A. Transcriptome profiling in human diseases: New advances and perspectives. *Int. J. Mol. Sci.* 2017;18:1652. doi: 10.3390/ijms18081652. [[PMC free article](#)] [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
3. Wheeler D.A., Wang L. From human genome to cancer genome: The first decade. *Genome Res.* 2013;23:1054–1062. doi: 10.1101/gr.157602.113. [[PMC free article](#)] [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
4. Lowe R., Shirley N., Bleackley M., Dolan S., Shafee T. Transcriptomics technologies. *PLoS Comput. Biol.* 2017;13:e1005457. doi: 10.1371/journal.pcbi.1005457. [[PMC free article](#)] [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
5. Wang Z., Gerstein M., Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 2009;10:57–63. doi: 10.1038/nrg2484. [[PMC free article](#)] [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
6. Jacquier A. The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.* 2009;10:833–844. doi: 10.1038/nrg2683. [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
7. Cieřlik M., Chinnaiyan A.M. Cancer transcriptome profiling at the juncture of clinical translation. *Nat. Rev. Genet.* 2018;19:93–109. doi: 10.1038/nrg.2017.96. [[PubMed](#)] [[CrossRef](#)] [[Google Scholar](#)]
8. Ying Lu (yinglu@uiuc.edu), Cancer Classification Using Gene Expression Data
9. Aigner, T., Zien, A., Gehrsitz, A., Gebhard, P. M., McKenna, L. (2010). Anabolic and catabolic gene expression pattern analysis in normal versus osteoarthritic cartilage using complementary DNA-array technology. *Arthritis Rheumatism* 44 (12), 2777–2789. doi: 10.1002/1529-0131(200112)44:12<2777::aid-art465>3.0.co;2-h
10. Celis, J. E., Kruhøffer, M., Gromova, I., Frederiksen, C., østergaard, M., Thykjaer, T., et al. (2000). Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett.* 480 (1), 2–16. doi: 10.1016/s0014-5793(00)01771-3
11. Chen, Y., Li, Y., Narayan, R., Subramanian, A., Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics.* 32 (12), 1832–1839. doi: 10.1093/bioinformatics/btw074

12. Ozerov, I. V., Lezhnina, K. V., Izumchenko, E., Artemov, A. V., Medintsev, S., Vanhaelen, Q., et al. (2016). In silico pathway activation network decomposition analysis (iPANDA) as a method for biomarker development. *Nat. Commun.* 7, 13427. doi: 10.1038/ncomms13427
13. Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313 (5795), 1929–1935. doi: 10.1126/science.1132939
14. Penfold, C. A., Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface Focus* 1 (6), 857–870. doi: 10.1098/rsfs.2011.0053
15. GDC Documentation: Encyclopedia
<https://docs.gdc.cancer.gov/Encyclopedia/pages/HTSeq-FPKM/>
16. Zarringhalam K, Degras D, Brockel C, Ziemek D. Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes. *Sci Rep.* 2018;8(1):1237. <https://doi.org/10.1038/s41598-018-19635-0>.
17. P. Russel. *Fundamentals of Genetics*. Addison Wesley Longman Inc., 2000.
18. A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. In *Proc. of the Fourth Annual Int. Conf. on Computational Molecular Biology*, 2000.
19. M. Schena, D. Shalon, R. Davi, and P. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.
20. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in PyTorch. In: *Proceedings of Neural Information Processing Systems*; 2017.
21. Ankita Shukla, Tiratha Raj Singh (2021). Structure based inference of functional single nucleotide polymorphism and its role in TGFβ1 allied colorectal cancer (CRC). *International Journal of Bioinformatics Research and Applications*, 17 (1), 80-99
22. Arvind Kumar Yadav, Tiratha Raj Singh (2021). Novel inhibitors design through structural investigations and simulation studies for human PKMTs (SMYD2) involved in cancer. *Molecular Simulation*, 47 (14), 1149-1158
23. Manika Sehgal, Rajinder Gupta, Ahmed Moussa, Tiratha Raj Singh (2016). An Integrative Approach for Mapping Differentially Expressed Genes and Network Components Using Novel Parameters to Elucidate Key Regulatory Genes in Colorectal Cancer. *PLoS ONE*, 10 (7), e0133901

24. Manika Sehgal, Tiratha Raj Singh (2014). Computational Approach for the Identification of Plausible Biomarkers from Composite Networks and Gene Expression data Associated with Colorectal Cancer. *International Journal of Basic and Applied Biology*, 1 (), 62-66

25. Ankita Shukla, Tiratha Raj Singh (2018). Network-based Approach to Understand Dynamic Behaviour of Wnt Signaling Pathway Regulatory Elements in Colorectal Cancer. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 7 (14),

26. Ankita Shukla, Ahmed Moussa, Tiratha Raj Singh (2016). DREMECELS: A Curated Database for Base Excision and Mismatch Repair Mechanisms Associated Human Malignancies. *PLoS ONE*, 11 (6), e0157031