Ain Shams University
Faculty of Computer &
Information Sciences
Credit Hours New Programs
Accredited Faculty

جامعة عين شمس
كلية الحاسبات والمعلومات
البرامج جديدة للتعليم العالي
بنظام الساعات المعتمدة
كلية معتمدة

A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer

2021

By:

Gehad sayed Ali

Omnia Khaled

Esraa Mahmoud

Alaa sayed Gameel

Rawan mostafa

Ayaa Abdelhakeem

Under Supervision of:

Dr: Walaa Gad

Dr: Marco Alfons

TA: Esraa Hamdy

ACKNOWLEDGMENT

Words are written with love

We write to express our gratitude and appreciation for all your exerted efforts throughout the past four academic years. Your efforts resulted in what we are today, citizens yearning to serve their country with all the knowledge and experience we acquired from our most efficient, beloved and parent like professors, and special thanks for our graduation project supervisors,

Dr: Walaa Gad, Dr Marco Alfonse and TA Esraa Hamdy.

We are glade to work with you along the year and learn a lot from your knowledge.

Feeling a twinge of envy for new students replacing us

Might we meet again to prove to you how your students, through your care, love and efforts were able to reach the sky.

ABSTRACT

We can find the right treatment via Genomic profiles among different breast cancer survivors who received similar treatment.

More specifically, such profiling may help personalize the treatment based on the patients' gene expression.

As we know that Genomic profile has all information of what happen for cell and what reaction it can take with any chemical or any effect on it so we will use this wonderful possibility to predict right treatment for specific patient passed on his/her genomic profile which has relation with previous patient took one therapy and he lived or even dead when he took wrong therapy because of poor information about genomic profile and cancer which still consider hidden monster.

We implement our project based on machine learning techniques at the first we search for correlation between genes because we have almost 24500 genes and each gene has its gene EXP value that we are search in, so we will find

the solution in sckitlearn library in python in ML like “Chi sqr” after feature selection implementation we have the most related genes that we called them “BIOMARKERS” our target.

Then we implement classifier models using BIOMARKERS as features.

Know we can predict for any new genomic profile with the right treatment from only 3 types of therapy “Hormone, Chemo and surgery”.

TABLE OF CONTENT

CONTENT **PAGE**

CONTENT PAGE

CHAPTER1

INTRODUCTION

➢ 1.1) MOTIVATION

Despite the fast increase in the breast cancer incidence rate, the survival rates have also increased due to improvements in the treatments because of new technologies (Siegel et al., 2016). Breast cancer, however, is still one of the leading causes of cancer-related death among women worldwide. The survival rates vary among the various treatment therapies that are currently used, which include surgery, chemotherapy, hormone therapy, and radiotherapy. Nevertheless, each patient's response to a specific treatment varies based on some factors that are being investigated (Miller et al., 2016).

➢ 1.2) PROBLEM DEFINITION

- Traditional laboratory techniques like CAT scans and magnetic resonance imaging (MRI) have been proven to be useful. However, they provide very little information about the mechanism of the cancer progression.

- On the contrary, advances in DNA microarray technology have provided high throughput samples of gene expression.

- Machine learning approaches have been utilized to detect breast cancer treatment or survivals (Mangasarian and Wolberg, 2000; Cardoso et al., 2016; Abou Tabl et al., 2017; Tang et al., 2017; Zeng et al., 2018).

- many researchers have used DNA microarray technology to study breast cancer survivability (Mangasarian and

Wolberg, 2000; Cardoso et al., 2016; Abou Tabl et al., 2017).

- Analyzing gene expression among breast cancer patients who undergo varying treatment types deepens the current understanding of the disease’s progression and prognosis. Many features complicate the computational model; the number of features is usually significantly larger than the number of samples, which is known as the curse of dimensionality problem, in which standard classifiers overfit the data, and hence, perform poorly. Therefore, feature selection techniques are proven to alleviate the curse of dimensionality by removing irrelevant and/or redundant features.

1.3) OBJECTIVES:

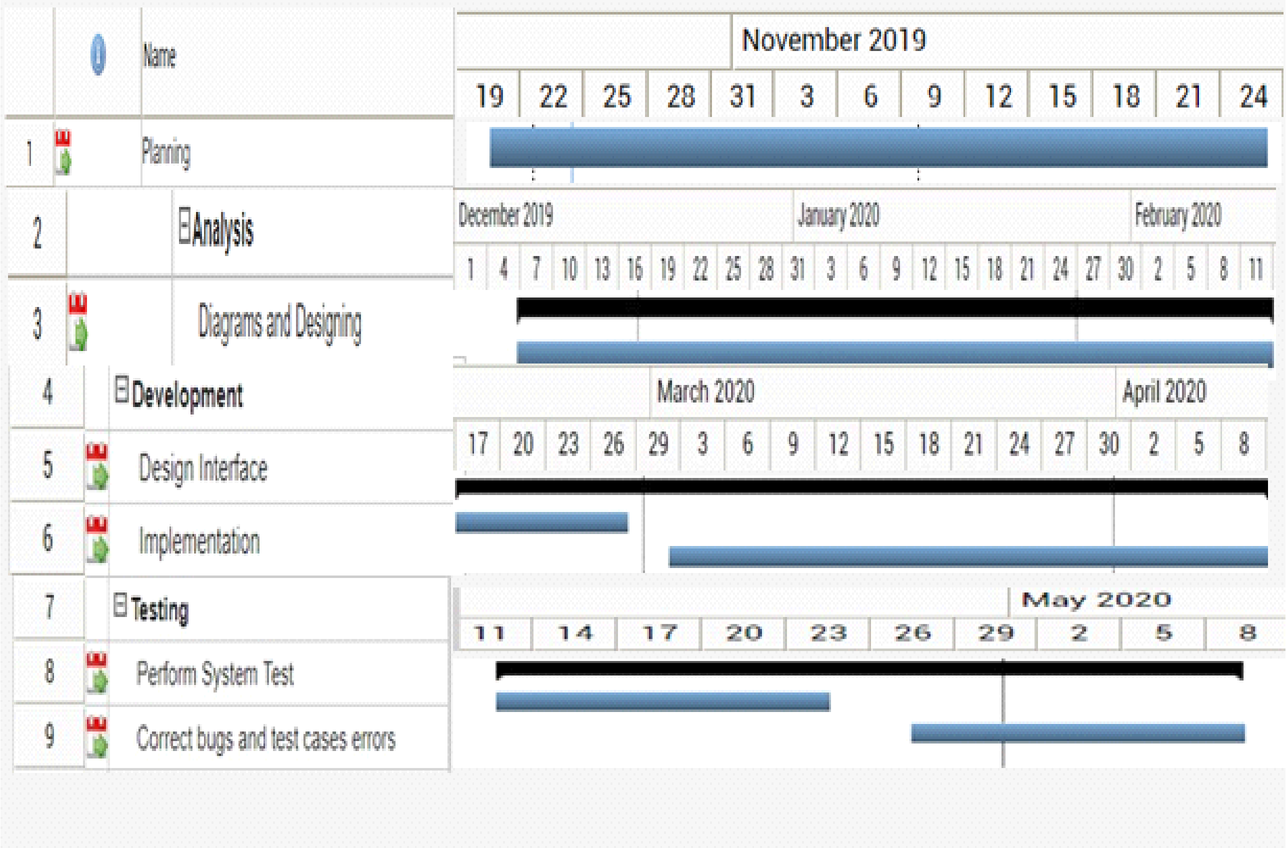
- **Genomic profiles among different breast cancer survivors who received similar treatment may provide clues about the key biological processes involved in the cells and finding the right treatment.**

- **By using machine learning approaches and by looking for different breast cancer survivors who received similar treatment We can find the right treatment**
- **treatment based on the patients' gene expression**
- **we will focus on breast cancer survivors who received hormone therapy, radiotherapy, or surgery treatment**
- **For the survivability information part, it defines whether the patient survives the 5-years interval or deceased.**
- **We will classify these six models and compare one class vs. the rest at each**

node, which makes the tree-based model creates five nodes.

- These classes are the combination of each treatment: surgery (S), hormone therapy (H), and radiotherapy (D) with a patient status as living (L) or deceased (D).
- Data from a total of 347 patients will include in this work.
- Totally from 24367 Genes.

1.4) TIME PLAN:



	Name
1	Planning
2	⊟Analysis
3	Diagrams and Designing
4	⊟Development
5	Design Interface
6	Implementation
7	⊟Testing
8	Perform System Test
9	Correct bugs and test cases errors

CHAPTER 2

2.1 SCIENTIFIC BACKGROUND

Introduction in genes:

- Learning some basic facts about DNA, RNA and proteins is helpful for understanding the importance of biomarkers in cancer.
- DNA, which stands for deoxyribonucleic acid, is a molecule inside the cell that carries genetic information and passes it on from one generation to the next.
- RNA, or ribonucleic acid, contains information that has been copied from DNA.
- Body cells make several different types of RNA molecules that are necessary for the synthesis of protein molecules. For example, mRNA, or messenger RNA molecules, serves as templates for the synthesis of proteins from amino acid building blocks, while tRNA, or transfer RNA molecules, bring the amino acid residues to the ribosome.

- Inside the ribosome – an organelle where the protein is being synthesized – tRNA “reads” the mRNA template in a process called translation.

What are biomarkers?

- **Biomarkers** are molecules that indicate normal or abnormal process taking place in your body and may be a sign of an underlying condition or disease.
- Various types of molecules, such as DNA (genes), proteins or hormones, can serve as biomarkers, since they all indicate something about your health.
- Biomarkers may be produced by the cancer tissue itself or by other cells in the body in response to cancer.
- They can be found in the blood, stool, urine, tumor tissue, or other tissues or bodily fluids. Notably, biomarkers are not limited to cancer.
- There are biomarkers for heart disease, multiple sclerosis, and many other diseases.

- Proteins help the body function properly and are the basis of body structures such as skin and hair.
- They have a wide range of functions inside the human body.
- Certain proteins speed up chemical reactions (enzymes), others affect the functioning of the immune system (cytokines), and yet others, known as antibodies, trigger specific immune responses in response to antigens – harmful substances that the body periodically has to overcome

Cancer biomarkers can include:

- Proteins
- Gene mutations (changes)
- Gene rearrangements
- Extra copies of genes
- Missing genes
- Other molecules

Functions of Cancer Biomarkers

There are many types of cancer biomarkers, and they each work differently within the body and react differently to treatments. In general, cancer biomarkers are classified by their different functions:

When people talk about cancer biomarkers they’re usually referring to proteins, genes, and other molecules that affect how cancer cells grow, multiply, die, and respond to other compounds in the body. In recent years, scientists have started to look at patterns of gene expression and changes in DNA as cancer biomarkers. While some cancer biomarkers can be used to predict how aggressively your cancer will grow, and are therefore useful for assessing your prognosis (outlook), the most promising use of biomarkers today is to identify which therapies a particular patient’s cancer may or may not respond to.

Detecting and Measuring Biomarkers to Develop a Personalized Anticancer Treatment Plan

- In order to determine if, and at what levels, certain biomarkers are present in your cancer, your doctor will need to take a sample of tumor tissue or bodily fluid and send it to a laboratory to conduct a series of advanced pathology and molecular profiling tests.
- Those tests will detect and measure the levels of your cancer’s specific biomarkers.
- Obtained information will then be matched with published research by the world’s leading cancer researchers to identify which treatments are and are not likely to work.
- Your doctor will then receive a report that lists all the biomarkers that have

been detected in the sample, along with the treatments that have been identified as positively and negatively associated with those biomarkers.

- This process allows your doctor to personalize your anticancer treatment plan based on your cancer's unique biomarker profile.

Difficult targets

Biomarkers don't always tell the full story. Discovery of a biomarker that might indicate an increased cancer risk doesn't mean a patient will get cancer. Not all cancers have identifiable biomarkers. And identifying a driving biomarker in a cancer does not necessarily lead to a treatment option. Some biomarkers for cancer have no corresponding targeted therapy or immunotherapy drug. For example:

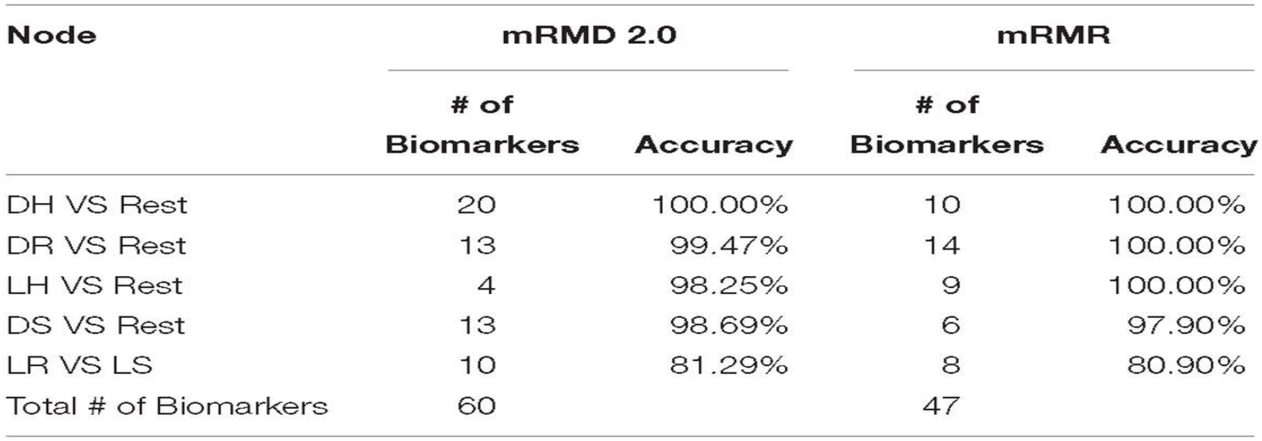
- TP53: Tumor protein 53 is a tumor suppressor gene designed to help stop cancer cells from growing. TP53 mutations are the most common found in cancer cells and may be found in most types of cancer.

- RAS: About 30 percent of all cancers, including 95 percent of all pancreatic cancers, have known mutations in the RAS family of genes that control cell death and growth.

2.2) RELATED WORK:

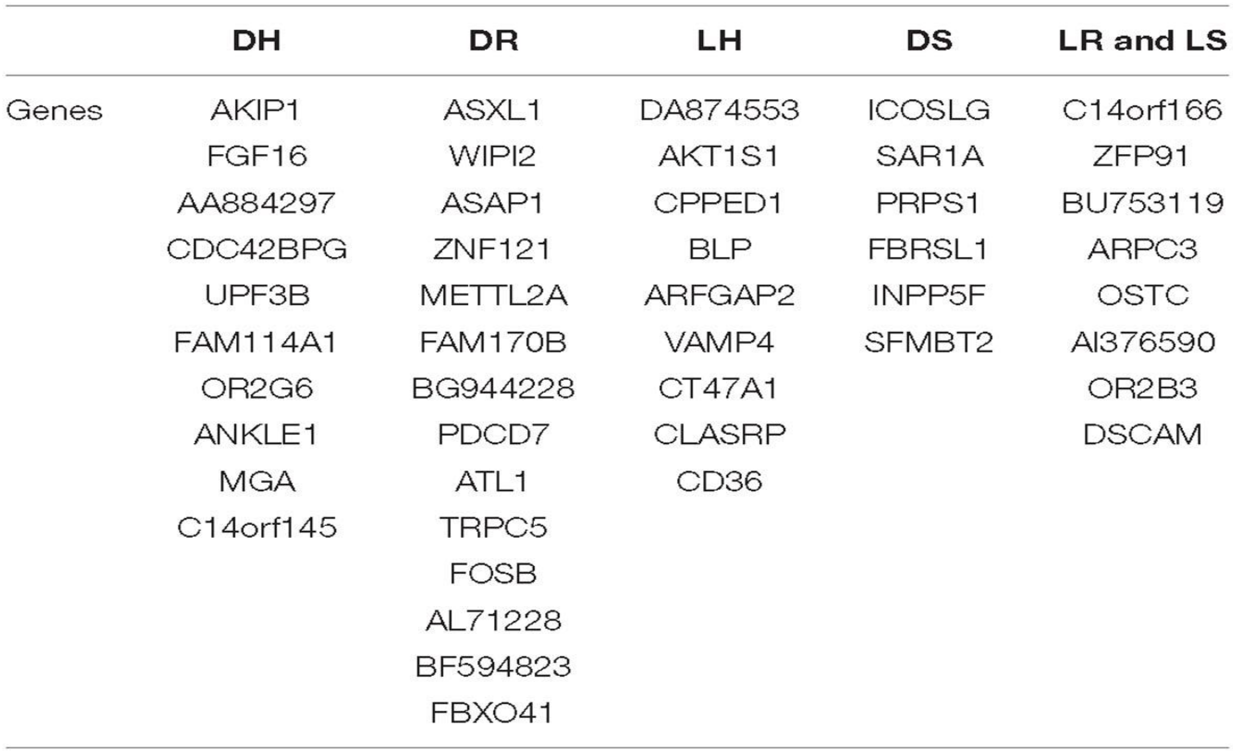
Title	Year	objective	Dataset	Accuracy
A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer	**March 2019**	finding the right treatment for breast cancer “find **Gene Biomarkers Guiding the Treatment of Breast Cancer** “	dataset contains 24,368 genes for each of the 347 samples	Table 1

TABLE 1:



Node	mRMD 2.0		mRMR	
	# of Biomarkers	Accuracy	# of Biomarkers	Accuracy
DH VS Rest	20	100.00%	10	100.00%
DR VS Rest	13	99.47%	14	100.00%
LH VS Rest	4	98.25%	9	100.00%
DS VS Rest	13	98.69%	6	97.90%
LR VS LS	10	81.29%	8	80.90%
Total # of Biomarkers	60		47	

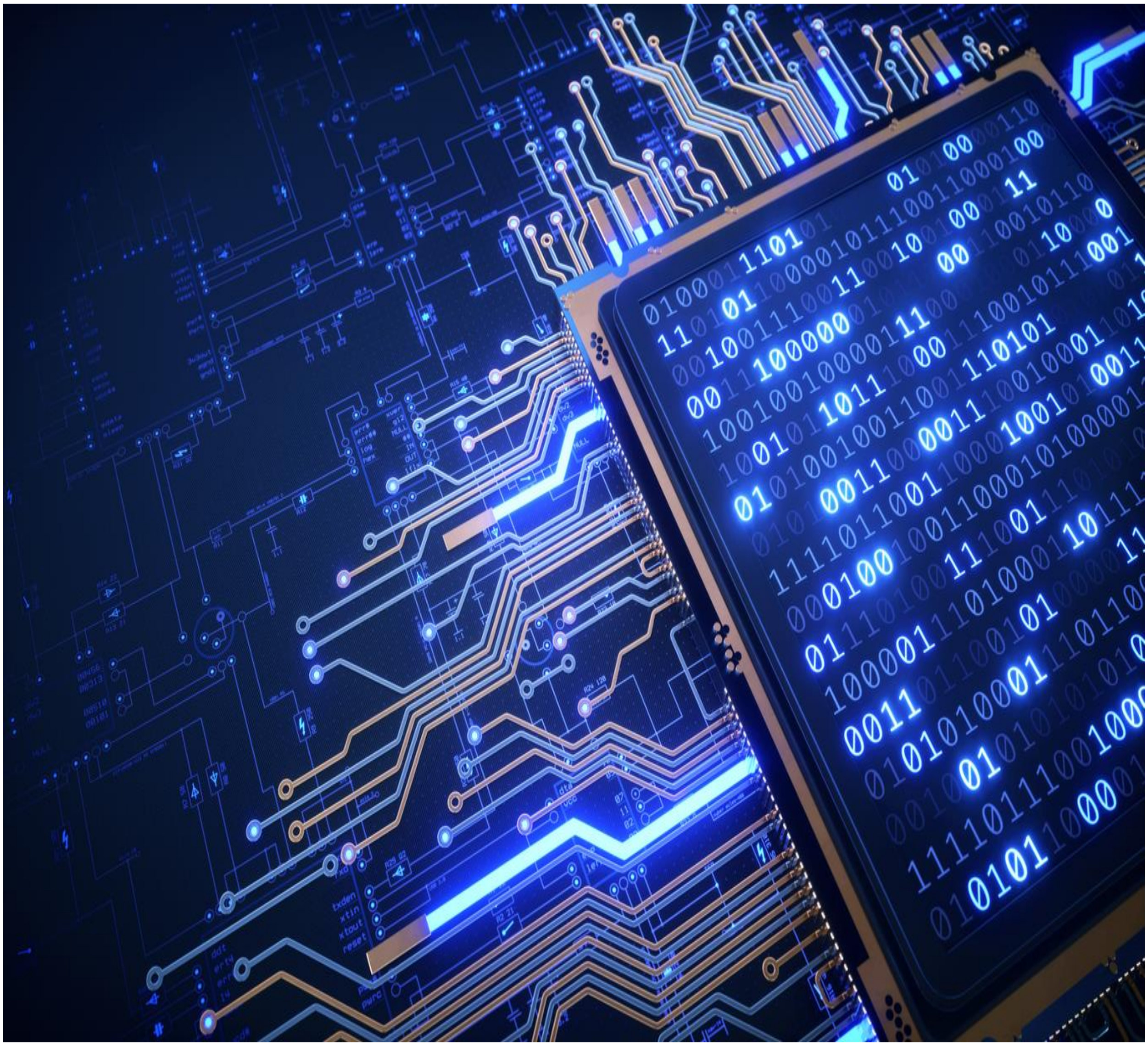
BIOMARKER RESULT OF THIS RESEARCH:



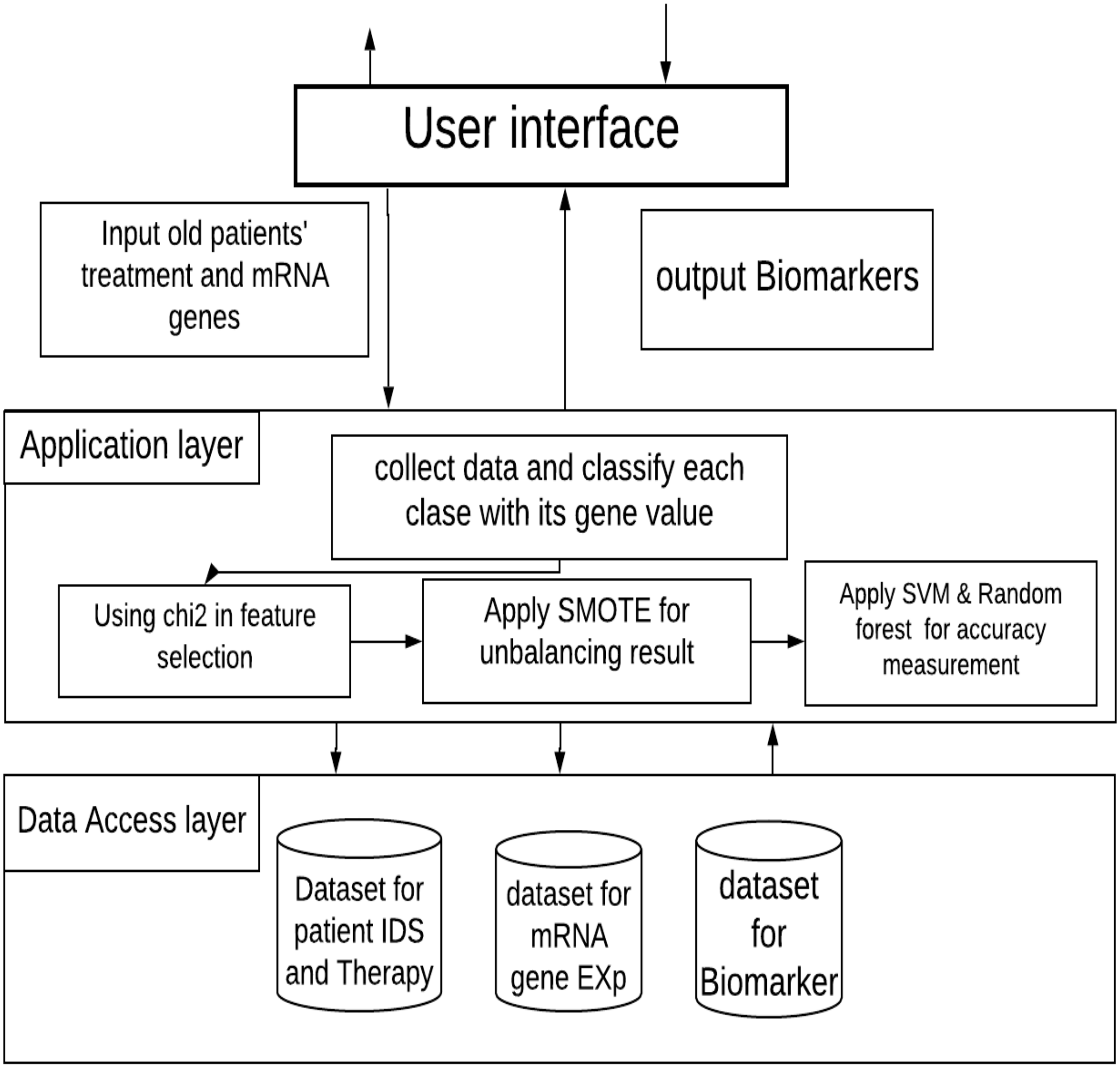
	DH	DR	LH	DS	LR and LS
Genes	AKIP1	ASXL1	DA874553	ICOSLG	C14orf166
	FGF16	WIPI2	AKT1S1	SAR1A	ZFP91
	AA884297	ASAP1	CPPED1	PRPS1	BU753119
	CDC42BPG	ZNF121	BLP	FBRSL1	ARPC3
	UPF3B	METTL2A	ARFGAP2	INPP5F	OSTC
	FAM114A1	FAM170B	VAMP4	SFMBT2	AI376590
	OR2G6	BG944228	CT47A1		OR2B3
	ANKLE1	PDCD7	CLASRP		DSCAM
	MGA	ATL1	CD36		
	C14orf145	TRPC5			
		FOSB			
		AL71228			
		BF594823			
		FBXO41			

CHAPTER 3

ANALYSIS AND DESIGN



SYSTEM ARCHITECTURE:


User interface
Input old patients' treatment and mRNA genes
output Biomarkers
Application layer
collect data and classify each clase with its gene value
Using chi2 in feature selection
Apply SMOTE for unbalancing result
Apply SVM & Random forest for accuracy measurement
Data Access layer
Dataset for patient IDS and Therapy
dataset for mRNA gene EXp
dataset for Biomarker

- Data access layer :

1. Patients IDs and therapy

   which is online database contain patient’s id, treatment which patient took and its clinical data.

2. mRNA Exp genes dataset

   It is dataset has all genes for previous patient we know from patient id dataset

3. Biomarker

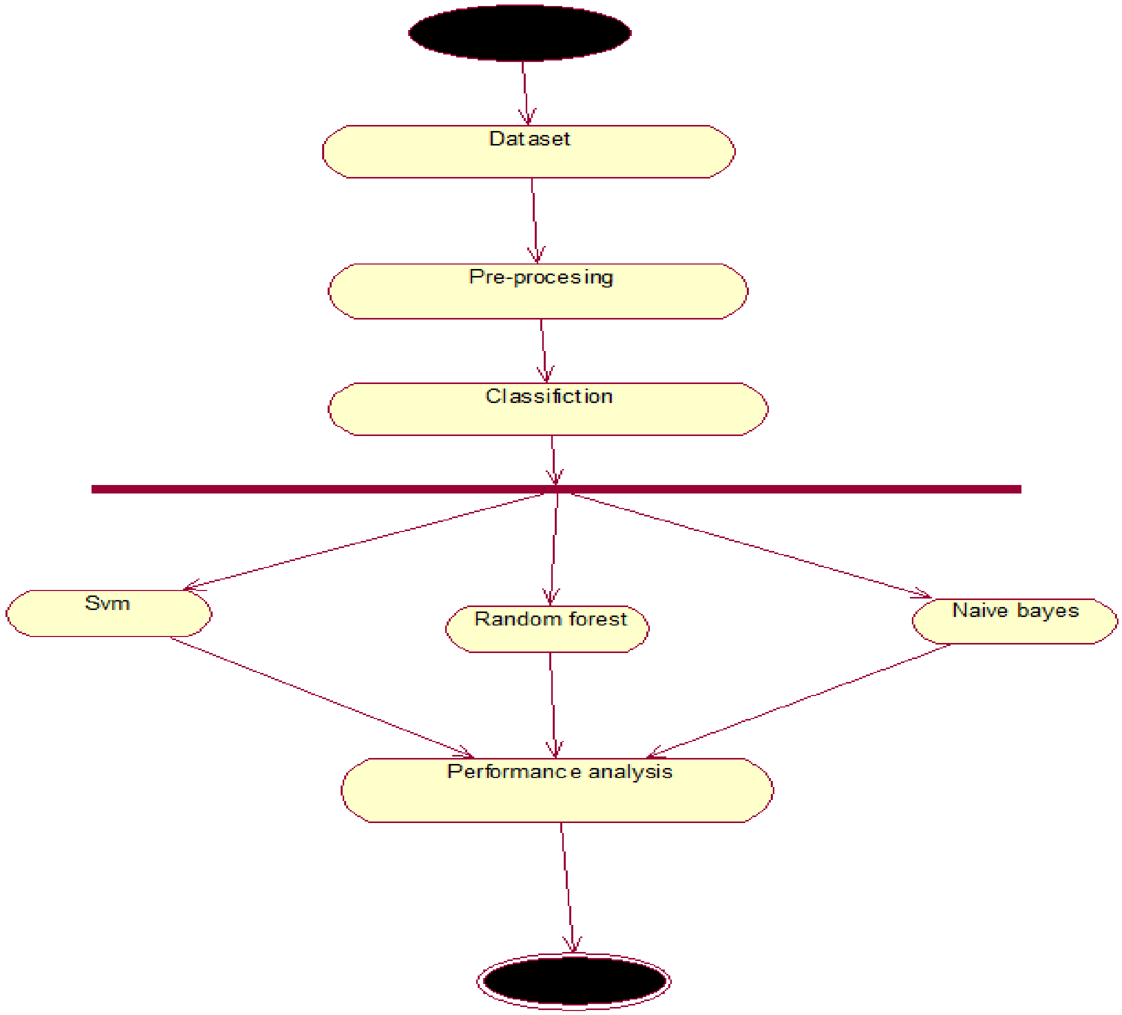
   result dataset of biomarkers

- Application layer :

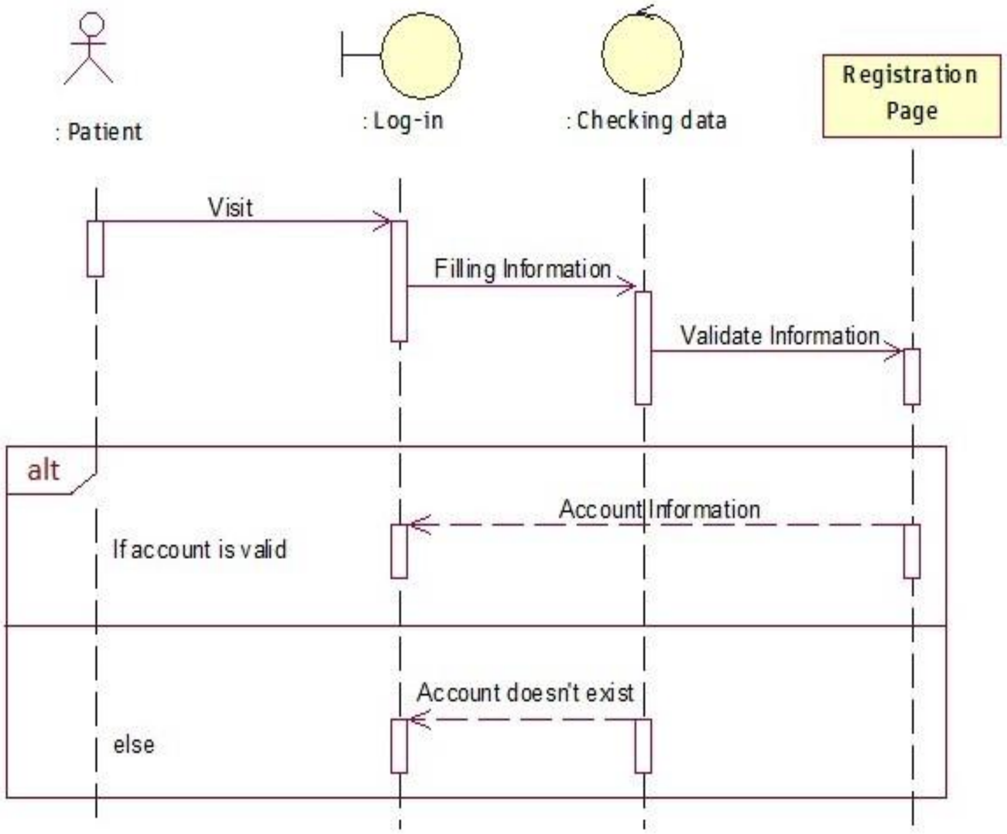
1. collect data and classify each class with its gene value
2. Using chi2 in feature selection
3. Apply SMOTE for unbalancing result
4. Apply SVM & Random forest for accuracy measurement

SYSTEM ANALYSIS & DESIGN

Activity diagram for Data processing



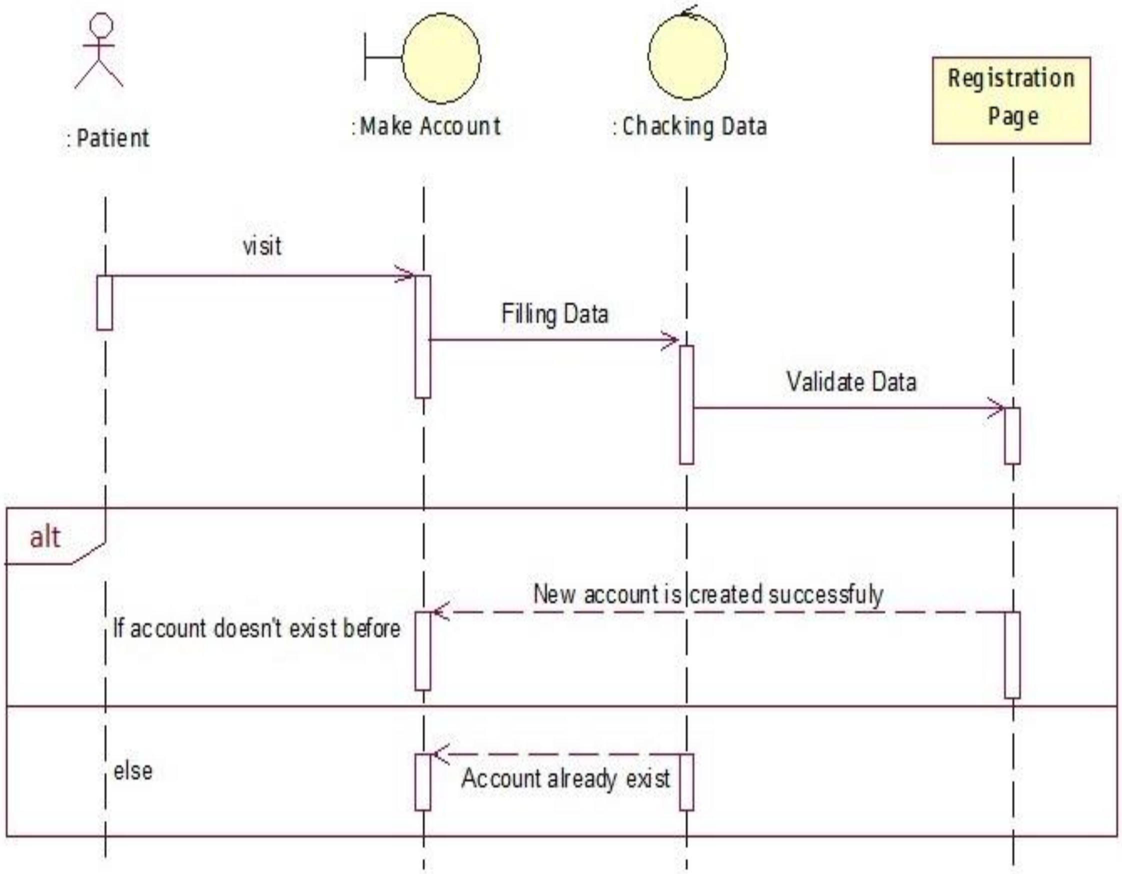
Sequence diagram:
-log in



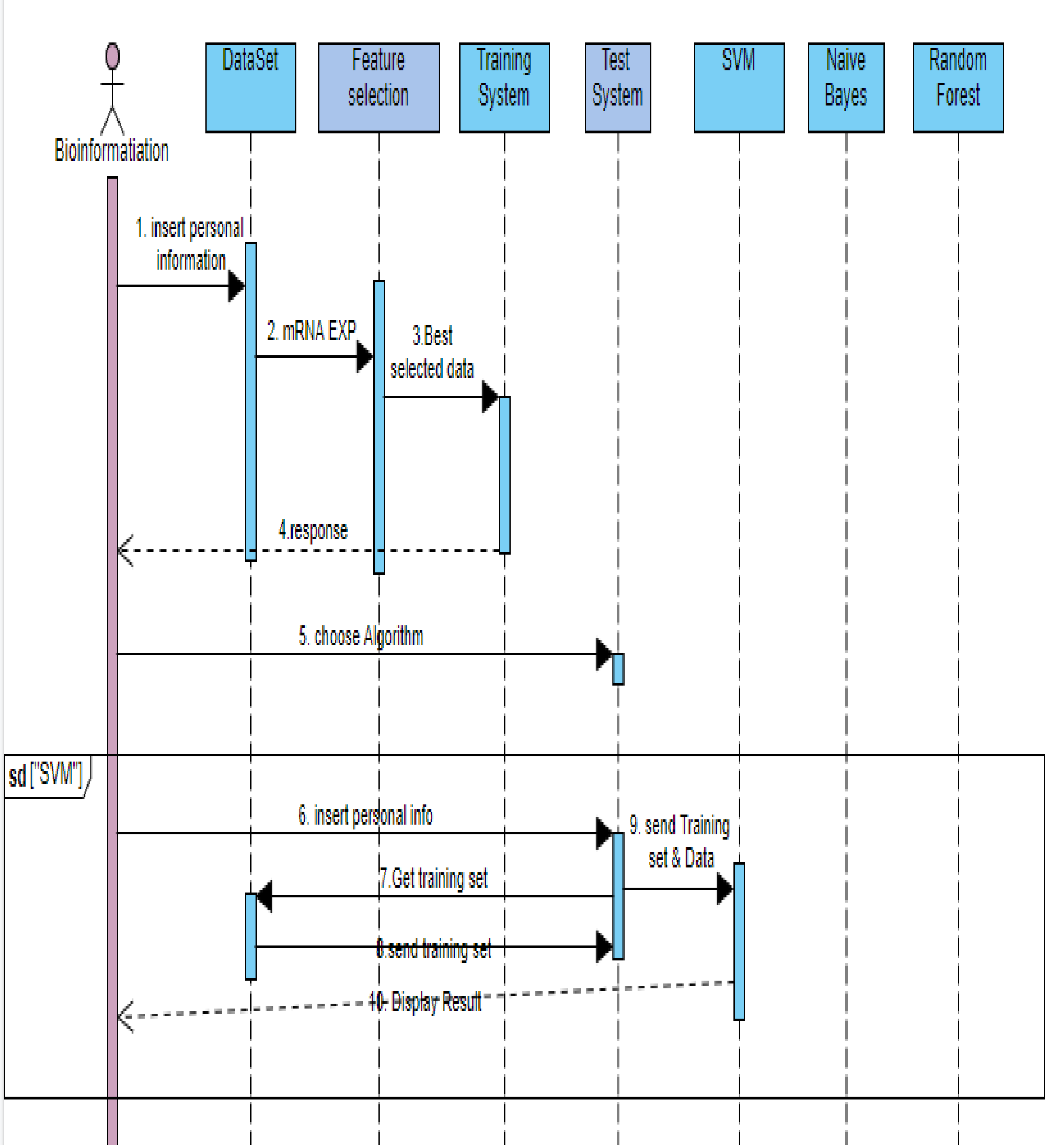
SYSTEM ANALYSIS & DESIGN

Sequence diagram:

Register

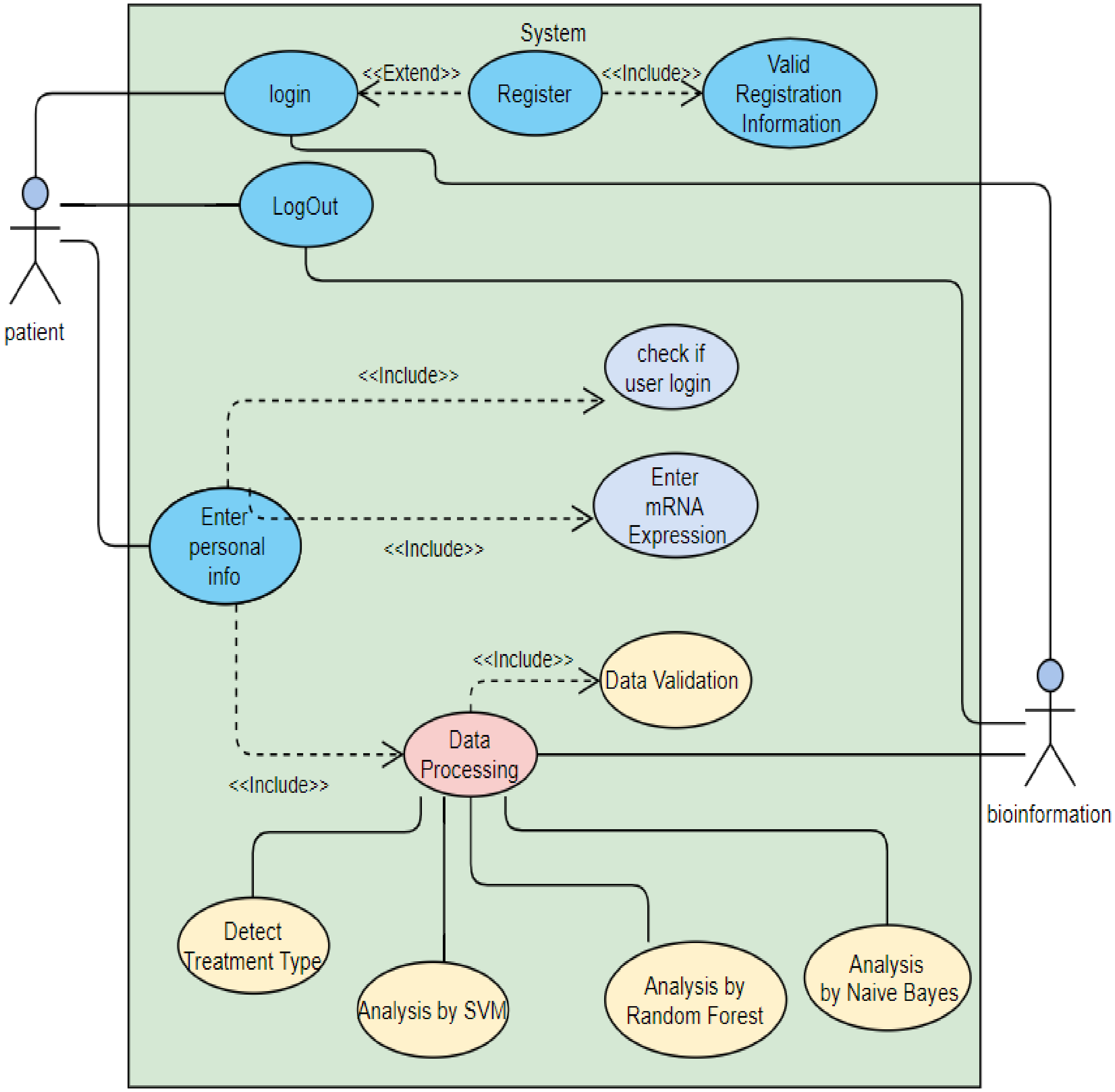


Sequence diagram:



SYSTEM ANALYSIS & DESIGN

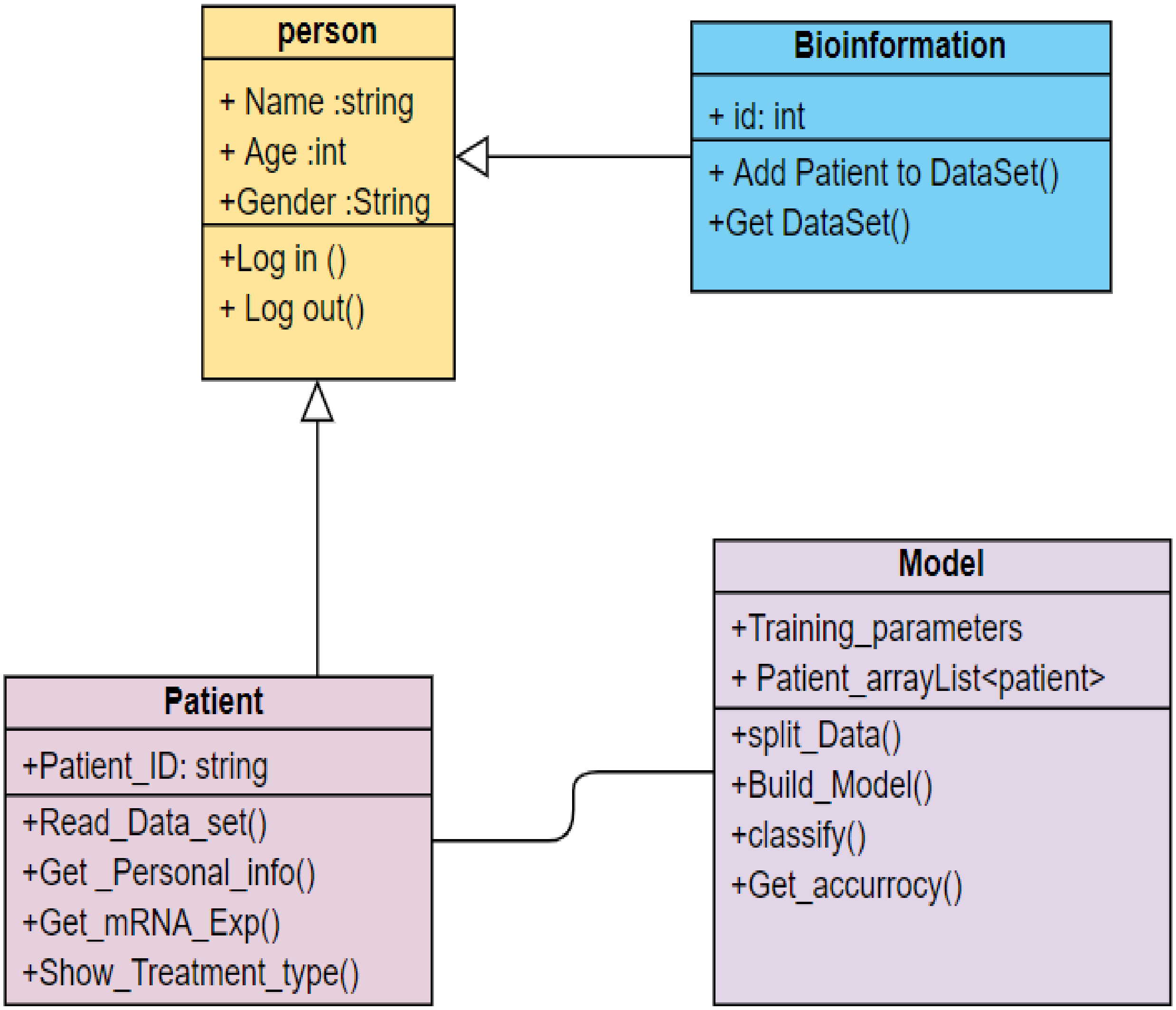
Use case:



SYSTEM ANALYSIS & DESIGN

Class diagram

UML:



CHAPTER 4

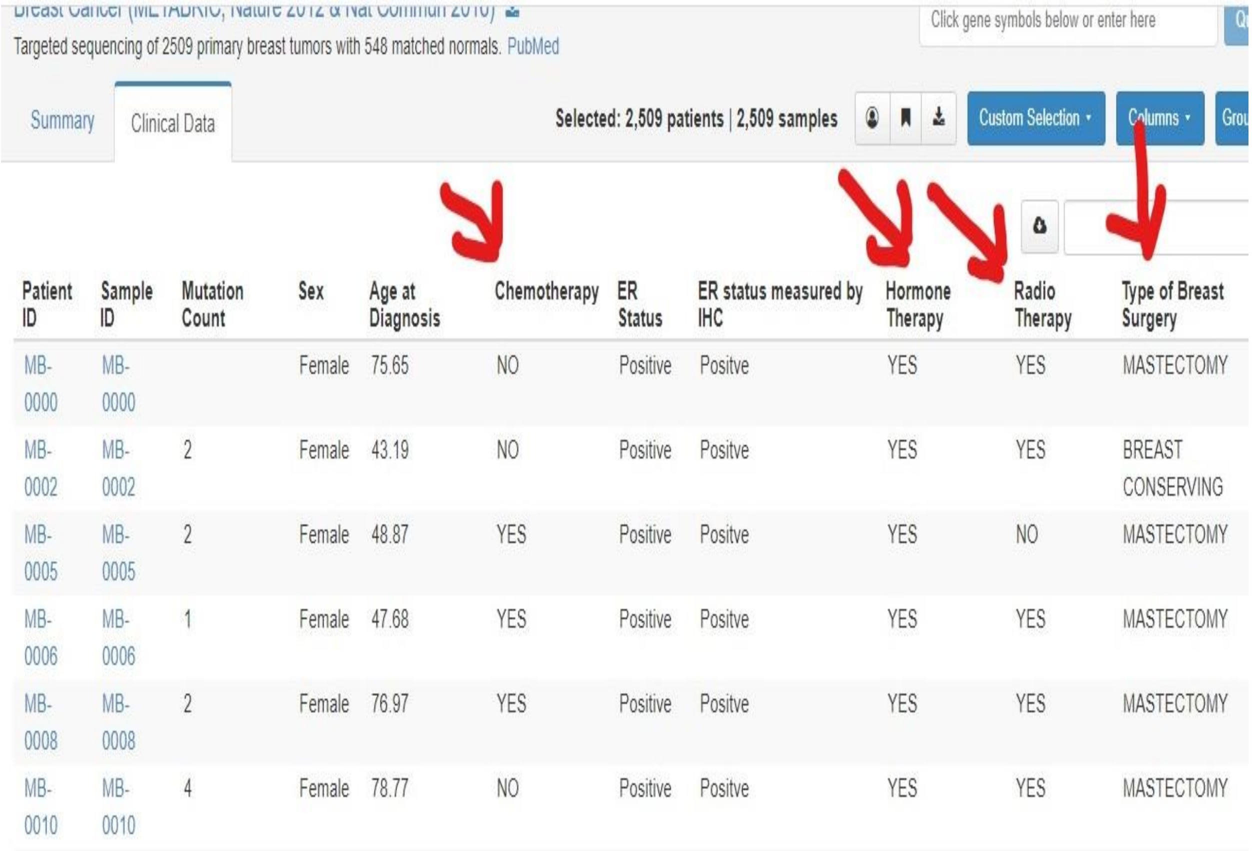
IDEA
STRATEGY
PLAN
Creativity
SUCCESS
Performance
STRATEGY
Business Performance
MANAGEMENT
IDEAS
MANAGE
STRATEGIC
FUTURE
Solutions
Ideas
RESULTS
BUSINESS PLAN
Implementation
SUCCESS
GOALS
Manage
TEAM
INNOVATION
VALUE-ADD
PLAN
Improve
PROBLEM SOLVING
Progressive
DYNAMIC
STRATEGY
FUTURE
MISSION
ACTIVITIES

4.1)DATASET PREPARATION:

we used online breast cancer dataset in “cBioPortal” website for online datasets.

FIRST…

- Select only patients who took one Therapy from dataset which contain clinical data for all patient



Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)

Targeted sequencing of 2509 primary breast tumors with 548 matched normals. PubMed

Selected: 2,509 patients | 2,509 samples

Patient ID	Sample ID	Mutation Count	Sex	Age at Diagnosis	Chemotherapy	ER Status	ER status measured by IHC	Hormone Therapy	Radio Therapy	Type of Breast Surgery
MB-0000	MB-0000		Female	75.65	NO	Positive	Positve	YES	YES	MASTECTOMY
MB-0002	MB-0002	2	Female	43.19	NO	Positive	Positve	YES	YES	BREAST CONSERVING
MB-0005	MB-0005	2	Female	48.87	YES	Positive	Positve	YES	NO	MASTECTOMY
MB-0006	MB-0006	1	Female	47.68	YES	Positive	Positve	YES	YES	MASTECTOMY
MB-0008	MB-0008	2	Female	76.97	YES	Positive	Positve	YES	YES	MASTECTOMY
MB-0010	MB-0010	4	Female	78.77	NO	Positive	Positve	YES	YES	MASTECTOMY

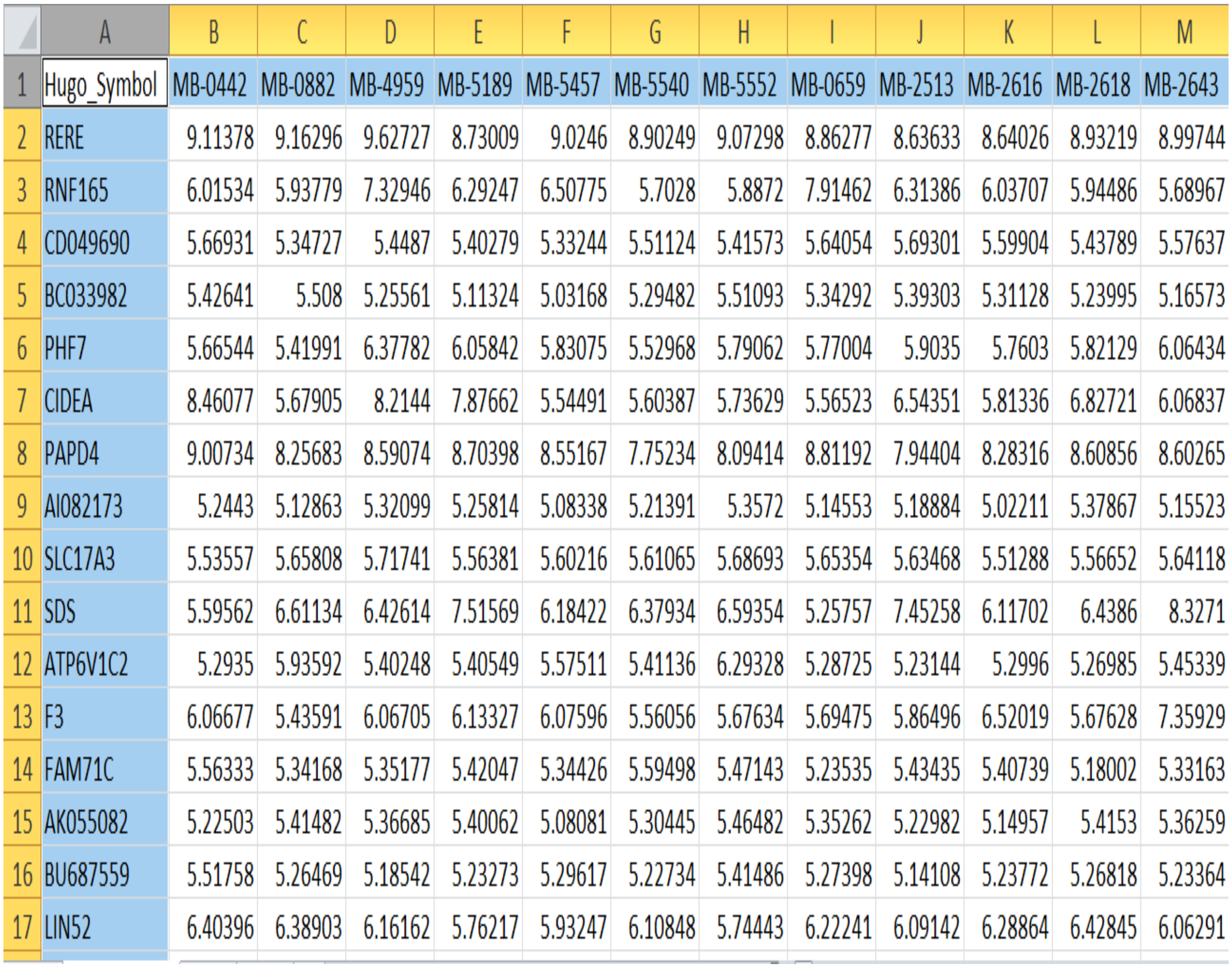
Here we must choose the patient ID who has one “Yes” in Therapies columns and the other Therapies columns equal “NO” that’s mean the patient had been treated with one therapy.

We don’t care even if patient lived or dead because of cancer, because we have to get “Biomarkers” for lived patients and dead patients

SECOND…

- Get mRNA Expression.
  We will get “Biomarkers” from mRNA EXp so we collect all these genes for every patient we had selected from the first clinical dataset

- Here we will see the dataset that contains mRNA EXp values for each patient



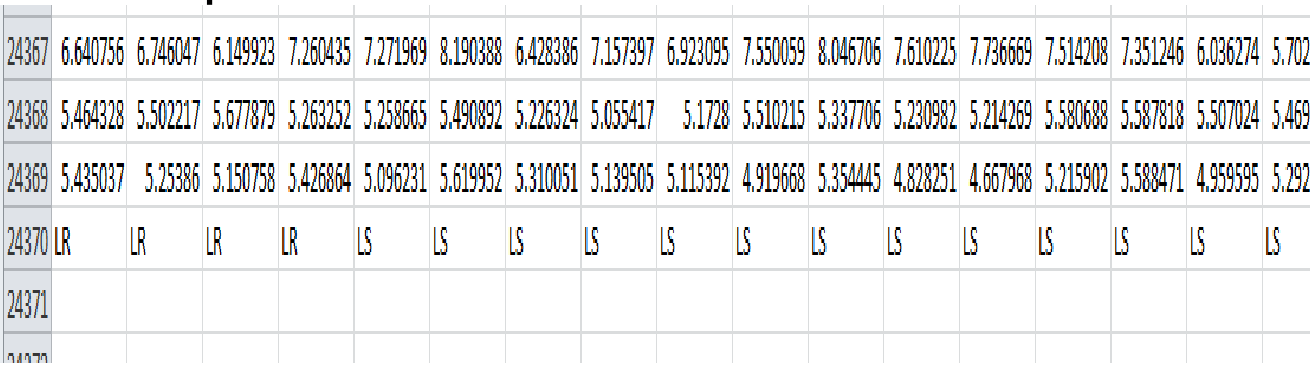
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Hugo_Symbol	MB-0442	MB-0882	MB-4959	MB-5189	MB-5457	MB-5540	MB-5552	MB-0659	MB-2513	MB-2616	MB-2618	MB-2643
2	RERE	9.11378	9.16296	9.62727	8.73009	9.0246	8.90249	9.07298	8.86277	8.63633	8.64026	8.93219	8.99744
3	RNF165	6.01534	5.93779	7.32946	6.29247	6.50775	5.7028	5.8872	7.91462	6.31386	6.03707	5.94486	5.68967
4	CD049690	5.66931	5.34727	5.4487	5.40279	5.33244	5.51124	5.41573	5.64054	5.69301	5.59904	5.43789	5.57637
5	BC033982	5.42641	5.508	5.25561	5.11324	5.03168	5.29482	5.51093	5.34292	5.39303	5.31128	5.23995	5.16573
6	PHF7	5.66544	5.41991	6.37782	6.05842	5.83075	5.52968	5.79062	5.77004	5.9035	5.7603	5.82129	6.06434
7	CIDEA	8.46077	5.67905	8.2144	7.87662	5.54491	5.60387	5.73629	5.56523	6.54351	5.81336	6.82721	6.06837
8	PAPD4	9.00734	8.25683	8.59074	8.70398	8.55167	7.75234	8.09414	8.81192	7.94404	8.28316	8.60856	8.60265
9	AI082173	5.2443	5.12863	5.32099	5.25814	5.08338	5.21391	5.3572	5.14553	5.18884	5.02211	5.37867	5.15523
10	SLC17A3	5.53557	5.65808	5.71741	5.56381	5.60216	5.61065	5.68693	5.65354	5.63468	5.51288	5.56652	5.64118
11	SDS	5.59562	6.61134	6.42614	7.51569	6.18422	6.37934	6.59354	5.25757	7.45258	6.11702	6.4386	8.3271
12	ATP6V1C2	5.2935	5.93592	5.40248	5.40549	5.57511	5.41136	6.29328	5.28725	5.23144	5.2996	5.26985	5.45339
13	F3	6.06677	5.43591	6.06705	6.13327	6.07596	5.56056	5.67634	5.69475	5.86496	6.52019	5.67628	7.35929
14	FAM71C	5.56333	5.34168	5.35177	5.42047	5.34426	5.59498	5.47143	5.23535	5.43435	5.40739	5.18002	5.33163
15	AK055082	5.22503	5.41482	5.36685	5.40062	5.08081	5.30445	5.46482	5.35262	5.22982	5.14957	5.4153	5.36259
16	BU687559	5.51758	5.26469	5.18542	5.23273	5.29617	5.22734	5.41486	5.27398	5.14108	5.23772	5.26818	5.23364
17	LIN52	6.40396	6.38903	6.16162	5.76217	5.93247	6.10848	5.74443	6.22241	6.09142	6.28864	6.42845	6.06291

- Now we have a data we will make preprocessing for it data consist of patient IDs and almost 24.369 genes for every patient

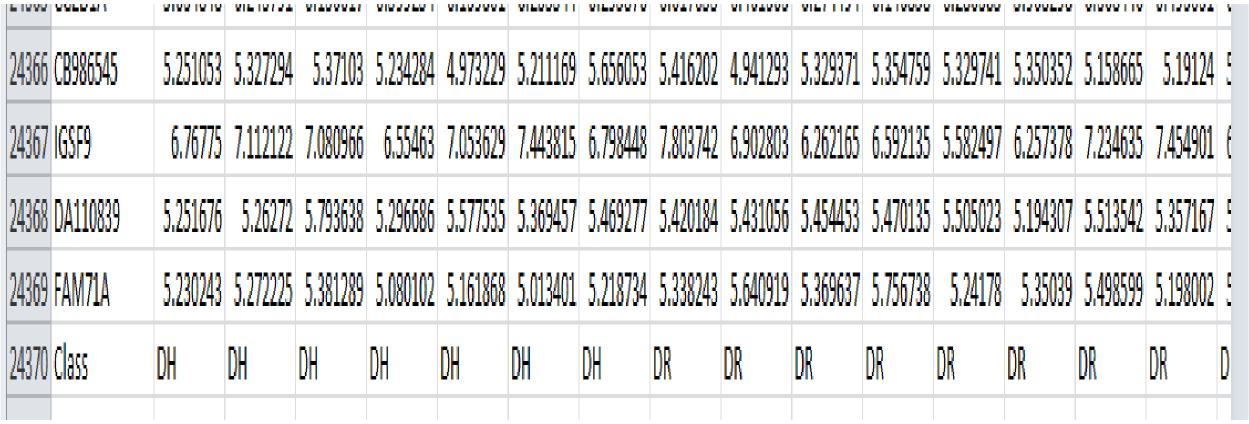
THIRD…

- We mention before that we have to classify Data into 6 Classes
  - 1 LIVE AND RADIO (LR).
  - 2 DEAD AND RADIO (DR).
  - 3 LIVE AND HORMONE (LH).
  - 4 DEAD AND HORMONE (DH).
  - 5 LIVE AND SURGERY (LS).
  - 6 DEAD AND SURGERY (DS).

- So we classified each patient with specific class
- This is part of classified Dataset.



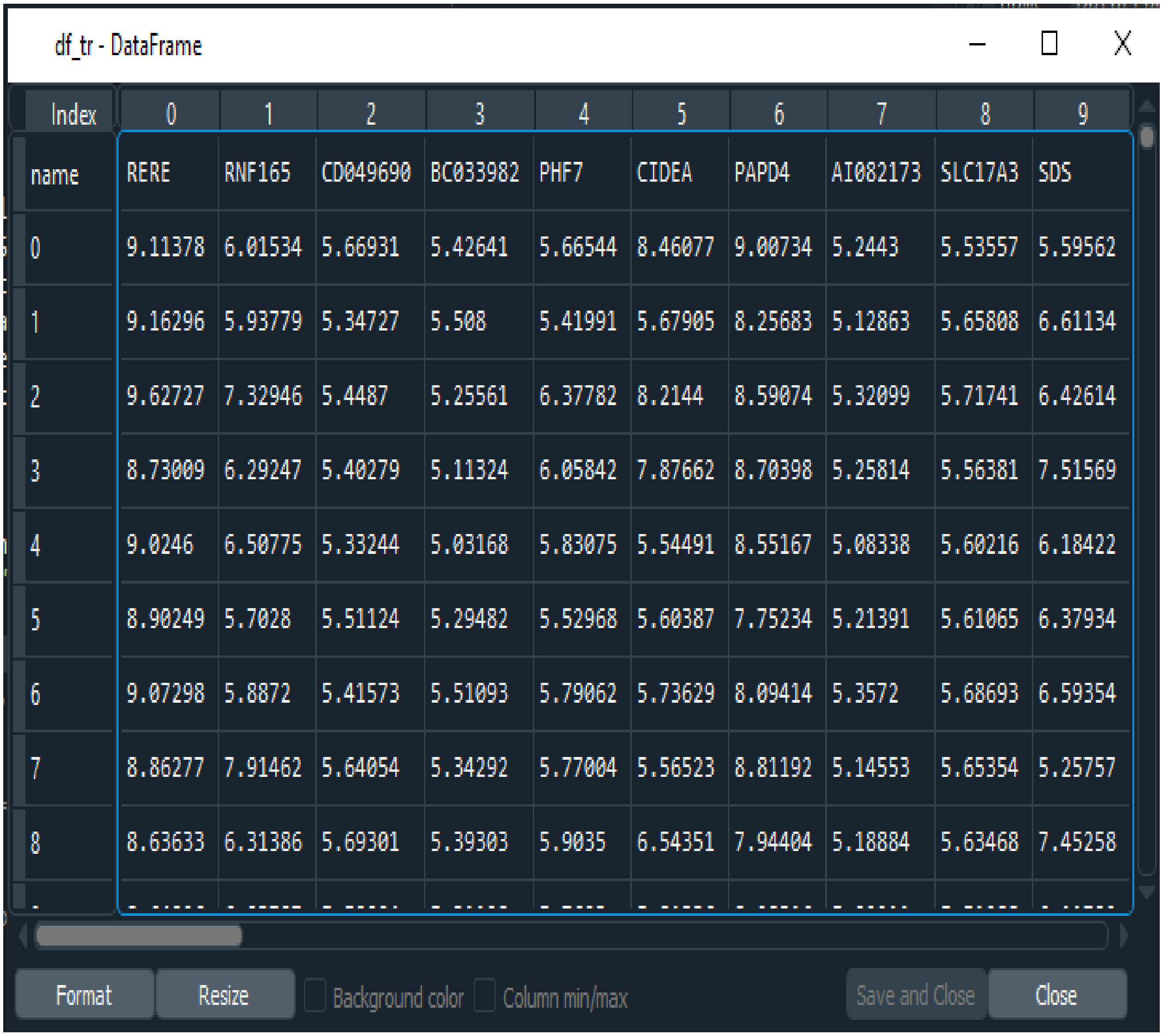
24367	6.640756	6.746047	6.149923	7.260435	7.271969	8.190388	6.428386	7.157397	6.923095	7.550059	8.046706	7.610225	7.736669	7.514208	7.351246	6.036274	5.702
24368	5.464328	5.502217	5.677879	5.263252	5.258665	5.490892	5.226324	5.055417	5.1728	5.510215	5.337706	5.230982	5.214269	5.580688	5.587818	5.507024	5.469
24369	5.435037	5.25386	5.150758	5.426864	5.096231	5.619952	5.310051	5.139505	5.115392	4.919668	5.354445	4.828251	4.667968	5.215902	5.588471	4.959595	5.292
24370	LR	LR	LR	LR	LS	LS	LS	LS	LS	LS	LS	LS	LS	LS	LS	LS	LS
24371																	
24372																	



[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
24366	CB986545	5.251053	5.327294	5.37103	5.234284	4.973229	5.211169	5.656053	5.416202	4.941293	5.329371	5.354759	5.329741	5.350352	5.158665	5.19124	5
24367	IGSF9	6.76775	7.112122	7.080966	6.55463	7.053629	7.443815	6.798448	7.803742	6.902803	6.262165	6.592135	5.582497	6.257378	7.234635	7.454901	6
24368	DA110839	5.251676	5.26272	5.793638	5.296686	5.577535	5.369457	5.469277	5.420184	5.431056	5.454453	5.470135	5.505023	5.194307	5.513542	5.357167	5
24369	FAM71A	5.230243	5.272225	5.381289	5.080102	5.161868	5.013401	5.218734	5.338243	5.640919	5.369637	5.756738	5.24178	5.35039	5.498599	5.198002	5
24370	Class	DH	DH	DH	DH	DH	DH	DH	DR	DR	DR	DR	DR	DR	DR	DR	D

FOURTH...

- Because of machine learning algorithms we used in this project we cannot use this data with this case
- So we had to transpose columns with rows
- Then we checked cell to make sure that no cell is empty to avoid algorithm mistakes
- The clear data
- And this is final Dataset we used after preprocessing.



df_tr - DataFrame

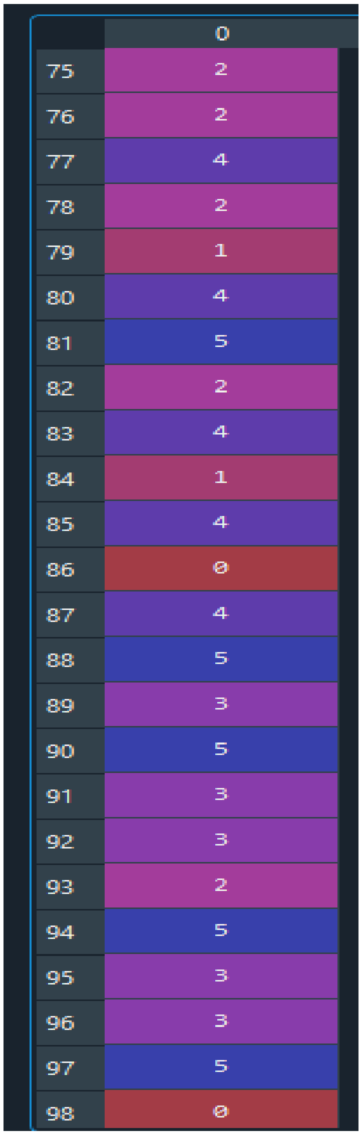
Index	0	1	2	3	4	5	6	7	8	9
name	RERE	RNF165	CD049690	BC033982	PHF7	CIDEA	PAPD4	AI082173	SLC17A3	SDS
0	9.11378	6.01534	5.66931	5.42641	5.66544	8.46077	9.00734	5.2443	5.53557	5.59562
1	9.16296	5.93779	5.34727	5.508	5.41991	5.67905	8.25683	5.12863	5.65808	6.61134
2	9.62727	7.32946	5.4487	5.25561	6.37782	8.2144	8.59074	5.32099	5.71741	6.42614
3	8.73009	6.29247	5.40279	5.11324	6.05842	7.87662	8.70398	5.25814	5.56381	7.51569
4	9.0246	6.50775	5.33244	5.03168	5.83075	5.54491	8.55167	5.08338	5.60216	6.18422
5	8.90249	5.7028	5.51124	5.29482	5.52968	5.60387	7.75234	5.21391	5.61065	6.37934
6	9.07298	5.8872	5.41573	5.51093	5.79062	5.73629	8.09414	5.3572	5.68693	6.59354
7	8.86277	7.91462	5.64054	5.34292	5.77004	5.56523	8.81192	5.14553	5.65354	5.25757
8	8.63633	6.31386	5.69301	5.39303	5.9035	6.54351	7.94404	5.18884	5.63468	7.45258

FINALLY…

We get classes in string and we have to work on float or integer data so we made

Label Encoder:

Here sample of result of label encoder



	0
75	2
76	2
77	4
78	2
79	1
80	4
81	5
82	2
83	4
84	1
85	4
86	0
87	4
88	5
89	3
90	5
91	3
92	3
93	2
94	5
95	3
96	3
97	5
98	0

#0->DH
#1->DR
#2->DS
#3->LH
#4->LR
#5->LS

SUMMARY OF PREPROCESSING AND PSEUDO CODE:

Function:

From data set of 350 breast cancer patient each patient has 24.367 genes we have to get biomarker for each class of 6 classes to get the best treatment from first time

PSEUDO CODE FOR PREPROCESSING:

- Get clinical data set
- Extract only patient took one therapy
- From patient ID get genes from genetic Dataset
- Classify data into 6 classes
    - In Excel
- Label data to make it supervised
- Make label Encoder
    - from sklearn.preprocessing import LabelEncoder

  - label=preprocessing.LabelEncoder()
  - df_tr['Class']=label.fit_transform(df_tr['Class'])
  - dfc=df_tr

- Transpose columns with rows
  - name =pd.read_csv("dd.csv")
  - df_tr = name.transpose()
  - new_header = df_tr.iloc[0] #grab the first row for the header
  - df_tr = df_tr[1:] #take the data less the header row
  - df_tr.columns = new_header #set the header row as the df header
  - df_tr.rename(columns=df_tr.iloc[0])

- check NaN in dataset
  - gg=(dfc['FAM71A'] == 0).sum()

Now we are finishing from preprocessing and clean data

4.2) IMPLEMENTATION:

- Based on the available data, only three treatment therapies are covered in this study: surgery, hormone therapy, and radiotherapy.

.

Over-Sampling with Synthetic Data

- Oversampling the minority class by using synthetic data generators.
- Several algorithms are used to achieve this. We used one of the most popular ones, SMOTE

THEN

- The pipeline starts with feature selection methods like Chi-square and information gain (IG) that are applied to limit the number of significant features (genes).

- Using many classifiers for choosing best accuracy like
  - **Naive Bayes**
  - **SVM linear and polynomial**
  - **Random forest**
  - **KNN**
  - **Logistic Regression**
- Finally we used the biomarkers genes that we get as features for our classifiers so we used “Cross validation”

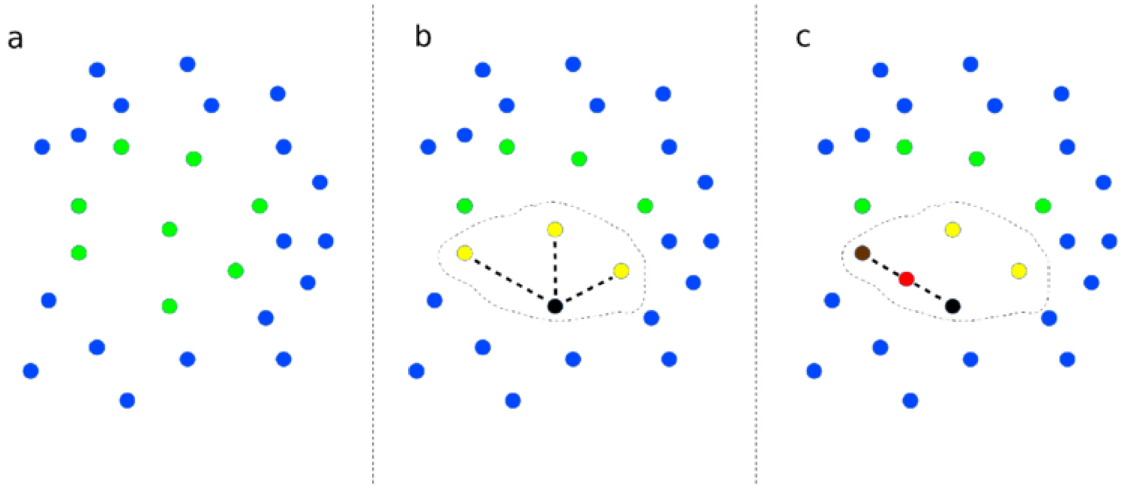
  It is a resampling procedure used to evaluate machine learning models on a limited data sample.

PSEUDO CODE:

- Data we collect is not similar in number of patient so we are facing unbalance data with SMOTE algorithm for oversampling

ALGORITHM DECLARATION:

- **SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem.**
  It aims to balance class distribution by randomly increasing minority class examples by replicating them.
  SMOTE synthesizes new minority instances between existing minority instances



Graphical representation of the SMOTE algorithm a:

- SMOTE starts from a set of positive (green points) and negative (blue points) examples;
- “b”: It then selects a positive example (black) and its k nearest neighbors among the positives (yellow points, with k = 3)
- “c”: Finally one of the k nearest neighbors is randomly selected (brown point) and a new synthetic positive example is added, by randomly generating an example (red point) along the straight line that connects the black and brown points. The procedure depicted in b and c is repeated for all the positives, by adding each time a new synthetic example similar (in an Euclidean sense) to the other positive examples.

PSEUDO CODE

```
from imblearn.over_sampling import SMOTE
x=dfc.iloc[:,0:24368]
y=dfc.iloc[:,24368:24374]
oversample = SMOTE()
x,y=oversample.fit_resample(x, y)
c=x.columns
c=x.columns
```

4.2.1) Chi Sqr Algorithm:

- The dataset we are using consist of almost 24.500 genes as features each gene has mRNA gene Expression so we have to get correlation between each gene and its mRNA gene Exp to get the "Biomarkers" that we will use them as features for our classifier models.
- And we use it for avoiding overfitting

ALGORITHM DECLARATION:

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested. The benefits of performing feature selection before modeling your data are:

- Avoid Overfitting: Less redundant data gives performance boost to the model and results in less opportunity to make decisions based on noise
- Reduces Training Time: Less data means that algorithms train faster

Chi sqr equation:

One common feature selection method that is used with text data is the Chi-Square feature selection. The χ2χ2 test is used in statistics to test the independence of two events. More specifically in feature selection we use it to test whether the occurrence of a specific term and the occurrence of a specific class are independent. More formally, given a document DD, we estimate the following quantity for each term and rank them by their score:

χ2(D,t,c)=∑et∈{0,1}∑ec∈{0,1}(Netec−Eetec)2/Eetec

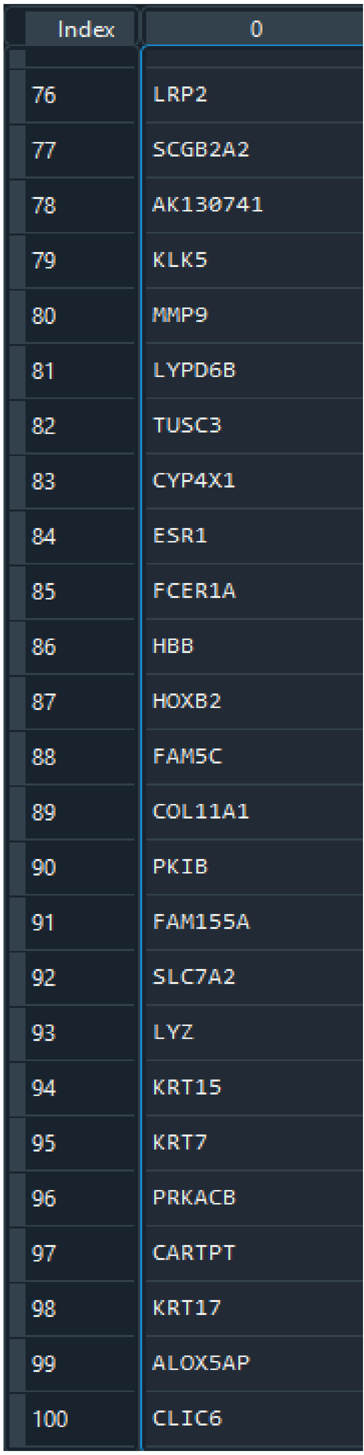
CODE:

- chi2_selector = SelectKBest(chi2, k=100)
- X_kbest = chi2_selector.fit_transform(x, y)
- cols = chi2_selector.get_support(indices=True)
- features_df_new = x.iloc[:,cols]
- c=features_df_new.columns
- x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20)
-
- cv = ShuffleSplit(n_splits=5, random_state=0)

Finally we get Biomarkers
We select only 200 genes as Biomarkers as “Features

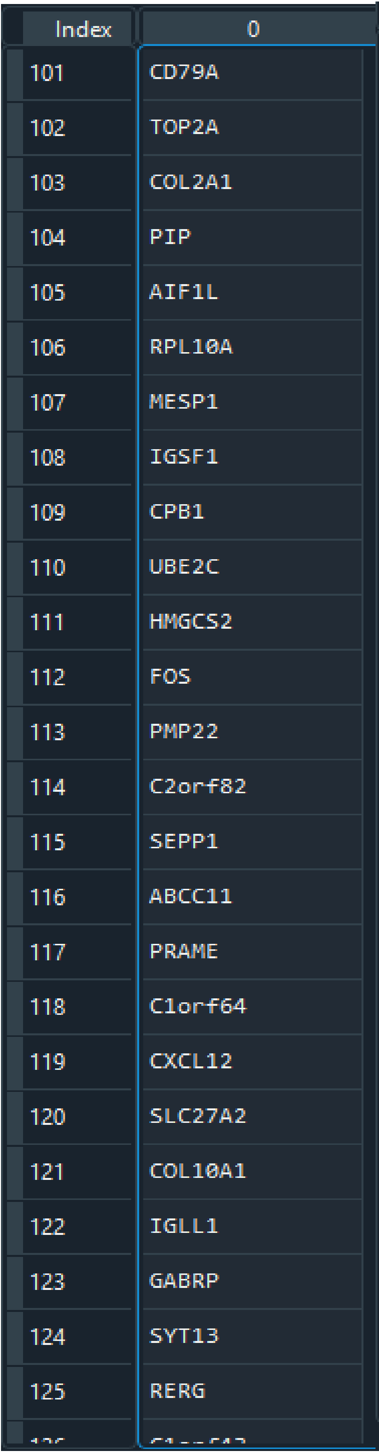
Here some of Biomarkers that we get

c - Index

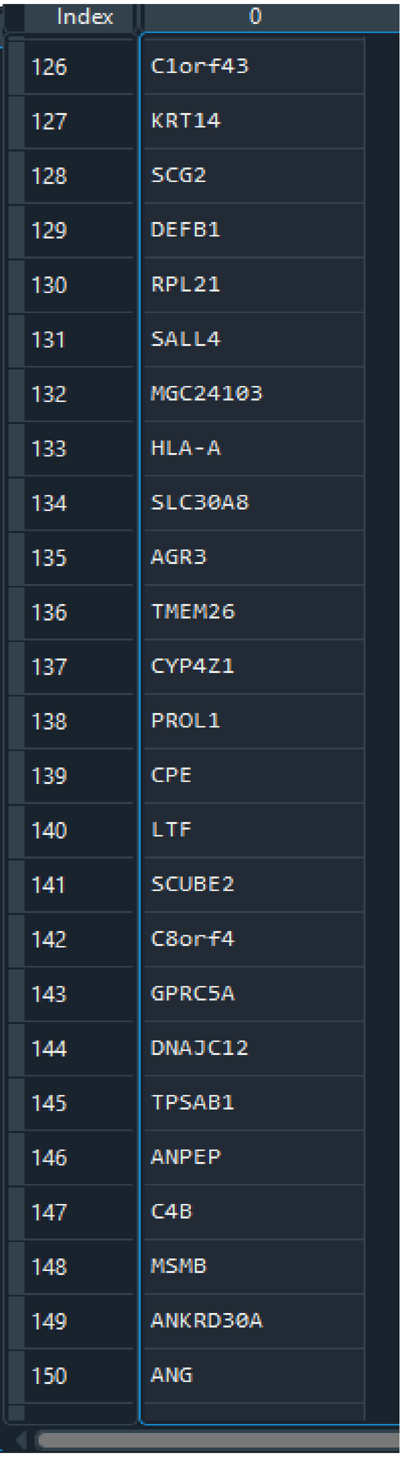


Index	0
76	LRP2
77	SCGB2A2
78	AK130741
79	KLK5
80	MMP9
81	LYPD6B
82	TUSC3
83	CYP4X1
84	ESR1
85	FCER1A
86	HBB
87	HOXB2
88	FAM5C
89	COL11A1
90	PKIB
91	FAM155A
92	SLC7A2
93	LYZ
94	KRT15
95	KRT7
96	PRKACB
97	CARTPT
98	KRT17
99	ALOX5AP
100	CLIC6

c - Index



Index	0
101	CD79A
102	TOP2A
103	COL2A1
104	PIP
105	AIF1L
106	RPL10A
107	MESP1
108	IGSF1
109	CPB1
110	UBE2C
111	HMGCS2
112	FOS
113	PMP22
114	C2orf82
115	SEPP1
116	ABCC11
117	PRAME
118	C1orf64
119	CXCL12
120	SLC27A2
121	COL10A1
122	IGLL1
123	GABRP
124	SYT13
125	RERG
126	C1orf43



Index	0
126	C1orf43
127	KRT14
128	SCG2
129	DEFB1
130	RPL21
131	SALL4
132	MGC24103
133	HLA-A
134	SLC30A8
135	AGR3
136	TMEM26
137	CYP4Z1
138	PROL1
139	CPE
140	LTF
141	SCUBE2
142	C8orf4
143	GPRC5A
144	DNAJC12
145	TPSAB1
146	ANPEP
147	C4B
148	MSMB
149	ANKRD30A
150	ANG

4.2.2) VarianceThreshold

We have to implement another feature selection algorithm to get the best correlations between genes ((Biomarkers))

ALGORITHM DECLARATION:

Feature selector that removes all low-variance features.

Goal

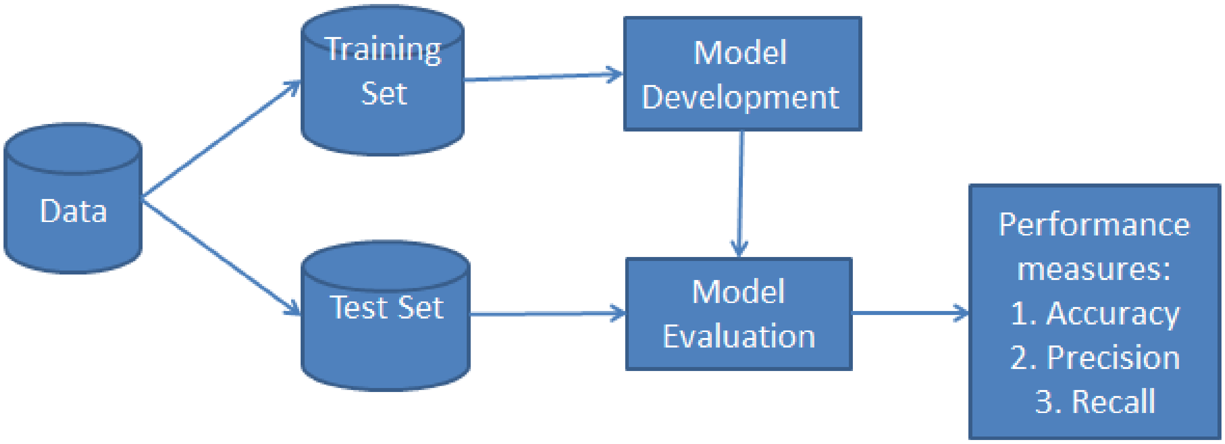
The goal of feature selection in machine learning is to find the best set of features that allows one to build useful models of studied phenomena.

Supervised Techniques: These techniques can be used for labeled data, and are used to identify the relevant features for increasing the efficiency of supervised models like classification and regression.

PSEUDO CODE:

```
selector =  VarianceThreshold(0.6)
selector.fit(x)
x[x.columns[selector.get_support(indices=True)]]
dfc=pd.concat([y,x],axis=1)
DF=pymrmr.mRMR(dfc,'MIQ',100)
fe=x[DF]
cv=fe.columns
x_train, x_test, y_train, y_test = train_test_split(fe, y, test_size=0.20)
```

Now we can build classifiers models with 200 Features



“CLASSIFIER NO. 1”

➢ Random forest:

- Random forest is a supervised learning algorithm.
- The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method.
- The general idea of the bagging method is that a combination of learning models increases the overall result.

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Why use Random Forest?

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Code for Random Forest:

- RandomForestClassifier(n_estimators=1000, class_weight='balanced')
- clf.fit(x_train, y_train)
- scores = cross_val_score(clf, x_train, y_train, cv=cv)
- print("%0.2f accuracy RandomForestClassifier with a standard deviation of %0.2f 1 :   " % (scores.mean(), scores.std()))
- Randomforce_pred=clf.predict(x_test)
- print("accuaracy RandomForestClassifier:",accuracy_score(y_test, Randomforce_pred))

ENTROPY IN RANDOM FOREST:

What is Entropy?

In the most layman terms, Entropy is nothing but the **measure of disorder.** (You can think of it as a measure of purity as well.

The Mathematical formula for Entropy is as follows -

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

- Where ‘Pi’ is simply the frequentist probability of an element/class ‘i’ in our data.
-

  Entropy is a measure of disorder or uncertainty and the goal of machine learning models and Data Scientists in general is to reduce uncertainty

“CLASSIFIER NO. 2”

NAÏVE BAYES GAUSSIANNB:

What is Naive Bayes Classifier?

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Attributes

*class_prior_*	array, shape = [n_classes]	Probability of each class.
*theta_*	array, shape = [n_classes, n_features]	mean of each feature per class
*sigma_*	array, shape = [n_classes, n_features]	variance of each feature per class

CODE FOR NAÏVE BAYES GAUSSIANNB:

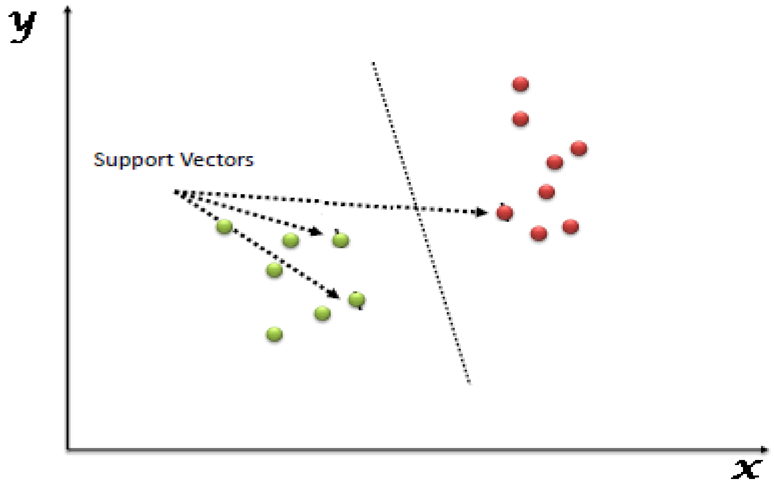
- from sklearn.naive_bayes import GaussianNB
- model = GaussianNB()
- model.fit(x_train, y_train)
- scores = cross_val_score(model, x_train, y_train, cv=cv)
- print("%0.2f accuracy GaussianNB with a standard deviation of %0.2f 5: " % (scores.mean(), scores.std()))
- GaussianNB_pred=model.predict(x_test)
- print("accuaracyGaussianNB ",accuracy_score(y_test, GaussianNB_pred)

"CLASSIFIER NO. 3"

"SVM LINEAR CLASSIFIER"

What is Support Vector Machine?

"Support Vector Machine" (SVM) is a supervised machine_learning_algorithm which can be used for either classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).



CODE FOR SVM LINEAR:

- rbf = svm.SVC(kernel='linear', C=0.1, class_weight='balanced').fit(x_train, y_train)
- scores = cross_val_score(rbf, x_train, y_train, cv=cv)
- print("%0.2f accuracy svmlinear with a standard deviation of %0.2f6: " % (scores.mean(), scores.std()))
- svm_pred=rbf.predict(x_test)
- print("accuaracysvmlinear ",accuracy_score(y_test, svm_pred))

“CLASSIFIER NO. 4”

“SVM POLYNOMIAL KERNEL”

In general, the polynomial kernel is defined as ;

$$K(X_1, X_2) = (a + X_1^T X_2)^b$$

b = degree of kernel & a = constant term.

in the polynomial kernel, we simply calculate the dot product by increasing the power of the kernel.

Example:
Let’s say originally X space is 2-dimensional such that
Xa = (a1 ,a2)
Xb = (b1 ,b2)
now if we want to map our data into higher dimension let’s say in Z space which is six-dimensional it may seem like

$$Z_a = \phi(X_a) = (1, a_1, a_2, a_1^2, a_2^2, a_1 * a_2)$$

$$Z_b = \phi(X_b) = (1, b_1, b_2, b_1^2, b_2^2, b_1 * b_2)$$

In order to solve this dual SVM we would require the dot product of (transpose) Za ^t and Zb.

Using kernel trick:

$$Z_a^T Z_b = k(X_a, X_b) = (1 + X_a^T X_b)^2$$
$$Z_a^T Z_b = 1 + a_1 b_1 + a_2 b_2 + a_1^2 b_1^2 + a_2^2 b_2^2 + a_1 b_1 a_2 b_2$$

In this method, we can simply calculate the dot product by increasing the value of power. Simple isn't it?

Code for SVM Polynomial Kernel:

- rbf = svm.SVC(kernel='poly', C=1.0 ,class_weight='balanced')
- rbf.fit(x_train, y_train)
- scores = cross_val_score(rbf, x_train, y_train, cv=cv)
- print("%0.2f accuracy svmpoly with a standard deviation of %0.2f6: " % (scores.mean(), scores.std()))
- svm_pred=rbf.predict(x_test)
- print("accuaracysvmpoly ",accuracy_score(y_test, svm_pred))

“CLASSIFIER NO.5”

“KNN”

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry so we used it in our project

K is a crucial parameter in the KNN algorithm. Some suggestions for choosing K Value are:

1. Using error curves:

2. Also, domain knowledge is very useful in choosing the K value.

3. K value should be odd while considering binary (two-class) classification.

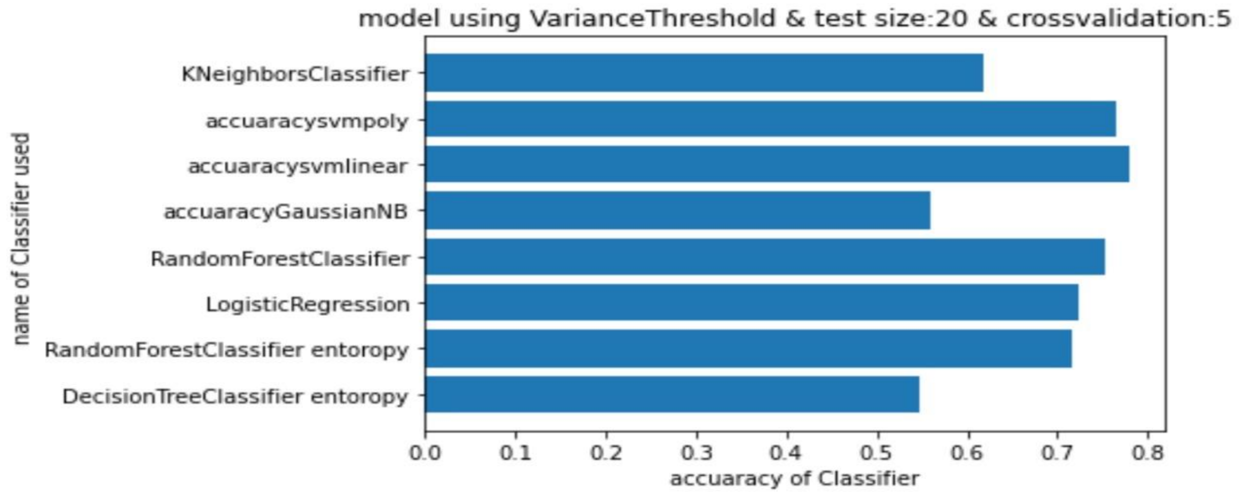
CODE:

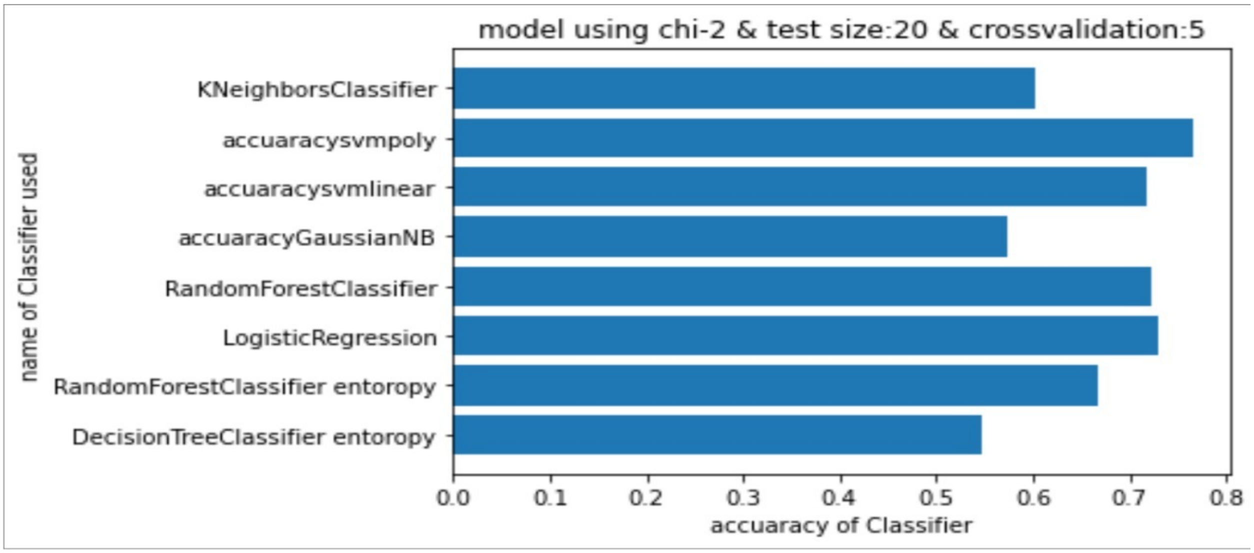
- from sklearn.neighbors import KNeighborsClassifier
- model = KNeighborsClassifier(n_neighbors=3)
- model.fit(x_train, y_train)
- scores = cross_val_score(model, x_train, y_train, cv=cv)
- print("%0.2f accuracy KNeighborsClassifier with a standard deviation of %0.2f 2: " % (scores.mean(), scores.std()))
- KNeighborsClassifier_pred=model.predict(x_test)
- print("accuaracyKNeighborsClassifier ",accuracy_score(y_test, KNeighborsClassifier_pred))

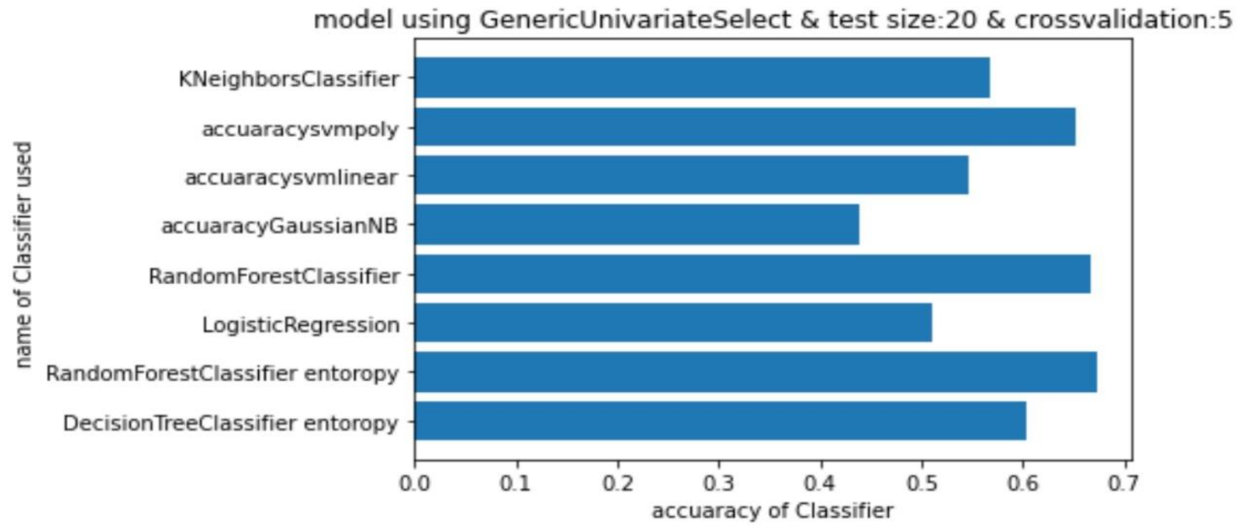
4.3) ACCURACY

	Chi sqr	Variance threshold	Generic Univariate Select
SVM POLY	76.5%	87.01%	65.7%
SVM LINEAR	71.4%	87.01%	54.6%
RANDOM FOREST	72.4%	73.78%	66.7%
LOGISTIC REGRESSION	73.3%	75.8%	51.7%
KNN	60.3%	62.4%	65.7%
RANDOM FOREST ENTROPY	66.7%	71.6%	67.7%
DECISION TREE ENTROPY	54.6%	60.2%	60.5%
GAUSSIANNB	58.0%	58.9%	43.9%

4.4) CHARTS


model using VarianceThreshold & test size:20 & crossvalidation:5
KNeighborsClassifier
accuaracysvmpoly
accuaracysvmlinear
accuaracyGaussianNB
RandomForestClassifier
LogisticRegression
RandomForestClassifier entoropy
DecisionTreeClassifier entoropy
name of Classifier used
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
accuaracy of Classifier

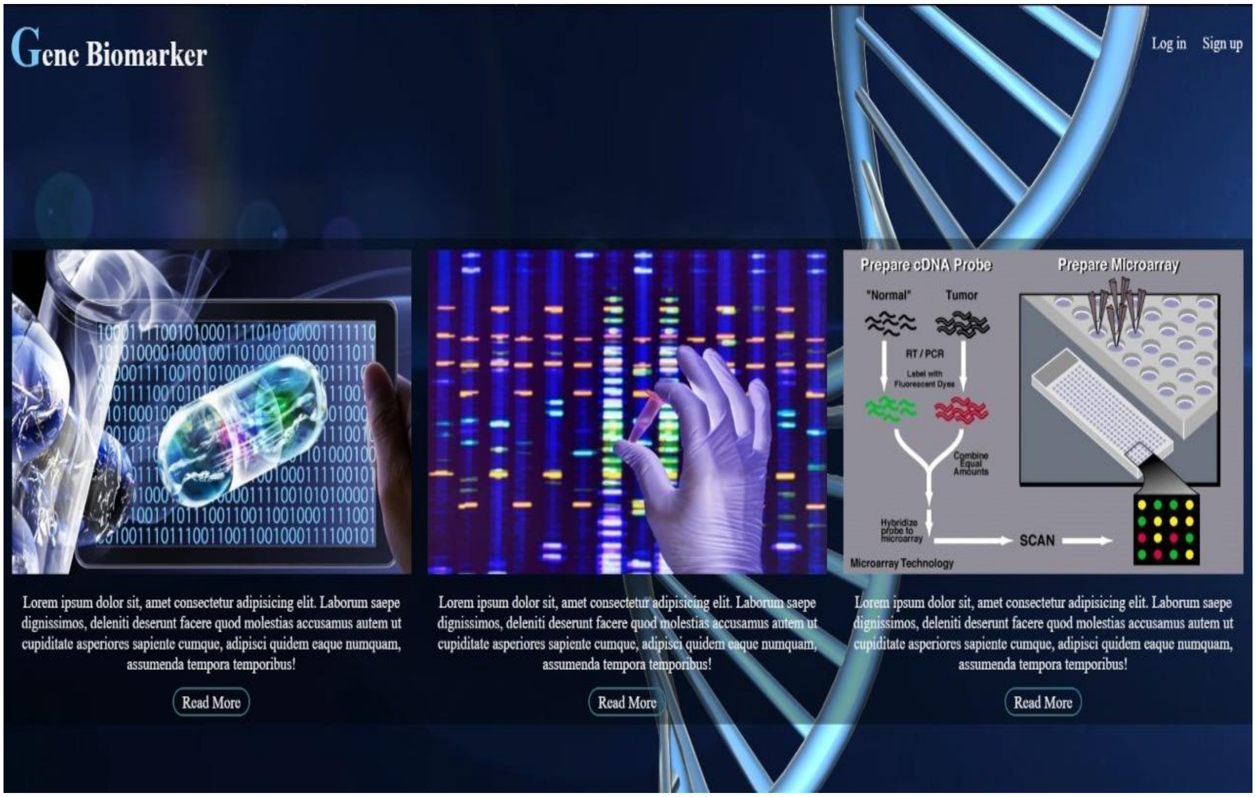

model using chi-2 & test size:20 & crossvalidation:5
KNeighborsClassifier
accuaracysvmpoly
accuaracysvmlinear
accuaracyGaussianNB
RandomForestClassifier
LogisticRegression
RandomForestClassifier entoropy
DecisionTreeClassifier entoropy
name of Classifier used
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
accuaracy of Classifier


model using GenericUnivariateSelect & test size:20 & crossvalidation:5
KNeighborsClassifier
accuaracysvmpoly
accuaracysvmlinear
accuaracyGaussianNB
RandomForestClassifier
LogisticRegression
RandomForestClassifier entoropy
DecisionTreeClassifier entoropy
name of Classifier used
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
accuaracy of Classifier

CHAPTER 5

WEB APP DESCRIPTION:

At the beginning of web home page we will find this page



Lorem ipsum dolor sit, amet consectetur adipisicing elit. Laborum saepe dignissimos, deleniti deserunt facere quod molestias accusamus autem ut cupiditate asperiores sapiente cumque, adipisci quidem eaque numquam, assumenda tempora temporibus!

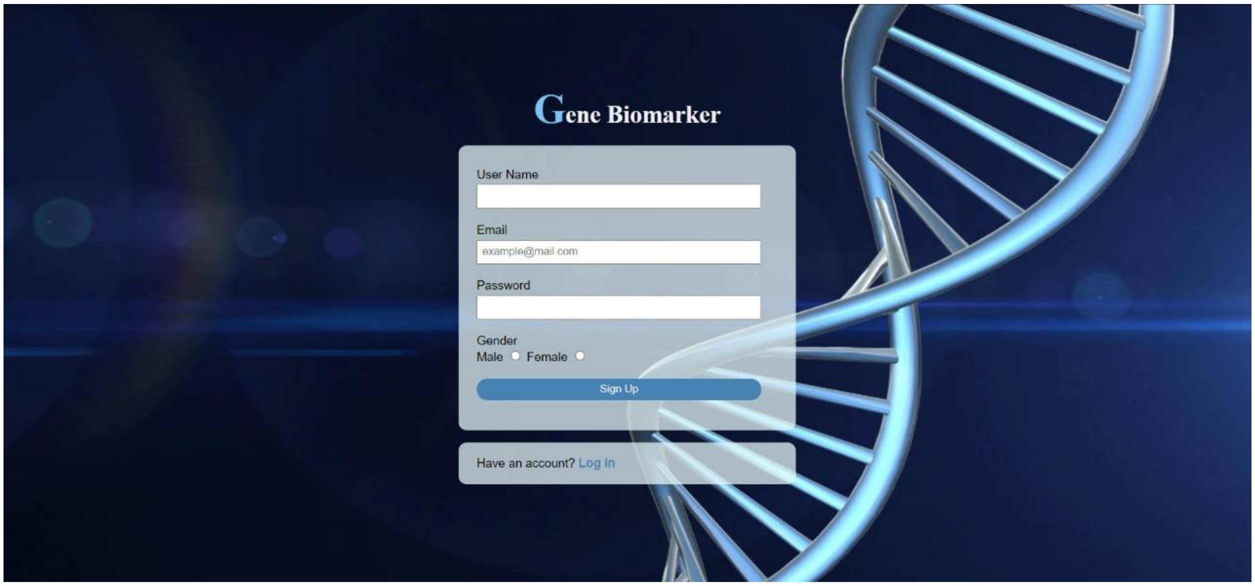
Lorem ipsum dolor sit, amet consectetur adipisicing elit. Laborum saepe dignissimos, deleniti deserunt facere quod molestias accusamus autem ut cupiditate asperiores sapiente cumque, adipisci quidem eaque numquam, assumenda tempora temporibus!

Lorem ipsum dolor sit, amet consectetur adipisicing elit. Laborum saepe dignissimos, deleniti deserunt facere quod molestias accusamus autem ut cupiditate asperiores sapiente cumque, adipisci quidem eaque numquam, assumenda tempora temporibus!

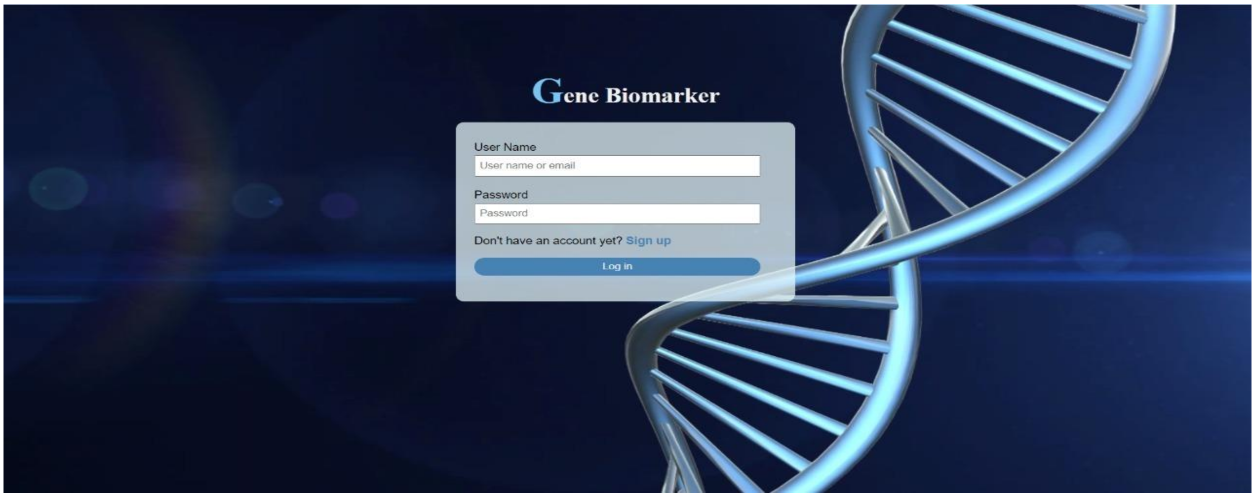
When user start using the app will find home page has a lot of information about gene expression, gene biomarkers and medical information.

At top left of page will find 2 important buttons to start using the app

- **The first is sign up used to save information about user**



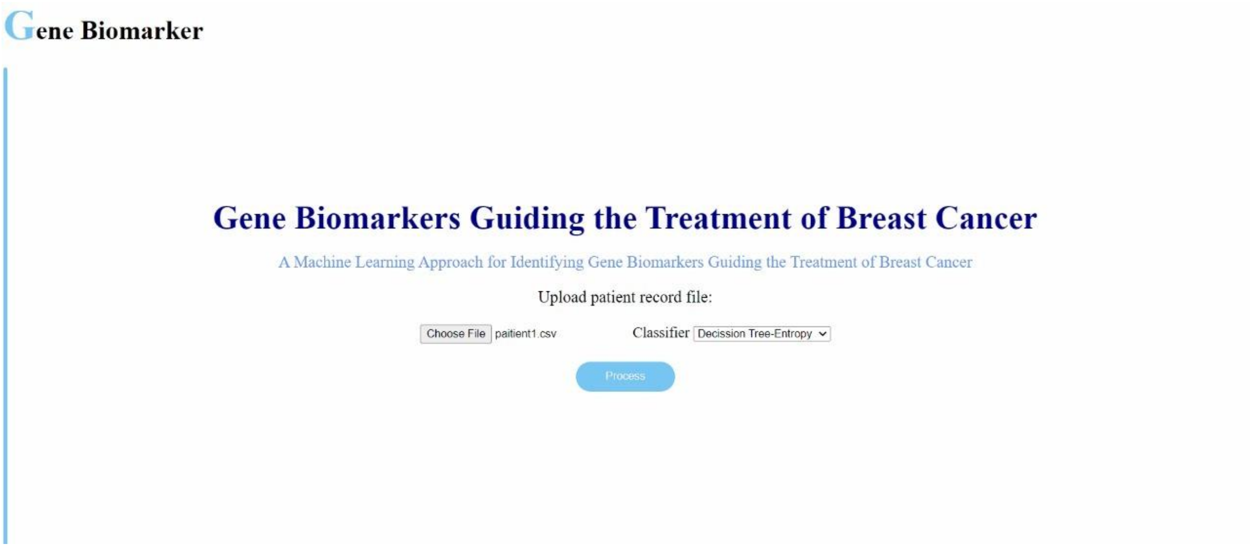
- **Second is log in used to enter well known users**



THE MAIN MACHINE LEARNING WEB PAGE:

1) Based on our project we will take from user the Genomic profile as “CSV file” from button choose file this button will make us choose “CSV” file from our files.
2) Then Doctor will choose the suitable ML classifier from 8 classifiers we implemented
3) Finally the machine learning project will render the suitable treatment based on genomic profile and dataset and render its accuracy

RUN FILE:



Gene Biomarkers Guiding the Treatment of Breast Cancer

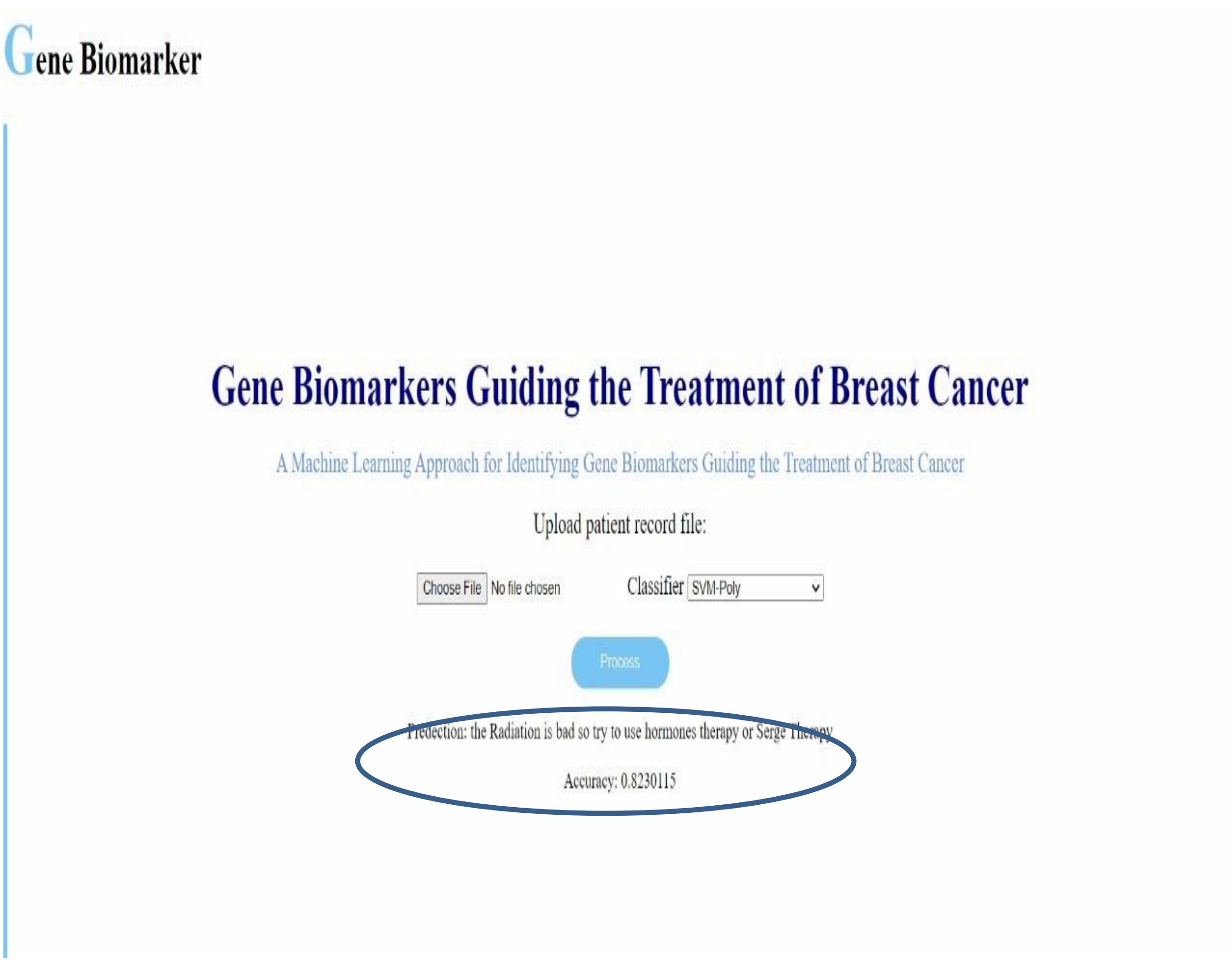
A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer

Upload patient record file:

Choose File paitient1.csv

Classifier Decission Tree-Entropy

SHOW RESULT:



Gene Biomarker

Gene Biomarkers Guiding the Treatment of Breast Cancer

A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer

Upload patient record file:

No file chosen

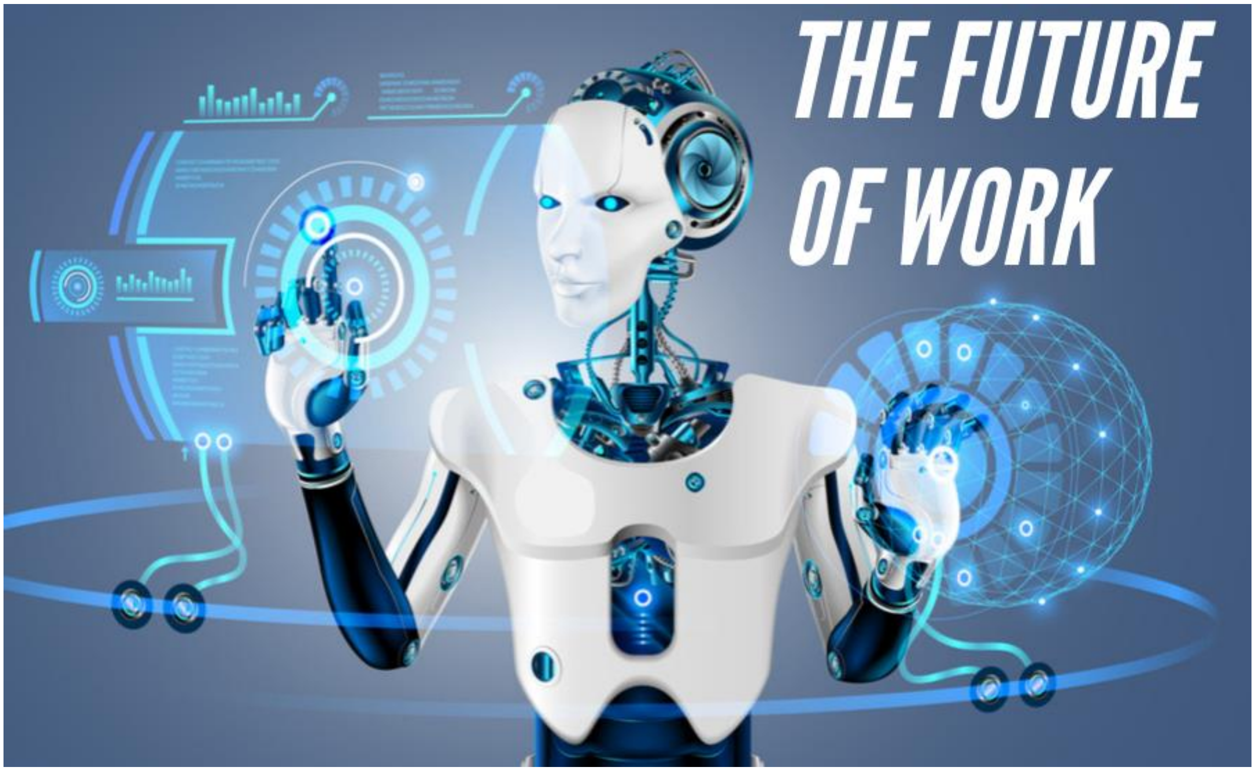
Classifier SVM-Poly

Predection: the Radiation is bad so try to use hormones therapy or Serge Therapy

Accuracy: 0.8230115

CHAPTER 6

CONCLUSION AND FUTURE WORK



6.1CONCLUSION

The goal for this project is finding right treatment for specific breast cancer patient based on previous survived patients who take only one therapy and (live or dead because of cancer disease).

We developed our project using machine learning techniques and algorithms that helped us too much to get Goal.

At the first we was looking for biomarker genes which have correlation so we used "Chi sqr feature selection model" & "Generic Univariate Select"

&" Variance Threshold " then we compared the final accuracy of all of them and we took the best model that has the best accuracy.

Then we implement a lot of ML classifiers like →SVM Linear, SVM polynomial, Random forest, KNN and Naive Bayes and get accuracy of all then dependent on model which has highest accuracy.

Now we have model can predict specific treatment for specific breast cancer patient based on his/her “Genomic profile”.

6.2 FUTURE WORK:

- Achieve high accuracy.
- Search in all cancer types not only breast cancer.
- Collect dataset contain huge number of patients who took only one Therapy.
- Test our project in real life.
- The project experience is supervised by a group of doctors and they correct our information about cancer for us so that we can complete the project and actually apply it.
- Use clinical data next to genomic data to be most accurate with each patient.

CHAPTER 7

REFERENCE:

1. Tsimberidou AM, Iskander NG, Hong DS, et al. Personalized medicine in a phase I clinical trials program: the MD Anderson Cancer Center Initiative. Clin Cancer Res. 2012;18(22):6373-83.
2. Von Hoff DD, Stephenson JJ Jr, Rosen P, et al. Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. J Clin Oncol. 2010;28(33):4877-83.
3. Why Are Cancer Rates Increasing? London, UK: Cancer Research UK; 2015. http://scienceblog.cancerresearchuk.org/2015/02/04/why-are-cancer-rates-increasing/. Accessed December 10, 2015
4. Abou Tabl, A., Alkhateeb, A., ElMaraghy, W., and Ngom, A. (2017). "Machine learning model for identifying gene biomarkers for breast cancer treatment survival," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, (New York, NY: ACM), 607. doi: 10.1145/3107411.3108217
5. PubMed Abstract | CrossRef Full Text | Google Scholar
6. Allegra, A., Alonci, A., Campo, S., Penna, G., Petrungaro, A., Gerace, D., et al. (2012). Circulating micrornas: new biomarkers in diagnosis, prognosis and treatment of cancer. *Int. J. Oncol.* 41, 1897–1912. doi: 10.3892/ijo.2012.1647
7. PubMed Abstract | CrossRef Full Text | Google Scholar
8. Bamberger, A. M., Methner, C., Lisboa, B. W., Städtler, C., Schulte, H. M., Löning, T., et al. (1999). Expression pattern of the ap-1 family in breast cancer: association of fosb expression with a well-differentiated, receptor-positive tumor phenotype. *Int. J. Cancer* 84, 533–538. doi: 10.1002/(SICI)1097-0215(19991022)84:5<533::AID-IJC16>3.0.CO;2-J
9. PubMed Abstract | CrossRef Full Text | Google Scholar

10.Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

11. CrossRef Full Text | Google Scholar

12.

SMOTE Algorithms declaration

Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol.30, 1095–1106 (2012).2. Veltman, J. A. & Lupski, J. R. From genes to genomes in the clinic. Genome Med. 7, 78 (2015).3.Ritchie, G. & Flicek, P. Functional Annotation of Rare Genetic Variants in Assessing Rare Variation in Complex Traits (ed.Zeggini, E. & Morris, A.) 57–70 (Springer New York, 2015

13 scikit learn

Data Availability

Publicly available datasets were analyzed in this study. This data can be found here: http://www.cbioportal.org/study?id=brca_metabric.

Finally, we wish a speedy recovery to all patients, and we hope that we have contributed their recovery, with what we have learned at the hands of our professors.