Breast Cancer Metastasis

Multi Class-Classification

1st Faisal Riaz
School of Business
UMT
Lahore, Pakistan
f2019313020@umt.edu.pk

2nd Muhammad Nauman
School of Business
UMT
Lahore, Pakistan
f2019313009@umt.edu.pk

3rd Talha Rasool
School of Business
UMT
Lahore, Pakistan
f2019313039@umt.edu.pk

***Index Terms*—Keywords: Breast cancer, Metastasis, Tumours, Gene Expression Omnibus (GEO), GEOparse**

I. Introduction

Cancer is a group of body cells that grow and proliferate abnormally and uncontrollably because of damaged DNA (deoxyribonucleic acid). This group of body cells, known as tumors and breast cancer originates in a breast tissue. It is the most frequently diagnosed cancer among women, and it is 100 times more common in women than in men.Across the globe, the breast cancer is the second major cause of female deaths resulting from cancer. There is no known way to prevent breast cancer,but mortality rate can be reduced with the help of predictions of early diagnosis symptoms of breast cancer. Moreover,Metastasis is the spread of cancer cells to new areas of the body, often by way of the lymph system or bloodstream. A metastatic cancer, or metastatic tumor, is one that has spread from the primary site of origin, or where it started, into different areas of the body which are called secondary tumor.For instance (Bone tumor,Brain Tumor, Lung tumor, Liver tumor).

Cancer metastasis is the spread of cancer cells to tissues and organs beyond where the tumor originated and the formation of new tumors (secondary and tertiary foci) is the single event that results in the death of most patients with cancer. Subsequently, metastasis is the most dangerous occasion in patients with malignant growth. The procedure is made out of various consecutive occasions which must be finished all together for the tumor cell to effectively metastasize, the supposed metastatic course.

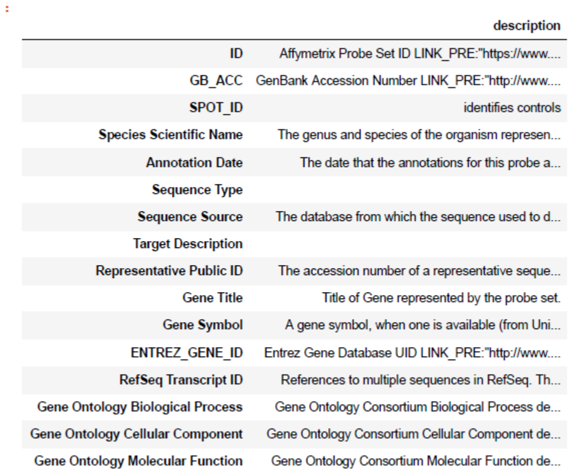
This procedure adds to the multifaceted nature of malignancy as a multiplex illness. During the metastatic course, changes in cell-cell and cell-lattice grip are significant.In this study we will discuss about Gene Expression Omnibus (GEO). The GEO is initiated by National Center for Biotechnology Information (NCBI) in 1999.It has an adaptable and open structure that permits the accommodation,stockpiling, and recovery of numerous kinds of informational collections.

II. Methods

In this study, We discussed the methods to solve the GEO data set. Every data set is nominated with the accession number. The accession number GSE 14020.

A. Data Details

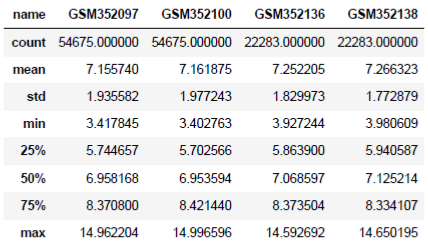
In that project we Geoparse library use because it facilitate the researchers in genome studies and allows downloading and loading the SOFT files from the Gene Expression Omnibus database The data is loaded in easily digestible data structures and analyze the Sample Data of **GPL570**.which has been shown in figure 1.



	description
ID	Affymetrix Probe Set ID LINK_PRE:"https://www....
GB_ACC	GenBank Accession Number LINK_PRE:"http://www....
SPOT_ID	identifies controls
Species Scientific Name	The genus and species of the organism represen...
Annotation Date	The date that the annotations for this probe a...
Sequence Type	
Sequence Source	The database from which the sequence used to d...
Target Description	
Representative Public ID	The accession number of a representative seque...
Gene Title	Title of Gene represented by the probe set.
Gene Symbol	A gene symbol, when one is available (from Uni...
ENTREZ_GENE_ID	Entrez Gene Database UID LINK_PRE:"http://www....
RefSeq Transcript ID	References to multiple sequences in RefSeq. Th...
Gene Ontology Biological Process	Gene Ontology Consortium Biological Process de...
Gene Ontology Cellular Component	Gene Ontology Consortium Cellular Component de...
Gene Ontology Molecular Function	Gene Ontology Consortium Molecular Function de...

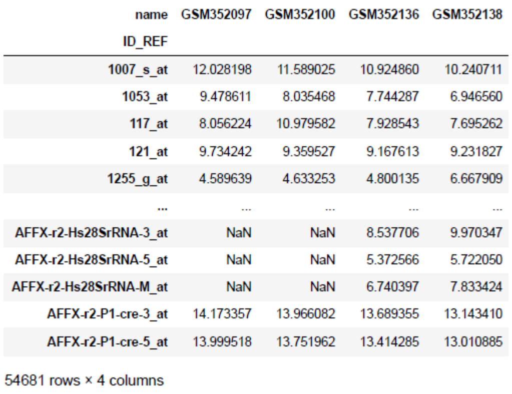
Fig:1 GPL Description

We took four samples to each from GPL570 and GPL96. Which describe the overall detail of descriptive data at we have to take care of.Samples from GPL570 and GPL96 has 54675 and 22283 genes respectively.Detail has been shown in Fig:2 and Fig:3.



name	GSM352097	GSM352100	GSM352136	GSM352138
count	54675.000000	54675.000000	22283.000000	22283.000000
mean	7.155740	7.161875	7.252205	7.266323
std	1.935582	1.977243	1.829973	1.772879
min	3.417845	3.402763	3.927244	3.980609
25%	5.744657	5.702566	5.863900	5.940587
50%	6.958168	6.953594	7.068597	7.125214
75%	8.370800	8.421440	8.373504	8.334107
max	14.962204	14.996596	14.592692	14.650195

Fig:2 Sample Descriptive



name	GSM352097	GSM352100	GSM352136	GSM352138
ID_REF				
1007_s_at	12.028198	11.589025	10.924860	10.240711
1053_at	9.478611	8.035468	7.744287	6.946560
117_at	8.056224	10.979582	7.928543	7.695262
121_at	9.734242	9.359527	9.167613	9.231827
1255_g_at	4.589639	4.633253	4.800135	6.667909
...	...	...	...	...
AFFX-r2-Hs28SrRNA-3_at	NaN	NaN	8.537706	9.970347
AFFX-r2-Hs28SrRNA-5_at	NaN	NaN	5.372566	5.722050
AFFX-r2-Hs28SrRNA-M_at	NaN	NaN	6.740397	7.833424
AFFX-r2-P1-cre-3_at	14.173357	13.966082	13.689355	13.143410
AFFX-r2-P1-cre-5_at	13.999518	13.751962	13.414285	13.010885

54681 rows × 4 columns

Fig:3 Gene Probs and Sample Data

Now, we have shown data distribution of 4 controlled sample in fig:4 in whihch Histogram shows that samples belongs to both GPLs are left skewed. Highest probs values are lying in between bin628 reaching 12000 genes and 5000 thousand for GPL507 and GPL96 in fig:4.

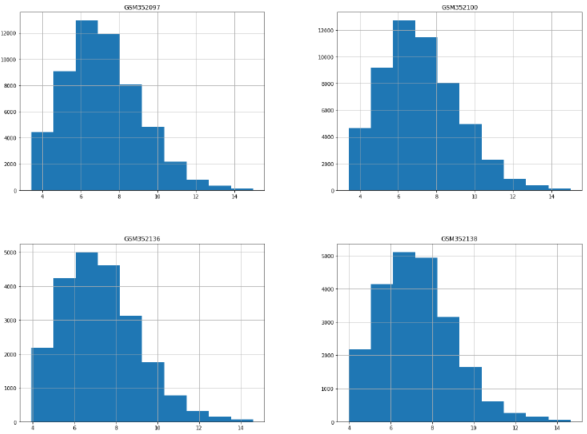
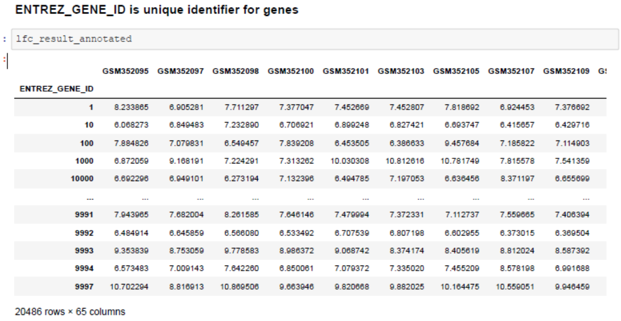


Fig:4 Samples Distribution

B. Data Transformation

Data normalization is a crucial preliminary step in analyzing genomic data sets. The goal of normalization is to remove global variation to make readings across different experiments comparable. In addition, most gnomic loci have non-uniform sensitivity to any given assay because of variation in local sequence properties. In micro array experiments, this non-uniform sensitivity is due to different DNA hybridization and cross-hybridization efficiencies, known as the probe effect. Now,After successful transformation of data and extracting metastasis data against each of 65 samples and from phenotype data as shown in fig:5 Metastasis data for each Sample and combined into one table and shape of transformed data achieved is **20486 rows × 65 columns** Matrix.



ENTREZ_GENE_ID is unique identifier for genes

```
: lfc_result_annotated
```

	GSM352095	GSM352097	GSM352098	GSM352100	GSM352101	GSM352103	GSM352105	GSM352107	GSM352109	G!
ENTREZ_GENE_ID										
1	8.233865	6.905281	7.711297	7.377047	7.452669	7.452807	7.818692	6.924453	7.376692	
10	6.068273	6.849483	7.232890	6.706921	6.899248	6.827421	6.693747	6.415657	6.429716	
100	7.884826	7.079831	6.549457	7.839208	6.453505	6.386633	9.457684	7.185822	7.114903	
1000	6.872059	9.168191	7.224291	7.313262	10.030308	10.812616	10.781749	7.815578	7.541359	
10000	6.692296	6.949101	6.273194	7.132396	6.494785	7.197053	6.636456	8.371197	6.655699	
...	...	...	...	...	...	...	...	...	...	
9991	7.943965	7.682004	8.261585	7.646146	7.479994	7.372331	7.112737	7.559665	7.406394	
9992	6.484914	6.645859	6.566080	6.533492	6.707539	6.807198	6.602955	6.373015	6.369504	
9993	9.353839	8.753059	9.778583	8.986372	9.068742	8.374174	8.405619	8.812024	8.587392	
9994	6.573483	7.009143	7.642260	6.850061	7.079372	7.335020	7.455209	8.578198	6.991688	
9997	10.702294	8.816913	10.869506	9.663946	9.820668	9.882025	10.164475	10.559051	9.946459	

20486 rows × 65 columns

Fig:5 ENTREZ GENE ID

Final shape of transformed data is **X[x1,x2,x3,........Xn] [y]** X and y. Metastasis distribution for 65 samples is shown in fig:6.

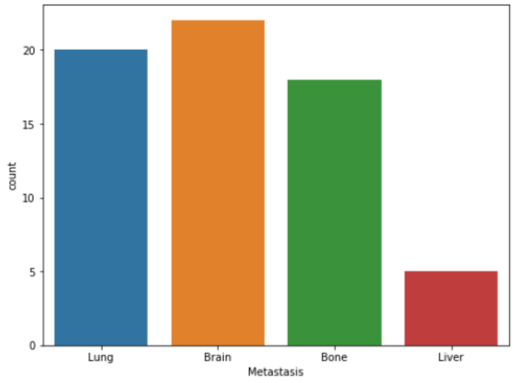


Fig:6 Metastasis Distribution

C. Missing Values Imputation

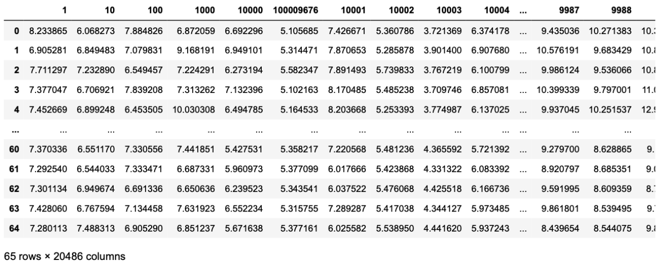
In this Dataset,We have used KNN imputation technique for imputing NAN values as described in- [3] The philosophy behind using KNN imputation is missing values are imputed based on N neighbours with respect to the mean of the euclidean distance of neighbours from the missing values location. Moreover,we have used number of neighbours=10, weight of each neighbour is governed by the inverse distance.

D. Normalization/ Standardization

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a **mean of zero and a standard deviation of 1**. This standardization is called a z-score, and data points can be standardized with the following formula:

$$Z_i = \frac{x_i - \bar{x}}{S} \tag{1}$$

Where: x_i is a data point $(x_1, x_2 \ldots x_n)$. $\bar{x}$ is the sample mean. **S** is the sample standard deviation. We have used standard scalar library from **scikit-learn**.



	1	10	100	1000	10000	100009676	10001	10002	10003	10004	...	9987	9988	
0	8.233865	6.068273	7.884826	6.872059	6.692296	5.105685	7.426671	5.360786	3.721369	6.374178	...	9.435036	10.271383	10.
1	6.905281	6.849483	7.079831	9.168191	6.949101	5.314471	7.870653	5.285878	3.901400	6.907680	...	10.576191	9.683429	10.
2	7.711297	7.232890	6.549457	7.224291	6.273194	5.582347	7.891493	5.739833	3.767219	6.100799	...	9.986124	9.536066	10.
3	7.377047	6.706921	7.839208	7.313262	7.132396	5.102163	8.170485	5.485238	3.709746	6.857081	...	10.399339	9.797001	11.
4	7.452669	6.899248	6.453505	10.030308	6.494785	5.164533	8.203668	5.253393	3.774987	6.137025	...	9.937045	10.251537	12.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
60	7.370336	6.551170	7.330556	7.441851	5.427531	5.358217	7.220568	5.481236	4.365592	5.721392	...	9.279700	8.628865	9.
61	7.292540	6.544033	7.333471	6.687331	5.960973	5.377099	6.017666	5.423868	4.331322	6.083392	...	8.920797	8.685351	9.
62	7.301134	6.949674	6.691336	6.650636	6.239523	5.343541	6.037522	5.476068	4.425518	6.166736	...	9.591995	8.609359	8.
63	7.428060	6.767594	7.134458	7.631923	6.552234	5.315755	7.289287	5.417038	4.344127	5.973485	...	9.861801	8.539495	9.
64	7.280113	7.488313	6.905290	6.851237	5.671638	5.377161	6.025582	5.538950	4.441620	5.937243	...	8.439654	8.544075	9.

65 rows × 20486 columns

Fig:7 Imputed and Standardized Data

E. *Feature Extraction/ Selection*

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve. Irrelevant or partially relevant features can negatively impact model performance. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

- Reduces Over-fitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster.

*1) **Recursive Feature Elimination**:* The Recursive Feature Elimination (RFE) method works by recursively removing attributes and building a model on those attributes that remain. It uses accuracy metric to rank the feature according to their importance. The RFE method takes the model to be used and the number of required features as input. It then gives the ranking of all the variables, 1 being most important. It also gives its support, True being relevant feature and False being irrelevant feature.We have used following two techniques.

- **RFECV-Random Forest** [1] Random forest (RF) is a machine-learning method that generally works well with high-dimensional problems and allows for nonlinear relationships between predictors; however, the presence of correlated predictors has been shown to impact its ability to identify strong predictors. The Random Forest-Recursive Feature Elimination algorithm (RFECV-RF) mitigates this problem in smaller data sets, but this approach has not been tested in high-dimensional omnibus data sets.
- **RFECV-SVM** [2] Support vector machines (SVM) are a powerful tool to analyze data with a number of predictors approximately equal or larger than the number of observations. However, originally, application of SVM to analyze biomedical data was limited because SVM was not designed to evaluate importance of predictor variables. Creating predictor models based on only the most relevant variables is essential in biomedical research. Currently, substantial work has been done to allow assessment of variable importance in SVM models but this work has focused on SVM implemented with linear kernels. [4],Therefore, RFECV-SVM is a greedy feature selection method that generates a ranking list of features and selects a subset of the top-ranked features. The ranking is built by a feature weight vector w obtained from the parameters of the hyperplane decision function of a SVM classifier and the top p features are selected.

F. *Removing Highly Correlated Independent Variables*

Before proceeding further,we remove highly correlated independent variables, which reduces the size of **X-Matrix 65 x 20486** to **65 x 12252**. Highly correlated variables criteria is **abs[Correlation greater than 0.8]** .

G. *RFECV-Random Forest*

Recursive feature elimination using Random Forest with startified kfold cross validation. The model successfully reduced features from **12252 to 18**. Which results on the optimal number of features 18.See fig:8 and fig:9

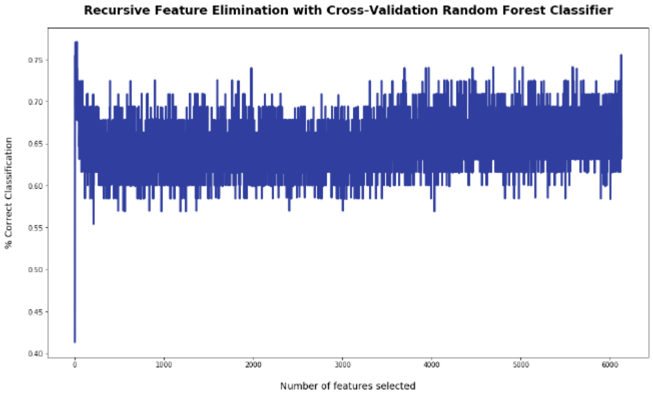


Fig:8 RFECV-Random Forest

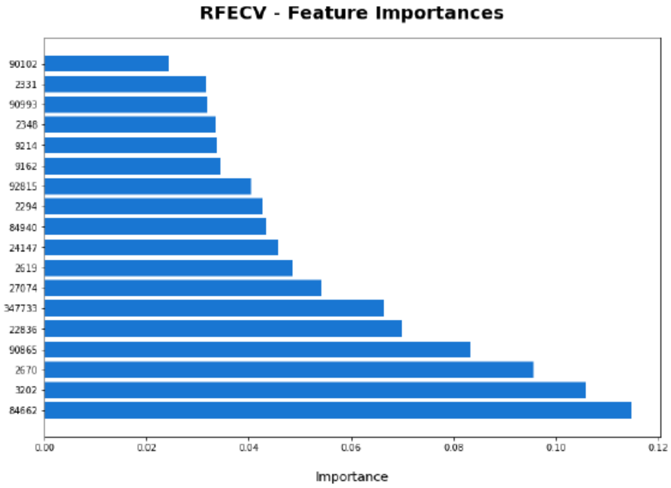


Fig:9 RFECV-Feature Importance

after feature selection modified **X** matrix **65 x 18** .

H. Sampling

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.We are using Stratified sampling technique. In statistics, stratified sampling is a method of sampling from a population which can be partitioned into sub populations. Similarly In our problem, we have the population and has sub population of genome. This strategy is best way to improve the coverage of genetics space.Three different ratios(Train/Test) that is 60/40, 70/30 and 80/20 are used.

I. Classification Types

We have trained five different models on the bases of five classification techniques. which are mentioned in below, whereas we achieved the results which has been shown in table:1

- Decision Tree classifier
- Random Forest Classifier
- Support Vector Machine Classifier
- K Nearest Neighbour Classifier
- eXtreme Gradient Boosting Classifier

Table I
RFECV-RANDOM FOREST

Model Accuray on Test Set			
Model	Split(60/40)	Split(70/30)	Split(80/20)
DTC	69%	70%	30%
RFC	84%	80%	84%
SVM	84%	80%	92%
KNN	92%	80%	92%
XGB	88%	80%	84%

Analysis of the above table shown that **KNN** has performed well in all three splits achieving the maximum accuracy of the model is **92 percent**. Classification report is shown in table2,Confusion Matrix and Classification Error is shown in fig:10 and fig:11 respectively.

Table II
KNN CLASSIFICATION REPORT

KNN Classification Report 80/20				
Metastasis	precision	recall	f1-score	support
Lung 0	1	1	1	4
Brain 1	1	1	1	4
Bone 2	0.8	1	0.89	4
Liver 3	0	0	0	1
accuracy			0.92	13
macro avg	0.7	0.75	0.72	13
weighted avg	0.86	0.92	0.89	13

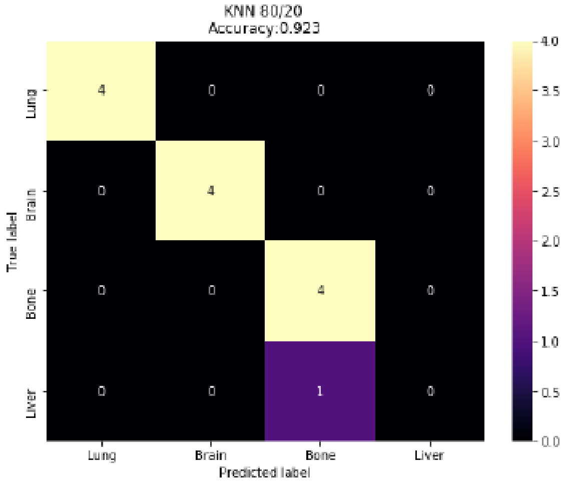


Fig:10 KNN Confusion Matrix

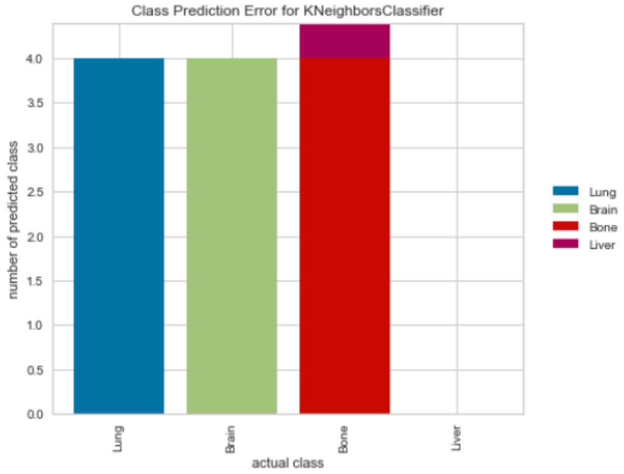


Fig:11 KNN-Classification Error

J. RFECV-Support Vector Machine

Recursive feature elimination using **SVM** with startified kfold cross validation. The model successfully reduced features from **12252 to 18**. Which results on the optimal number of features 46.See fig:12 and fig:13

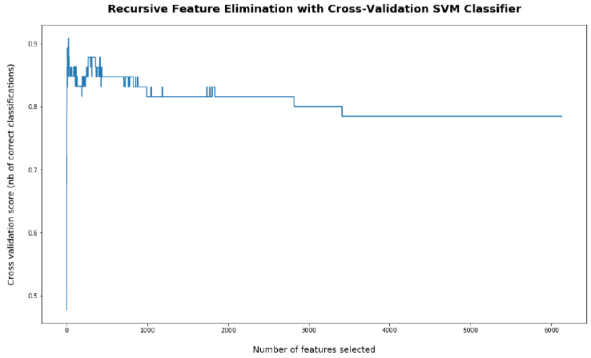


Fig:12 RFECV-SVM

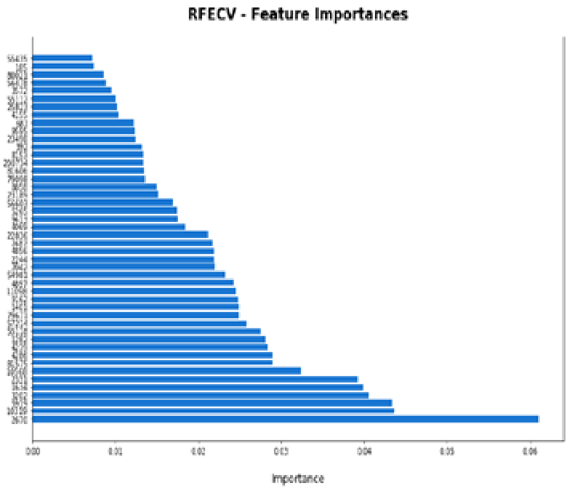


Fig:13 RFECV-Feature Importance

after feature selection modified **X** matrix **65 x 46** .We trained five different models and achieved following results shown in table3. Analysis of the above table shows that **SVM** has performed well in all three splits achieving the maximum accuracy of the model is **100 percent**. Classification report is shown in table:4,Confusion Matrix and classification Error is shown in fig:14 and fig:15 respectively

Table III
RFECV- SVM

Model Accuray on Test Set			
Model	Split(60/40)	Split(70/30)	Split(80/20)
DTC	69%	75%	69%
RFC	88%	80%	92%
SVM	100%	95%	92%
KNN	96%	95%	92%
XGB	84%	75%	92%

Table IV
SVM Classification Report 60/40

SVM Classification Report 60/40				
Metastasis	precision	recall	f1-score	support
Lung 0	1	1	1	8
Brain 1	1	1	1	9
Bone 2	1	1	1	7
Liver 3	1	1	1	2
accuracy			1	26
macro avg	1	1	1	26
weighted avg	1	1	1	26

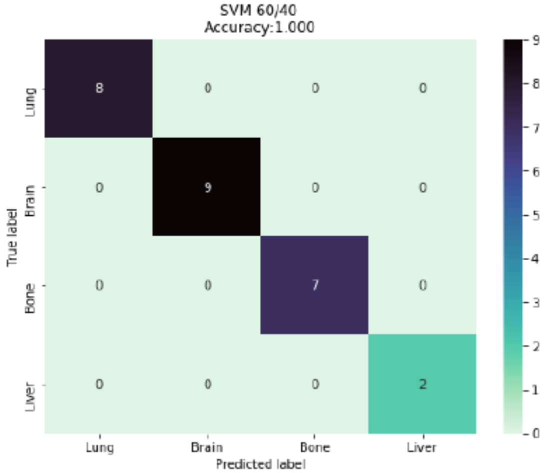


Fig:14 SVM Confusion Matrix

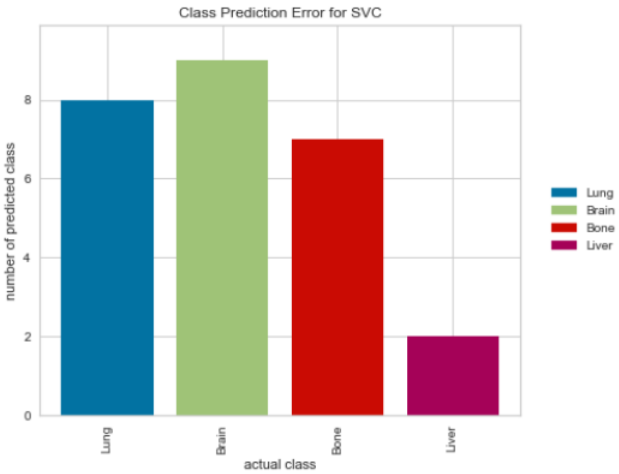


Fig:15 SVM-Classification Error

K. Results and Conclusion

Complex diseases such as breast cancer remain the greatest threat to human life. The growth of microarray data and the development of statistical methods have provided new possibilities for the prediction and treatment of such diseases. Feature selection and classification are the core technologies of microarray data analysis. They both play key roles in genes recognition and diseases diagnosis. Limited to the characteristics of microarray data, many typical methods in this field still need to be paid more attentions to overcome their disadvantages.

Feature selection and cross validation is a typical method, To reduce the time consumption of time, we firstly tries to reduce the recursion times by a large step size, and keep the step size decreasing while the number of features to be eliminated is getting smaller and by this way to ensure the quality of the meaningful genes selected.The advantages of SVM and reduces unnecessary computational cost for large-scale linear separable data such as microarray data becomes an efficient and effective feature selector compared with the existing methods and has potential in the gene selection field.

Random forest and SVM performed well on this Dataset, Further research needs to be done to improve the accuracy of classification prognosis breast cancer metastasis. due to limited computational capacity we were not able to explore the Dataset with further details.In-future, We would like to improve our models by fine tuning model parameters and we expect achieve more accurate, precise and less error prone classification.

References

[1] Burcu F Darst, Kristen C Malecki, and Corinne D Engelman. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics*, 19(1):65, 2018.

[2] Zifa Li, Weibo Xie, and Tao Liu. Efficient feature selection and classification for microarray data. *PloS one*, 13(8):e0202167, 2018.

[3] Alan Wee-Chung Liew, Ngai-Fong Law, and Hong Yan. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in bioinformatics*, 12(5):498–513, 2011.

[4] Hector Sanz, Clarissa Valim, Esteban Vegas, Josep M Oller, and Ferran Reverter. Svm-rfe: selection and visualization of the most relevant features through non-linear kernels. *BMC bioinformatics*, 19(1):1–18, 2018.