Cancer Diagnosis and Prognosis using Gene Expression

Group 3

Supervised by

Prof. Olubisi Runsewe

Problem Formulation

Identify liver cancer biomarkers using gene expression data to facilitate early diagnosis and monitoring disease prognosis.

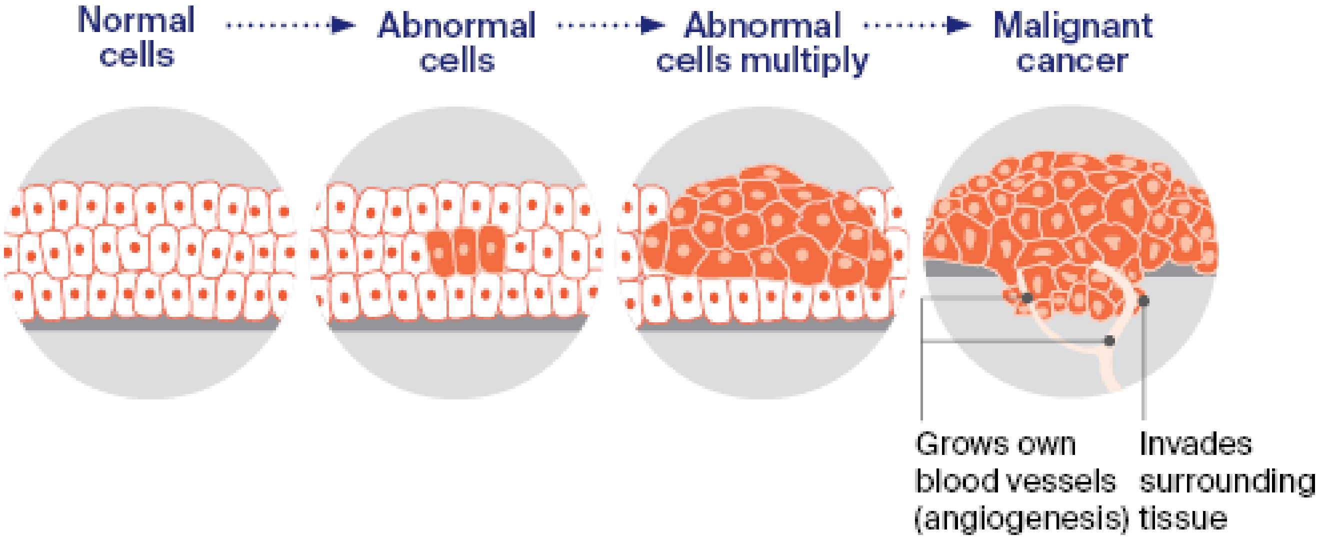
Agenda

- Introduction
- Project steps
- Results
- Conclusion
- Observations & Future work

1.

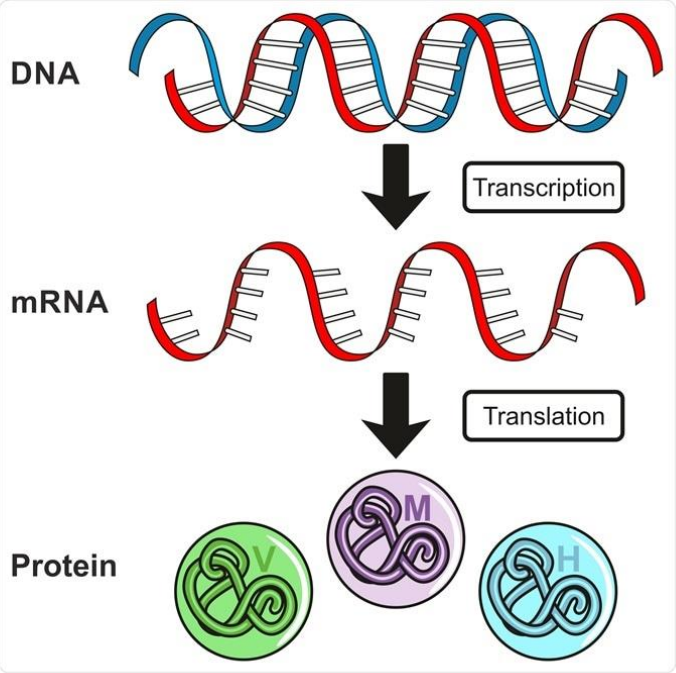
Introduction

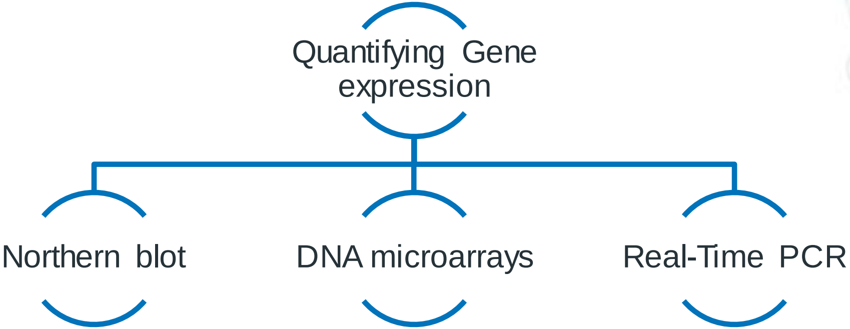
What is cancer ?



https://www.cancercouncil.com.au/cancer-information/understanding-cancer/what-is-cancer/

What is Gene Expression?





Quantifying Gene expression

Northern blot

DNA microarrays

Real-Time PCR

https://www.azolifesciences.com/article/A-Guide-to-Understanding-Gene-Expression.aspx

What are DNA Microarrays?

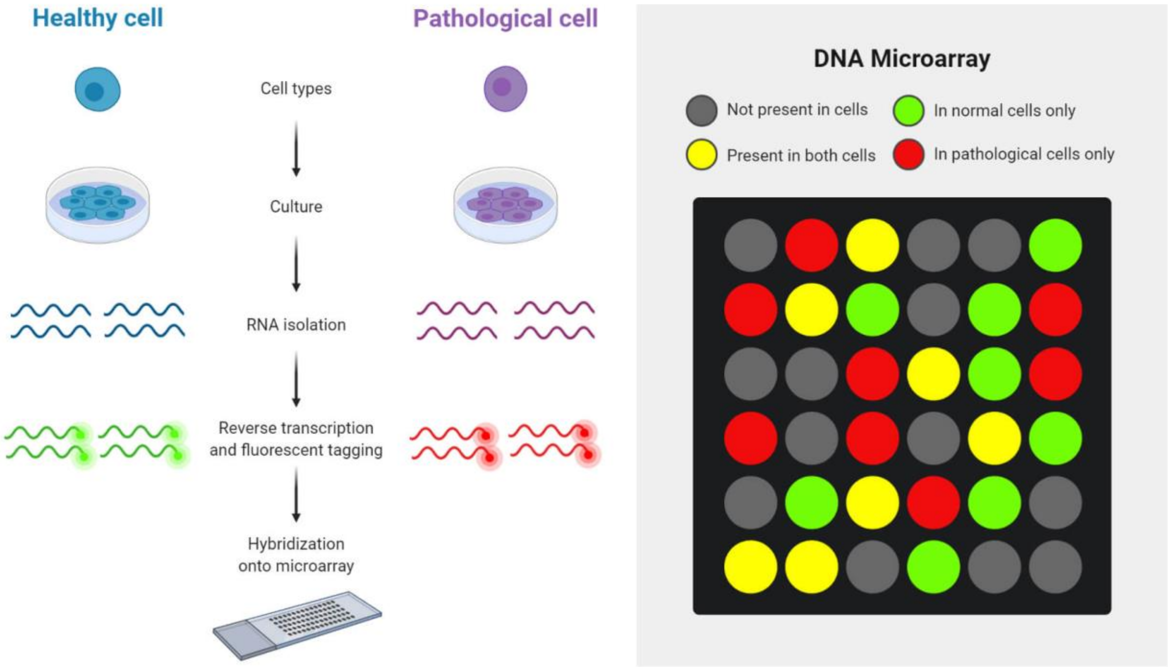


Image By Sagar Aryal, created using biorender.com

https://microbenotes.com/dna-microarray/

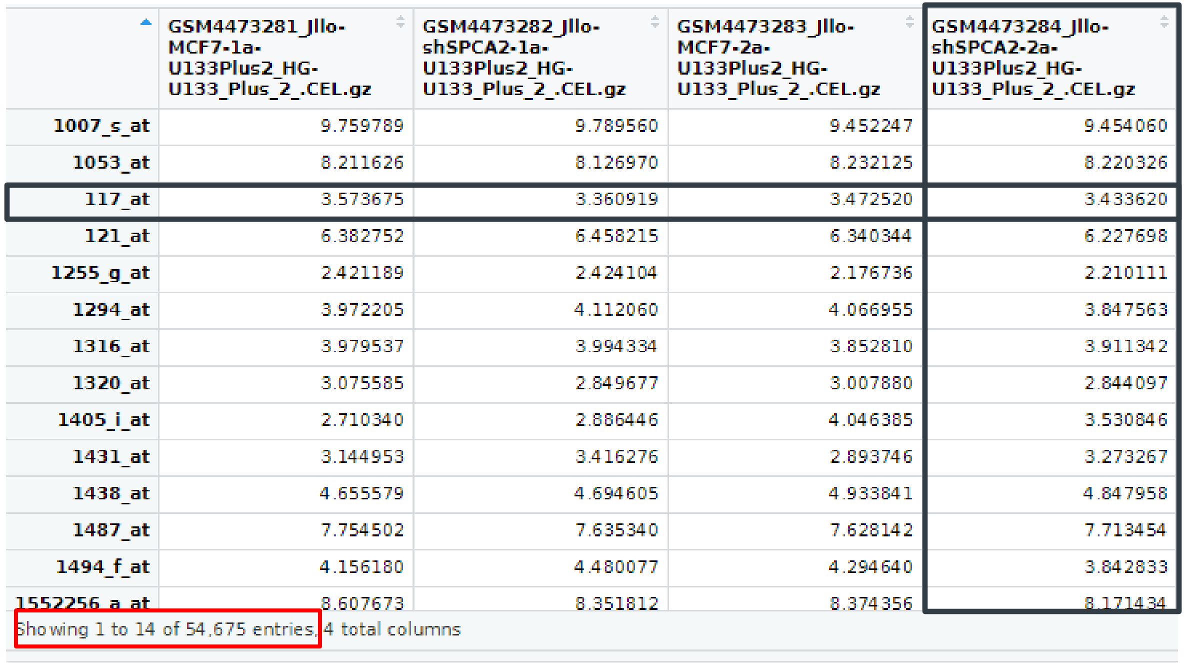
DNA Microarrays, Cancer tissues & Biomarkers

- Patterns of altered microarray expression profiles in cancer can serve as biomarkers for tumor diagnosis, prognosis of disease-specific outcomes, and prediction of treatments responses. [2]
- Microarray datasets containing expression profiles of several miRNAs or genes are used to identify biomarkers [2]
- Microarrays chips can have various sizes, they can have up to 2000 probes. [3]

What is Gene Expression Omnibus ?

- GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community.
- Microarray datasets have samples, their microarrays & meta data to describe other characteristics about tissue owner.

What is Gene Expression Omnibus ?



Tissue sample

Probe ID

	GSM4473281_Jllo-MCF7-1a-U133Plus2_HG-U133_Plus_2_.CEL.gz	GSM4473282_Jllo-shSPCA2-1a-U133Plus2_HG-U133_Plus_2_.CEL.gz	GSM4473283_Jllo-MCF7-2a-U133Plus2_HG-U133_Plus_2_.CEL.gz	GSM4473284_Jllo-shSPCA2-2a-U133Plus2_HG-U133_Plus_2_.CEL.gz
1007_s_at	9.759789	9.789560	9.452247	9.454060
1053_at	8.211626	8.126970	8.232125	8.220326
117_at	3.573675	3.360919	3.472520	3.433620
121_at	6.382752	6.458215	6.340344	6.227698
1255_g_at	2.421189	2.424104	2.176736	2.210111
1294_at	3.972205	4.112060	4.066955	3.847563
1316_at	3.979537	3.994334	3.852810	3.911342
1320_at	3.075585	2.849677	3.007880	2.844097
1405_i_at	2.710340	2.886446	4.046385	3.530846
1431_at	3.144953	3.416276	2.893746	3.273267
1438_at	4.655579	4.694605	4.933841	4.847958
1487_at	7.754502	7.635340	7.628142	7.713454
1494_f_at	4.156180	4.480077	4.294640	3.842833
1552256_a_at	8.607673	8.351812	8.374356	8.171434

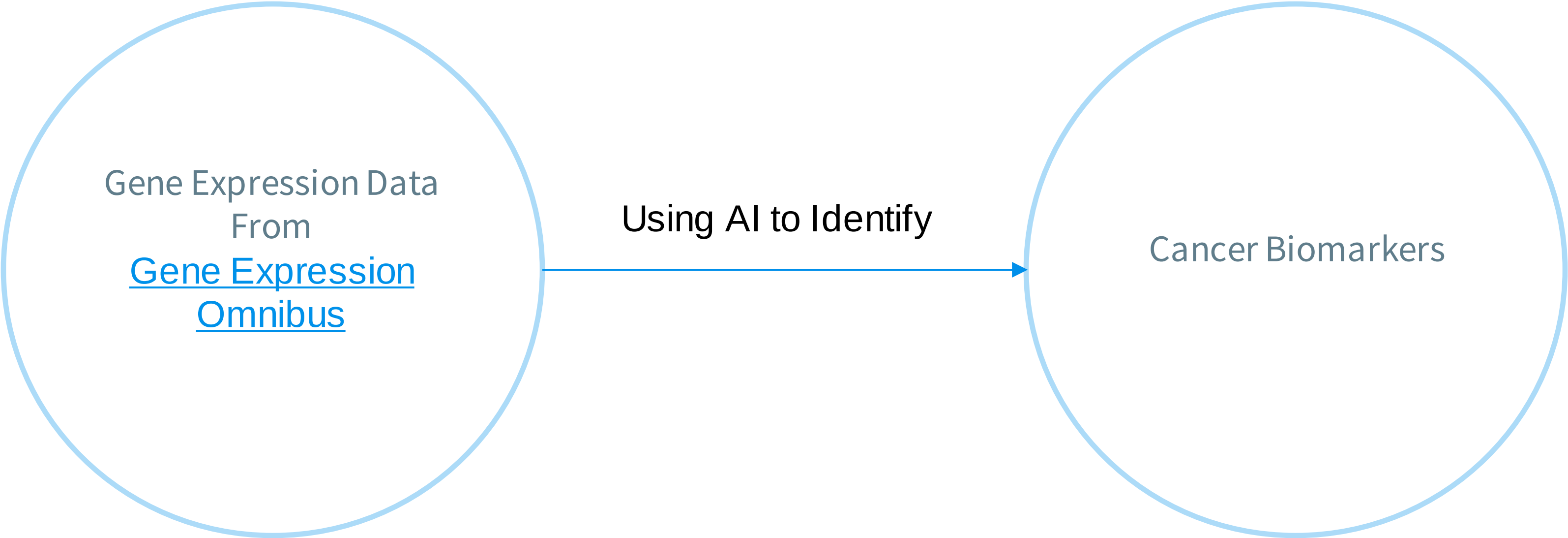
Showing 1 to 14 of 54,675 entries, 4 total columns

An example of data with ID = “GSE148537” downloaded from Gene Expression Omnibus

Cancer Diagnosis & prognosis

- **Diagnosis:**
  - The process of identifying a disease, condition, or injury from its signs and symptoms.
- **Prognosis:**
  - predicting the likely or expected development of a disease.
- **Methods Used in both:**
  - Lab Test
    - Blood, Tissue samples …
  - Imaging Tests
    - MRI, CT …

Another Approach for cancer diagnosis & prognosis



Gene Expression Data From Gene Expression Omnibus

Using AI to Identify

Cancer Biomarkers

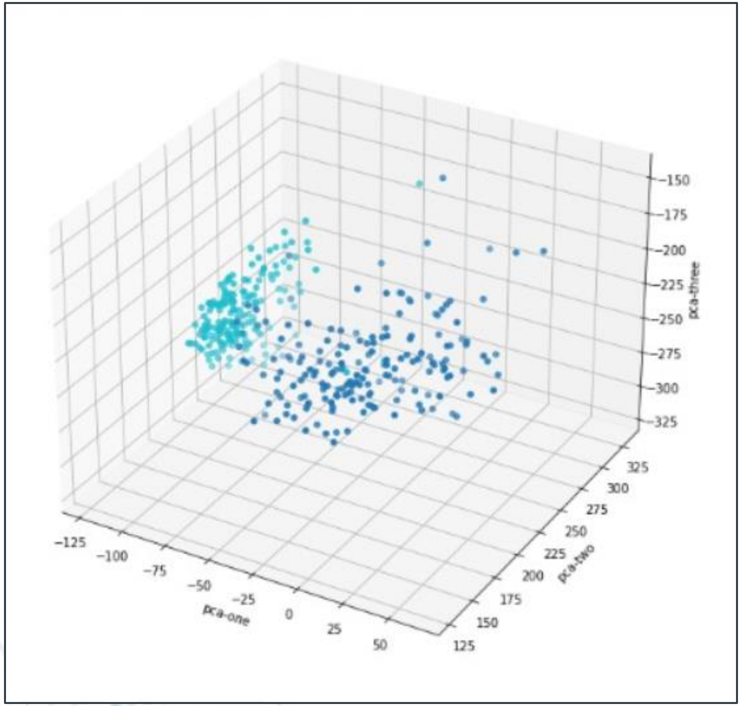
2.

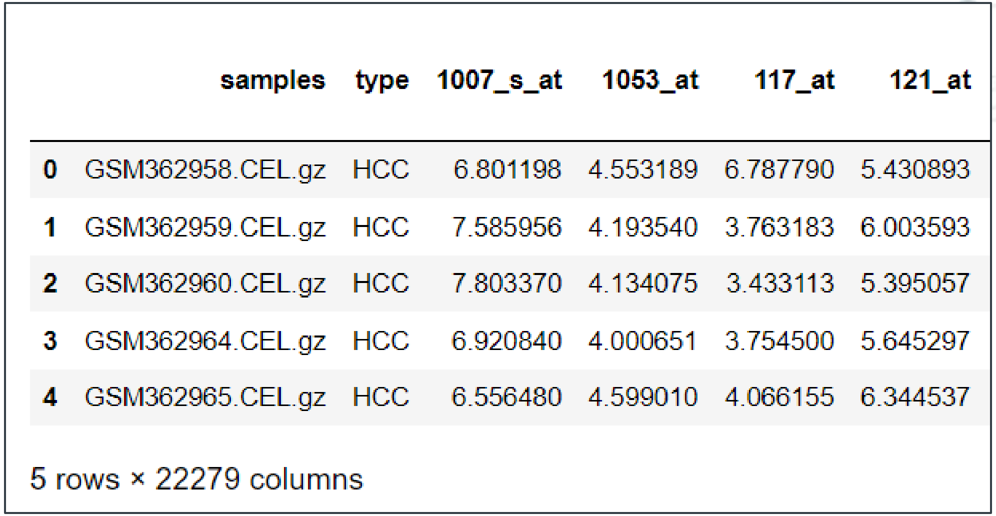
Project Steps

1- Dataset

- **Dataset:**
  - Challenges
    - Different environments
    - Different protocols & different metadata structure
    - HUGE number of features (Genes) per sample
    - Small number of samples
- **Selected dataset:**
  - Data with ID “GSE14520” from “**CuMiDa**” project that includes normalized microarrays of 375 samples of healthy and cancerous liver tissues
  - Raw data can be found here

2- Data visualization & Exploration





	samples	type	1007_s_at	1053_at	117_at	121_at
0	GSM362958.CEL.gz	HCC	6.801198	4.553189	6.787790	5.430893
1	GSM362959.CEL.gz	HCC	7.585956	4.193540	3.763183	6.003593
2	GSM362960.CEL.gz	HCC	7.803370	4.134075	3.433113	5.395057
3	GSM362964.CEL.gz	HCC	6.920840	4.000651	3.754500	5.645297
4	GSM362965.CEL.gz	HCC	6.556480	4.599010	4.066155	6.344537

5 rows × 22279 columns

3- System Architecture

- **Using different techniques to identify biomarkers**
  - PCA
  - Feature Selection using Chi-square
- **Using Different modeling techniques**
  - SVM
  - Naïve Bias
- **Evaluating different models**
  - Using K-Fold cross validation

4- System Evaluation

- **Evaluating best model & dimensionality reduction strategy**
  - Confusion matrix.
  - F1-score & Accuracy scores.
- **Evaluating acquired Biomarkers**
  - Checking the selected biomarkers against famous biomarkers for this cancer type.

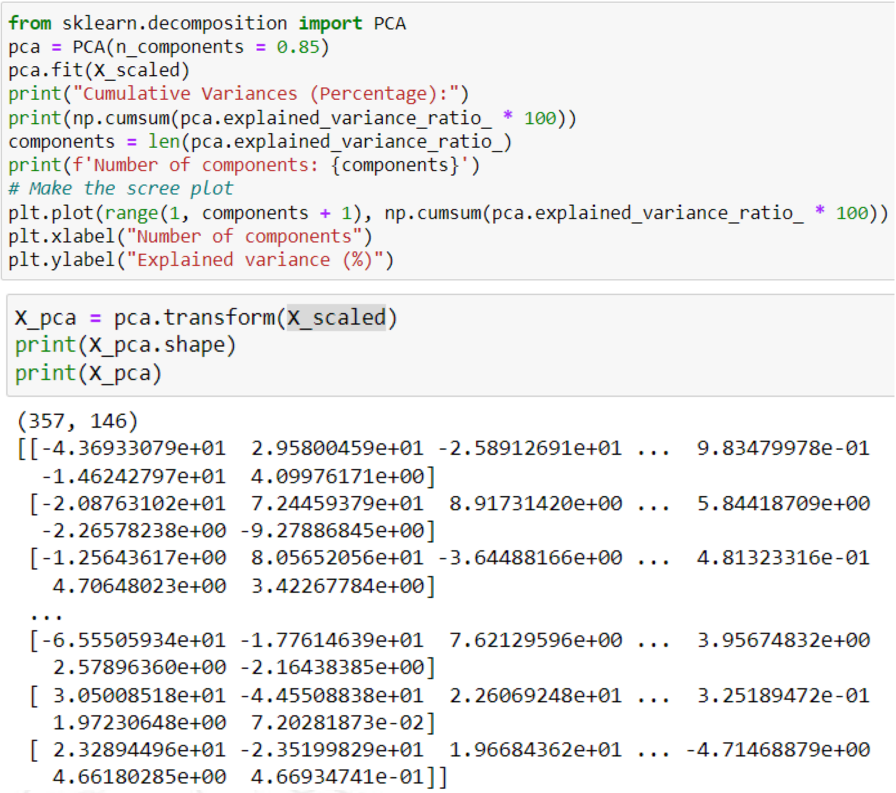
Using Proposed System For Prognosis & Diagnosis

- **Diagnosis**
  - The identified biomarkers will be checked for every new patient liver tissue sample, if they exist in patient tissue samples then this patient may get liver cancer.
- **Prognosis**
  - The identified biomarkers will be monitored while the patient is undergoing treatment, if their values are getting into normal range then the treatment is working.

3.

Results

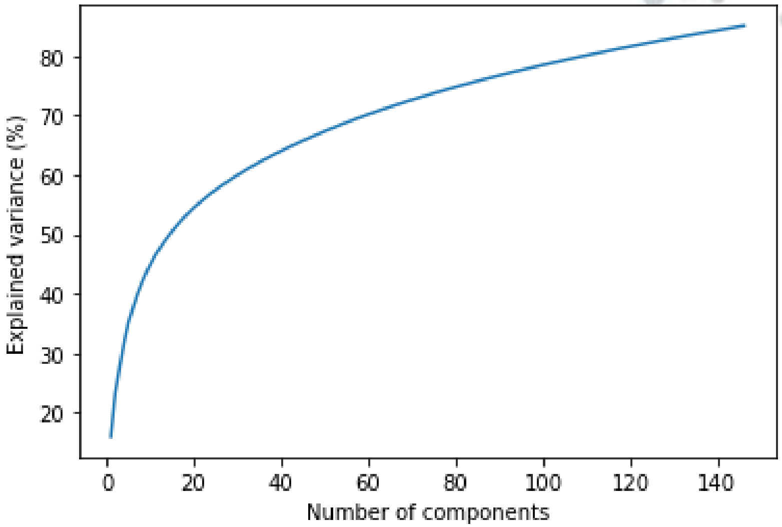
Apply PCA Analysis



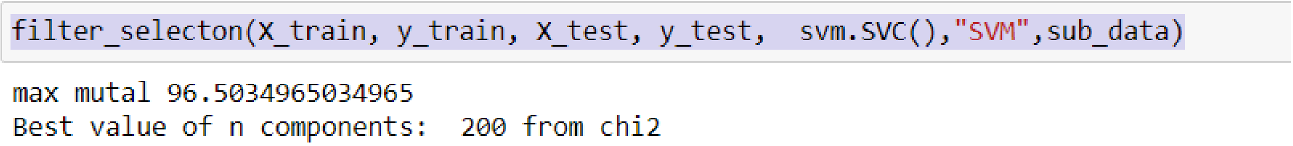
```
from sklearn.decomposition import PCA
pca = PCA(n_components = 0.85)
pca.fit(X_scaled)
print("Cumulative Variances (Percentage):")
print(np.cumsum(pca.explained_variance_ratio_ * 100))
components = len(pca.explained_variance_ratio_)
print(f'Number of components: {components}')
# Make the scree plot
plt.plot(range(1, components + 1), np.cumsum(pca.explained_variance_ratio_ * 100))
plt.xlabel("Number of components")
plt.ylabel("Explained variance (%)")
```

```
X_pca = pca.transform(X_scaled)
print(X_pca.shape)
print(X_pca)
```

```
(357, 146)
[[-4.36933079e+01  2.95800459e+01 -2.58912691e+01 ...  9.83479978e-01
  -1.46242797e+01  4.09976171e+00]
 [-2.08763102e+01  7.24459379e+01  8.91731420e+00 ...  5.84418709e+00
  -2.26578238e+00 -9.27886845e+00]
 [-1.25643617e+00  8.05652056e+01 -3.64488166e+00 ...  4.81323316e-01
   4.70648023e+00  3.42267784e+00]
 ...
 [-6.55505934e+01 -1.77614639e+01  7.62129596e+00 ...  3.95674832e+00
   2.57896360e+00 -2.16438385e+00]
 [ 3.05008518e+01 -4.45508838e+01  2.26069248e+01 ...  3.25189472e-01
   1.97230648e+00  7.20281873e-02]
 [ 2.32894496e+01 -2.35199829e+01  1.96684362e+01 ... -4.71468879e+00
   4.66180285e+00  4.66934741e-01]]
```

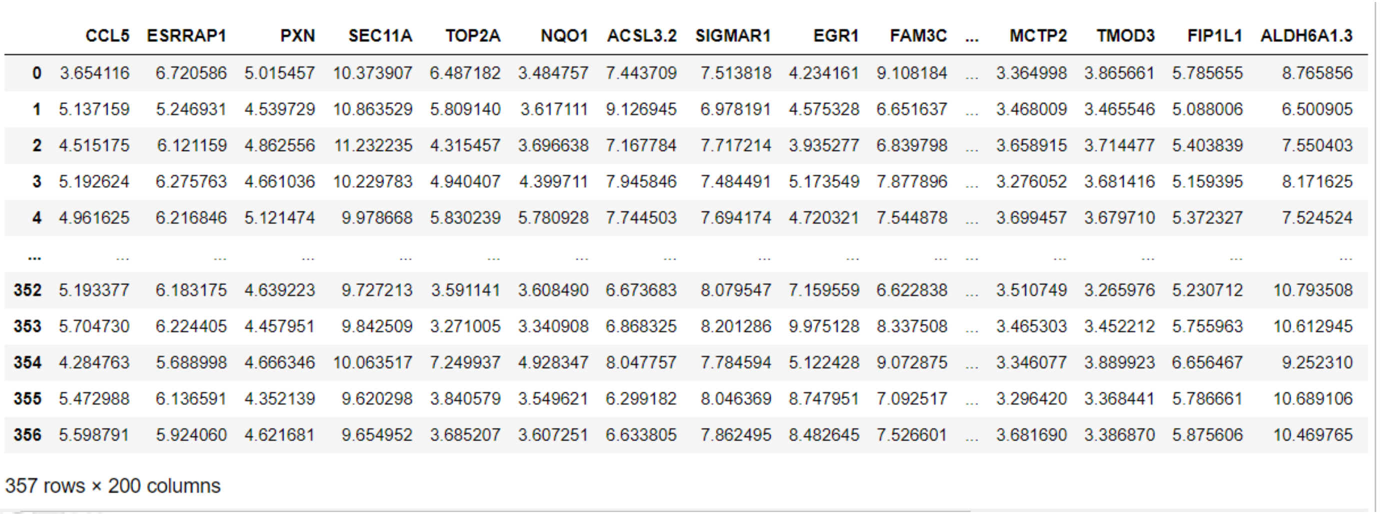


Apply Filter Selection Method using Chi square



```
filter_selecton(X_train, y_train, X_test, y_test,  svm.SVC(),"SVM",sub_data)
```

```
max mutal 96.5034965034965
Best value of n components:  200 from chi2
```

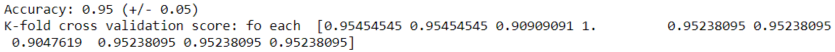


	CCL5	ESRRAP1	PXN	SEC11A	TOP2A	NQO1	ACSL3.2	SIGMAR1	EGR1	FAM3C	...	MCTP2	TMOD3	FIP1L1	ALDH6A1.3
0	3.654116	6.720586	5.015457	10.373907	6.487182	3.484757	7.443709	7.513818	4.234161	9.108184	...	3.364998	3.865661	5.785655	8.765856
1	5.137159	5.246931	4.539729	10.863529	5.809140	3.617111	9.126945	6.978191	4.575328	6.651637	...	3.468009	3.465546	5.088006	6.500905
2	4.515175	6.121159	4.862556	11.232235	4.315457	3.696638	7.167784	7.717214	3.935277	6.839798	...	3.658915	3.714477	5.403839	7.550403
3	5.192624	6.275763	4.661036	10.229783	4.940407	4.399711	7.945846	7.484491	5.173549	7.877896	...	3.276052	3.681416	5.159395	8.171625
4	4.961625	6.216846	5.121474	9.978668	5.830239	5.780928	7.744503	7.694174	4.720321	7.544878	...	3.699457	3.679710	5.372327	7.524524
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
352	5.193377	6.183175	4.639223	9.727213	3.591141	3.608490	6.673683	8.079547	7.159559	6.622838	...	3.510749	3.265976	5.230712	10.793508
353	5.704730	6.224405	4.457951	9.842509	3.271005	3.340908	6.868325	8.201286	9.975128	8.337508	...	3.465303	3.452212	5.755963	10.612945
354	4.284763	5.688998	4.666346	10.063517	7.249937	4.928347	8.047757	7.784594	5.122428	9.072875	...	3.346077	3.889923	6.656467	9.252310
355	5.472988	6.136591	4.352139	9.620298	3.840579	3.549621	6.299182	8.046369	8.747951	7.092517	...	3.296420	3.368441	5.786661	10.689106
356	5.598791	5.924060	4.621681	9.654952	3.685207	3.607251	6.633805	7.862495	8.482645	7.526601	...	3.681690	3.386870	5.875606	10.469765

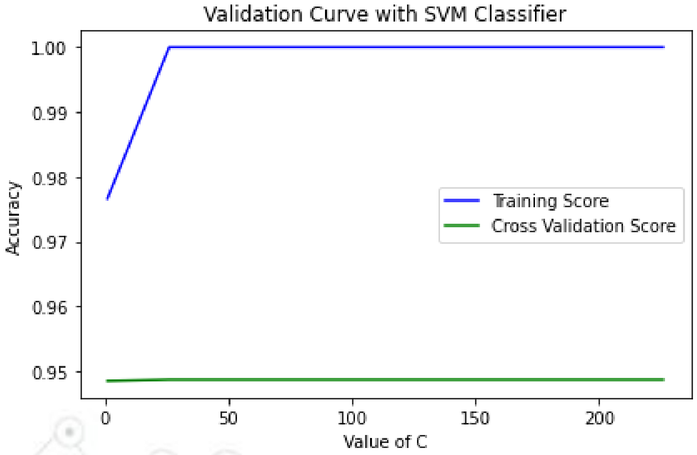
357 rows × 200 columns

Apply Cross Validation on different models using PCA data

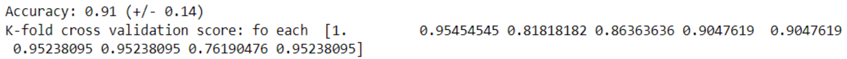
SVM Model



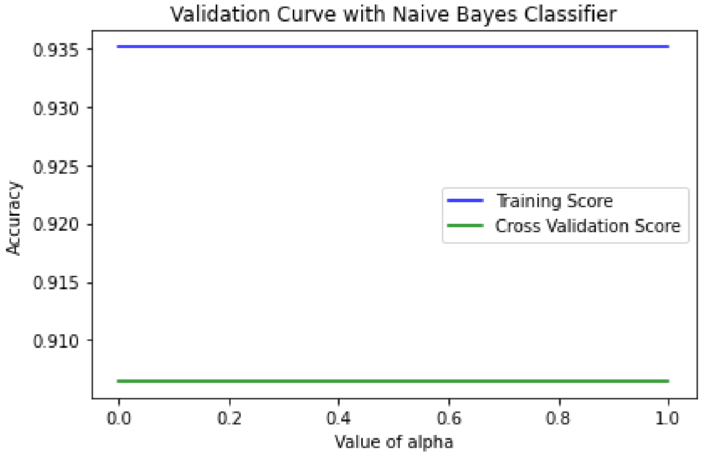
```
Accuracy: 0.95 (+/- 0.05)
K-fold cross validation score: fo each  [0.95454545 0.95454545 0.90909091 1.         0.95238095 0.95238095
 0.9047619  0.95238095 0.95238095 0.95238095]
```



Naïve Bayes

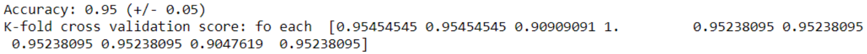


```
Accuracy: 0.91 (+/- 0.14)
K-fold cross validation score: fo each  [1.         0.95454545 0.81818182 0.86363636 0.9047619  0.9047619
 0.95238095 0.95238095 0.76190476 0.95238095]
```

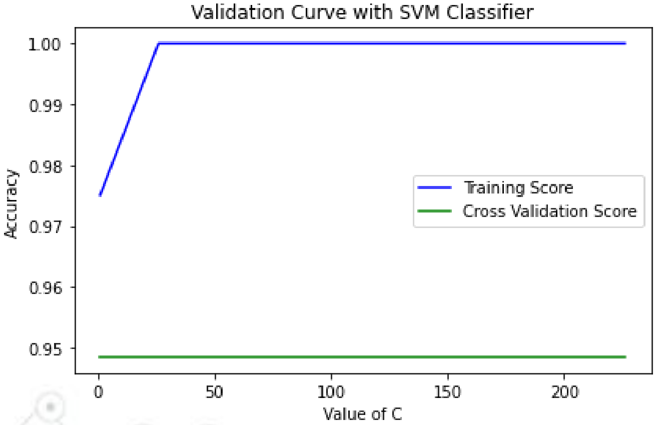


Apply Cross Validation on different models using Filter selection data

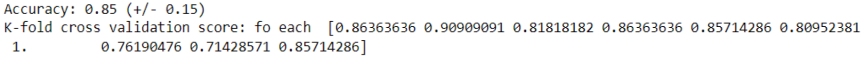
SVM Model



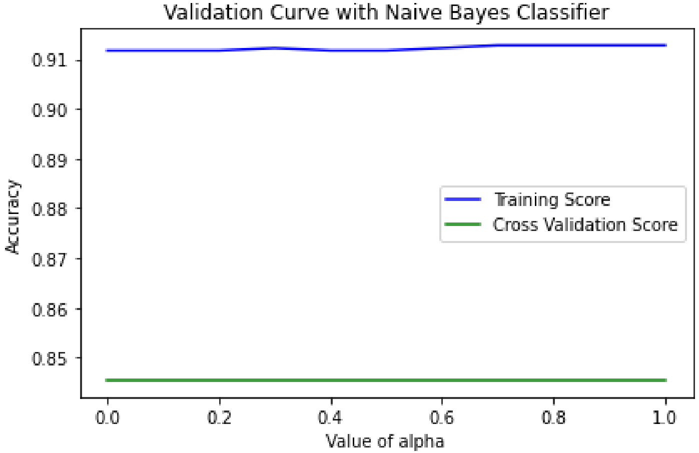
```
Accuracy: 0.95 (+/- 0.05)
K-fold cross validation score: fo each  [0.95454545 0.95454545 0.90909091 1.         0.95238095 0.95238095
 0.95238095 0.95238095 0.9047619  0.95238095]
```



Naïve Bayes

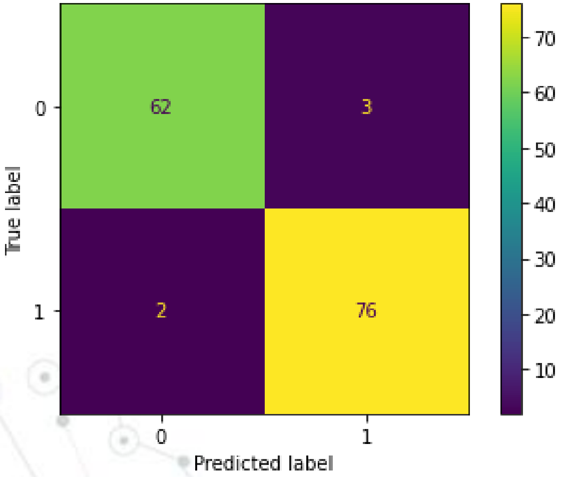


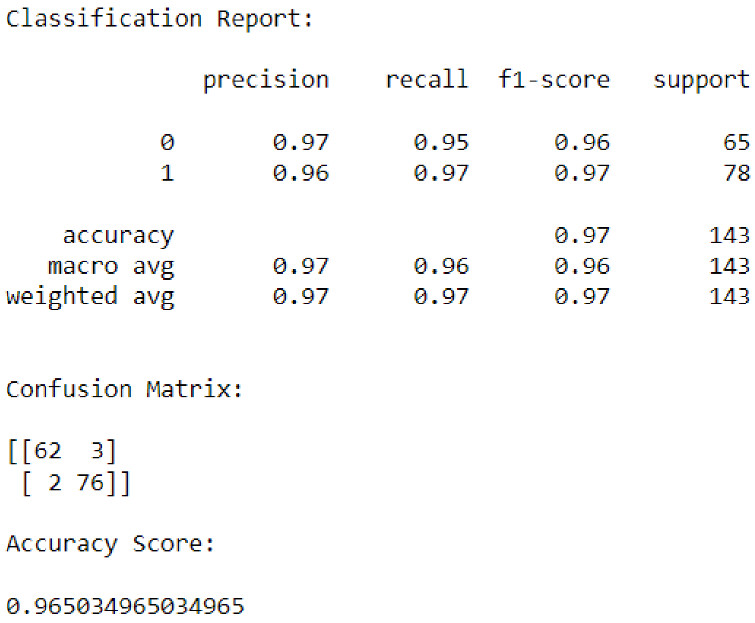
```
Accuracy: 0.85 (+/- 0.15)
K-fold cross validation score: fo each  [0.86363636 0.90909091 0.81818182 0.86363636 0.85714286 0.80952381
 1.         0.76190476 0.71428571 0.85714286]
```



Choose Champion Model

- After Evaluation cross validation we choose SVM model using PCA data as a champion model





```
Classification Report:

              precision    recall  f1-score   support

           0       0.97      0.95      0.96        65
           1       0.96      0.97      0.97        78

    accuracy                           0.97       143
   macro avg       0.97      0.96      0.96       143
weighted avg       0.97      0.97      0.97       143


Confusion Matrix:

[[62  3]
 [ 2 76]]

Accuracy Score:

0.965034965034965
```

4.

Conclusion

Conclusion

The following topics were covered

- Gene expression data and its challenges.
- Biomarkers and their role in cancer diagnosis and prognosis.
- Proposed a system for identifying cancer biomarkers.

5.

Future Trends

Future trends & observations

- Developing AI frameworks to identify biomarkers with the advances in computing efficiency.
- Gene expression data will increase & its accuracy will be better as technology is advancing.

Thank You

References

- https://www.cancer.gov/publications/dictionaries/cancer-terms/def/diagnosis
- https://en.wikipedia.org/wiki/Prognosis
- https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis
- https://www.cancercouncil.com.au/cancer-information/understanding-cancer/what-is-cancer/
- https://www.azolifesciences.com/article/A-Guide-to-Understanding-Gene-Expression.aspx
- https://microbenotes.com/dna-microarray/

References

[1] A. Mohammed, G. Biegert, J. Adamec, and T. Helikar, “Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers,” *Oncotarget*, vol. 8, no. 49, p. 85692, Oct. 2017, doi: 10.18632/ONCOTARGET.21127.

[2] D. Chakraborty and U. Maulik, “Oncology Identifying Cancer Biomarkers From Microarray Data Using Feature Selection and Semisupervised Learning”, doi: 10.1109/JTEHM.2014.2375820.

[3] Z. He, L. Wu, M. W. Fields, and J. Zhou, “Use of Microarrays with Different Probe Sizes for Monitoring Gene Expression,” *Applied and Environmental Microbiology*, vol. 71, no. 9, p. 5154, Sep. 2005, doi: 10.1128/AEM.71.9.5154-5162.2005.