CSE 601
Data Mining and Bioinformatics

Programming Assignment 1
PCA

Submitted By
Amit There - 50247978 - amitpadm
Sandesh Patnurkar - 50246311 - sandeshv

Principal Component Analysis:

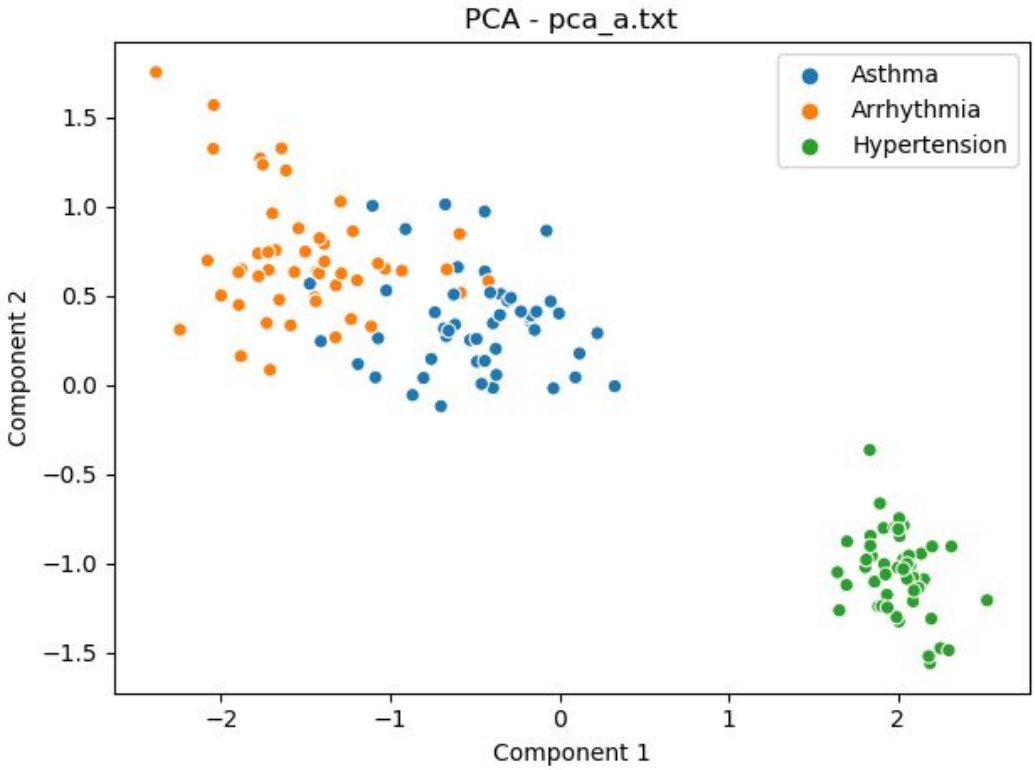
Principal Component Analysis is a method that transfers a set of correlated dimensions into a new set of uncorrelated dimensions. It reduces the number of dimensions of the data. This is especially helpful when the data has a lot of dimensions and you wish to visualize it in a single or a 2 dimensional space or maybe a 3 dimensional space.

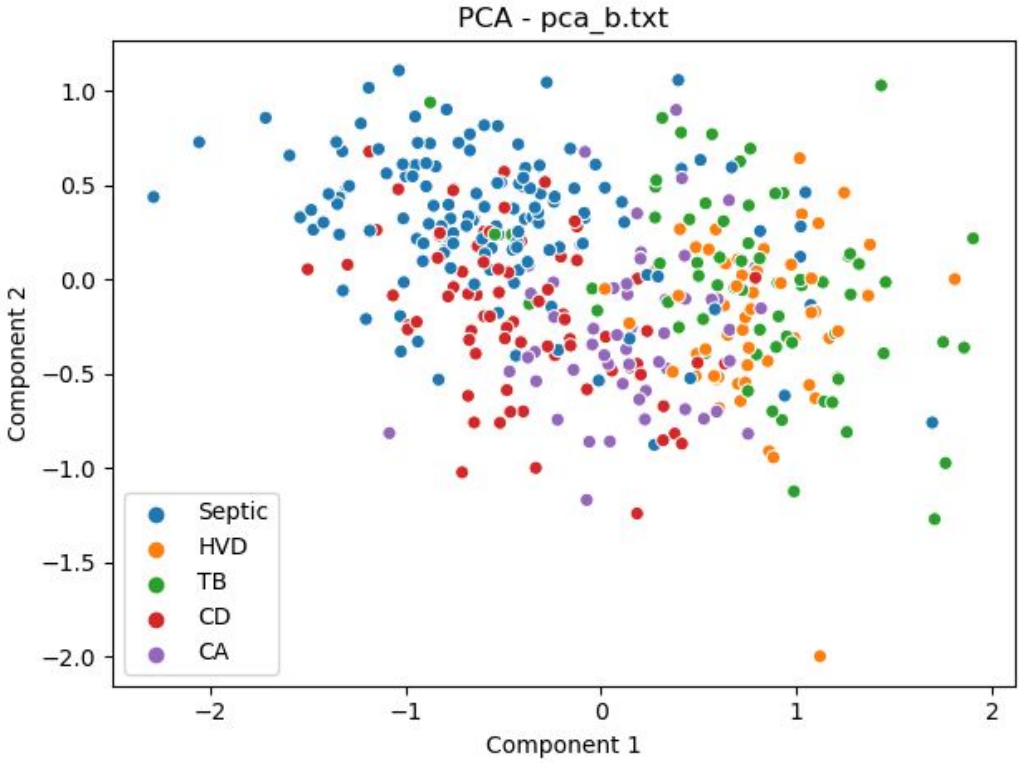
Steps:

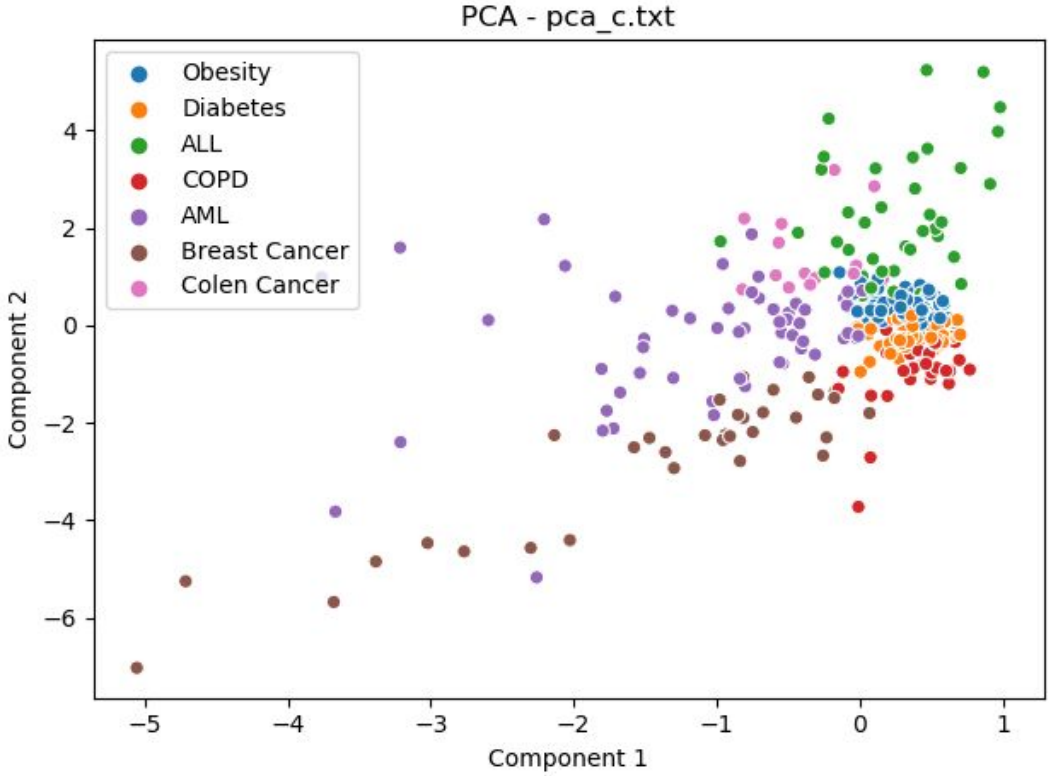
This is done by following the following sequence of steps in the code:

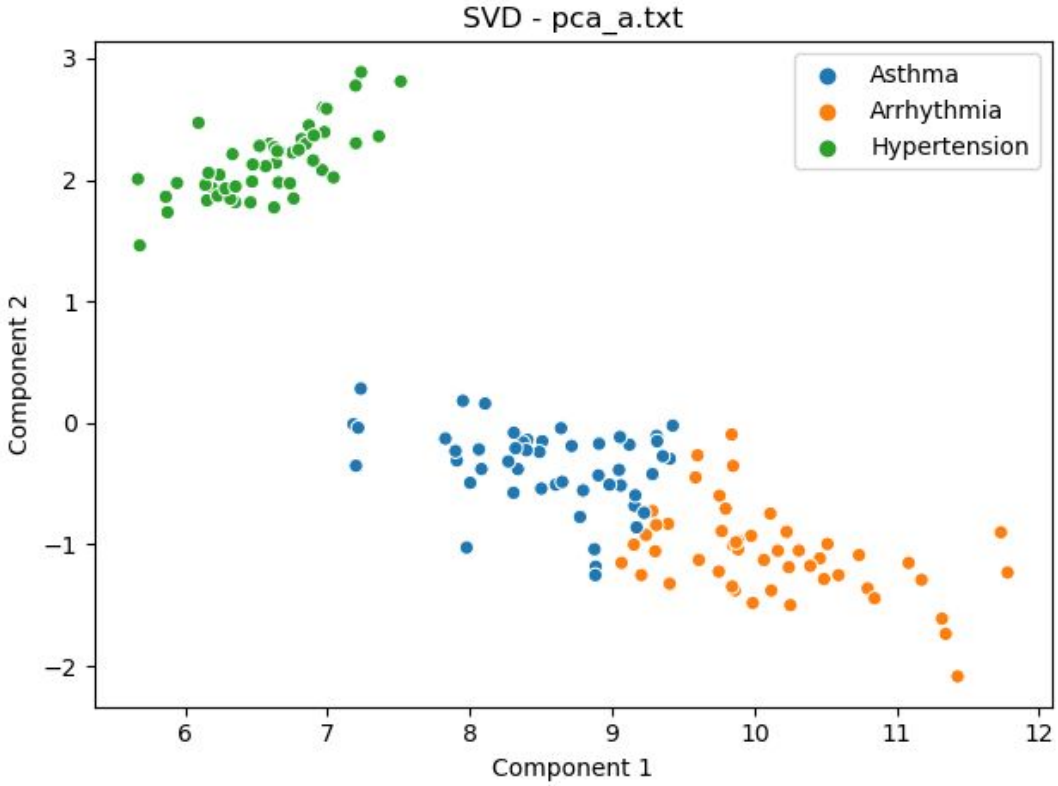
1) Read the file in a matrix.
2) We find a mean vector i.e. taking the mean of all the rows.
3) Normalize the data using this mean vector.
4) Compute the covariance matrix of this normalized vector.
5) Finding eigenvalues and eigenvectors of the covariance matrix.
6) Sorting on eigenvalues and taking the top d eigenvectors which become the directions with largest variances.
7) Rotate the data from original direction to the ones represented by these eigenvectors in 2 dimensional space.
8) Plot the rotated data.
9) We use the unique diseases from the last column in each file as colours for plotting the data points.

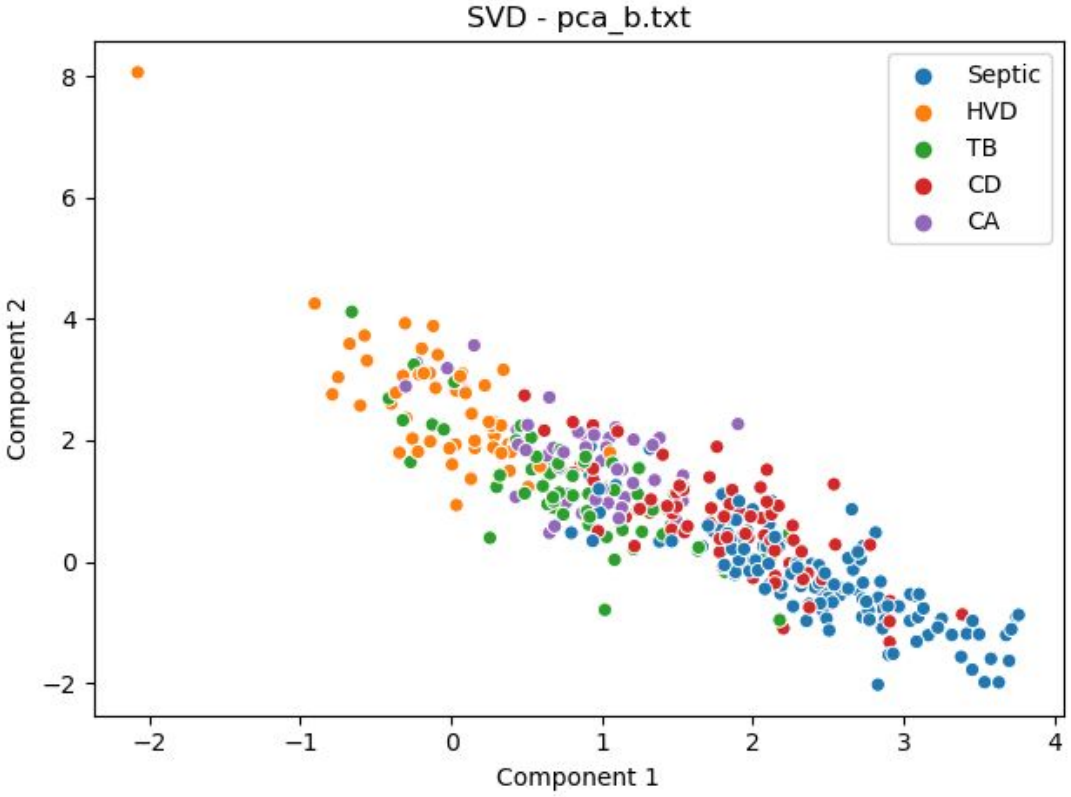
Plots:

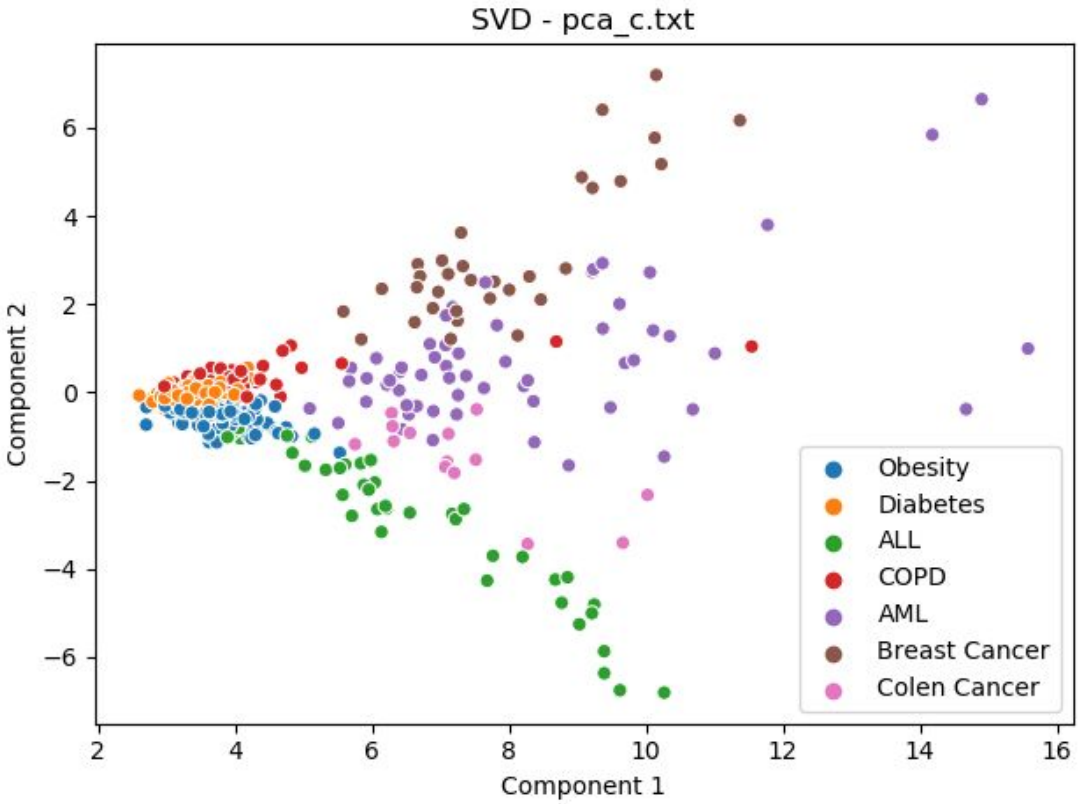

PCA - pca_a.txt
Asthma
Arrhythmia
Hypertension
Component 2
Component 1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1
2

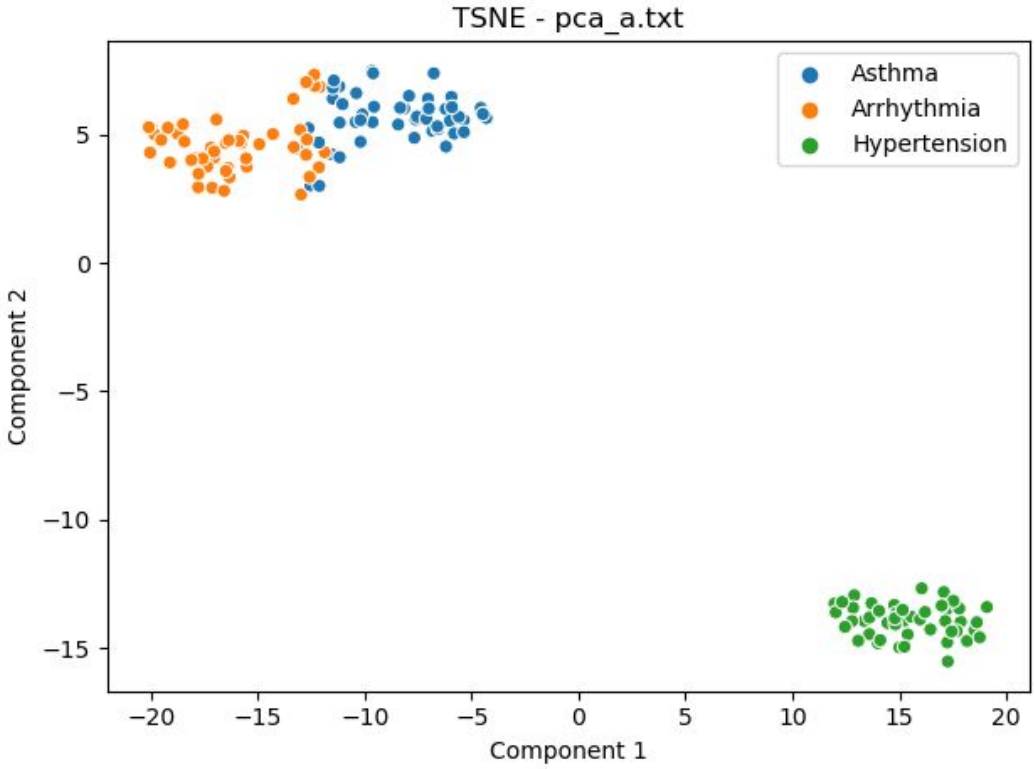

PCA - pca_b.txt
Septic
HVD
TB
CD
CA
Component 2
Component 1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2

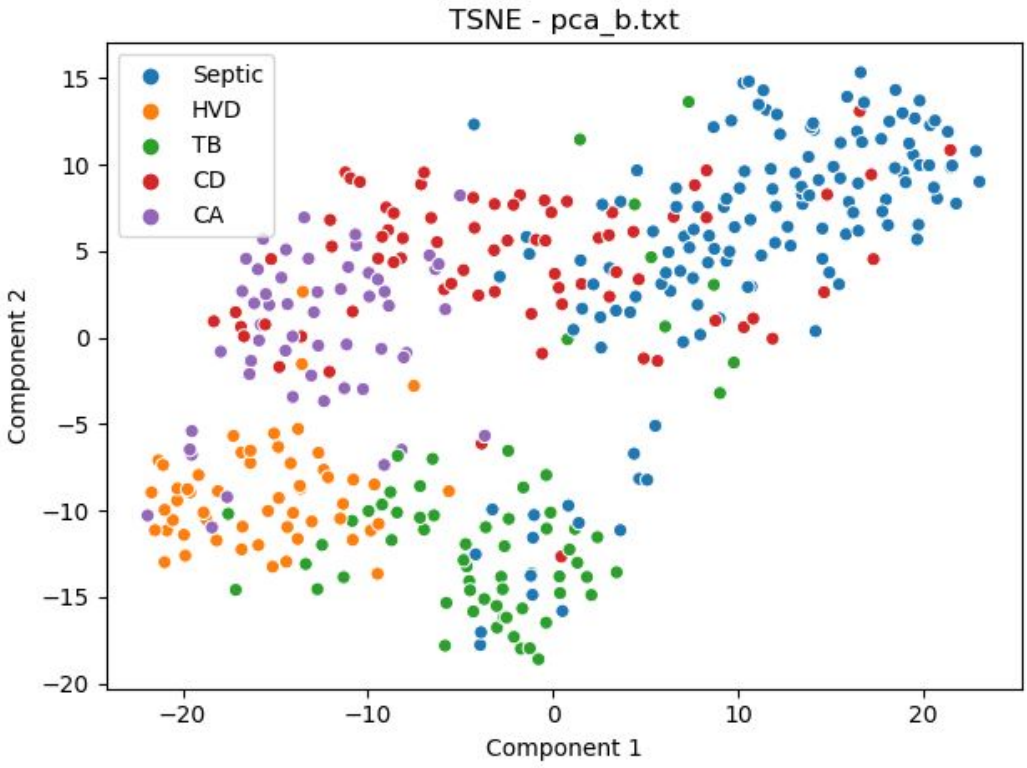

PCA - pca_c.txt
Obesity
Diabetes
ALL
COPD
AML
Breast Cancer
Colen Cancer
Component 2
Component 1
4
2
0
−2
−4
−6
−5
−4
−3
−2
−1
0
1

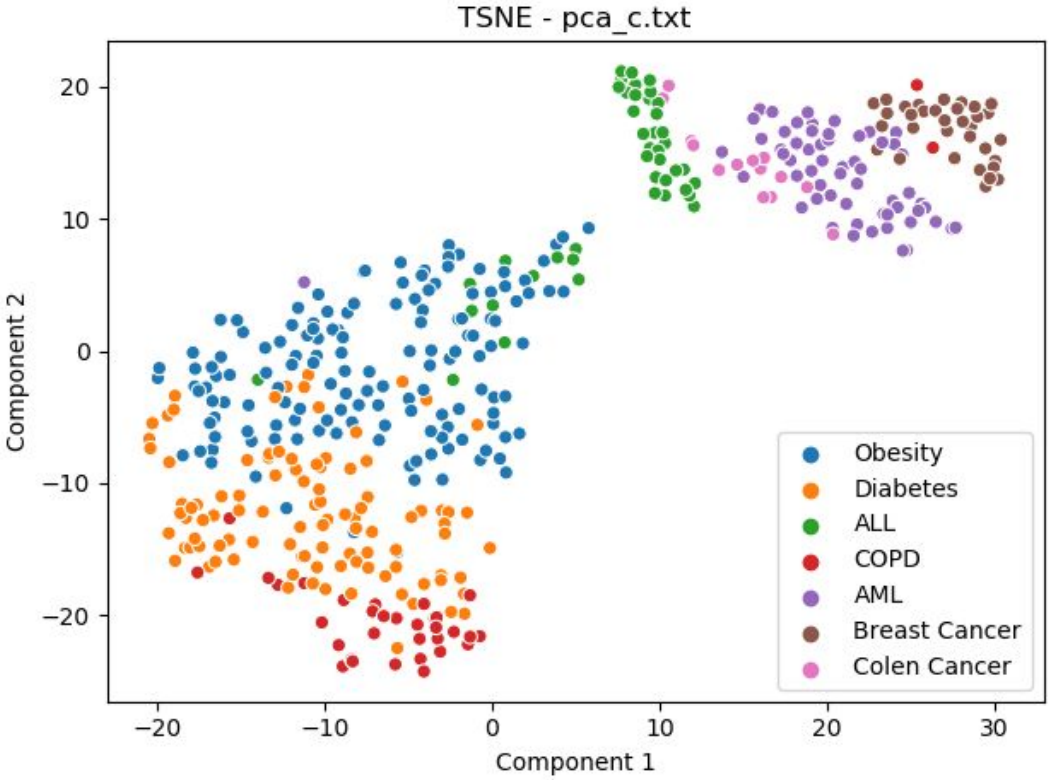

SVD - pca_a.txt
Asthma
Arrhythmia
Hypertension
Component 2
Component 1
3
2
1
0
−1
−2
6
7
8
9
10
11
12


SVD - pca_b.txt
Septic
HVD
TB
CD
CA
Component 2
Component 1
8
6
4
2
0
−2
−2
−1
0
1
2
3
4


SVD - pca_c.txt
Obesity
Diabetes
ALL
COPD
AML
Breast Cancer
Colen Cancer
Component 2
Component 1
6
4
2
0
−2
−4
−6
2
4
6
8
10
12
14
16


TSNE - pca_a.txt
Asthma
Arrhythmia
Hypertension
Component 2
Component 1
5
0
−5
−10
−15
−20
−15
−10
−5
0
5
10
15
20


TSNE - pca_b.txt
Septic
HVD
TB
CD
CA
Component 2
Component 1
15
10
5
0
−5
−10
−15
−20
−20
−10
0
10
20


TSNE - pca_c.txt
Component 2
Component 1
20
10
0
−10
−20
−20
−10
0
10
20
30
Obesity
Diabetes
ALL
COPD
AML
Breast Cancer
Colen Cancer