CSE 601:DATA MINING & BIOINFORMATICS

Dimensionality Reduction(PCA)

Pradeep Singh Bisht	50247429	pbisht2@buffalo.edu
Ashutosh Ahmad Alexandar	50248859	alexanda@bufalo.edu
Dilip Reddy Gaddam	50248867	dilipred@buffalo.edu

Principal Component Analysis:

Principal component analysis is a dimensionality reduction tool i.e. a dataset with large number of variables can be converted to dataset with small number of variables without losing most of information present in the dataset.

In our project we reduce the given n-dimensional data set to 2 dimensional data and plot the data.

Steps for PCA:

1.Given data is standardized around the mean by subtracting the mean from original data. Mean is calculated across every column.

$$X = X - \bar{X}$$

2.Calculate the covariance of data matrix. Covariance matrix is calculated using 'np.cov'(python)

$$S = \frac{1}{n} XX^T$$

3.Compute Eigen Values and Eigen vectors from covariance matrix using the function: 'np.linalg.eig'

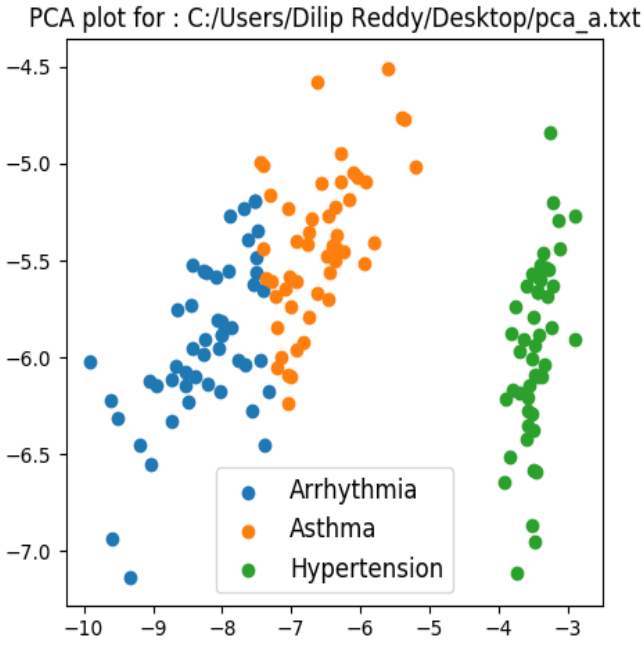
$$Sa = \lambda a$$

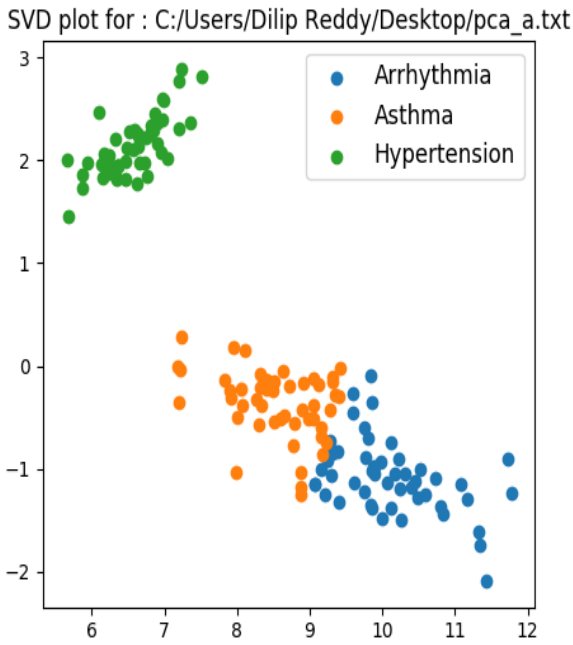
4.Sort the Eigen values in descending order and take top 2 Eigen values and corresponding Eigen vectors which for principal components.

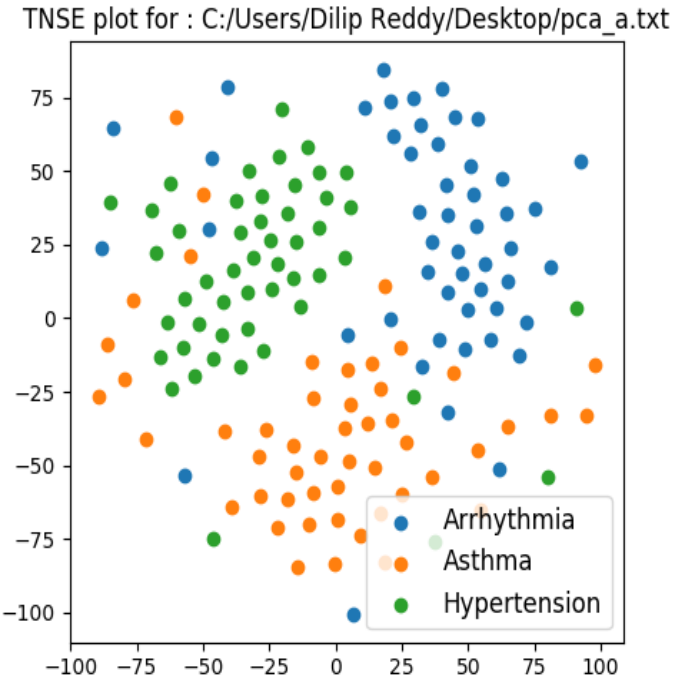
5. Generate new 2 dimensional data Using the principal components and original data.

Output of Dimensionality Reduction:

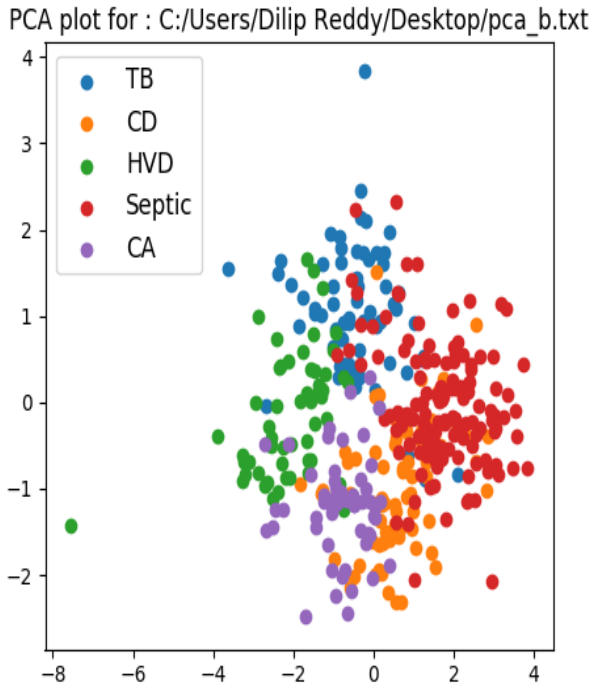
Plots for pca_a.txt:

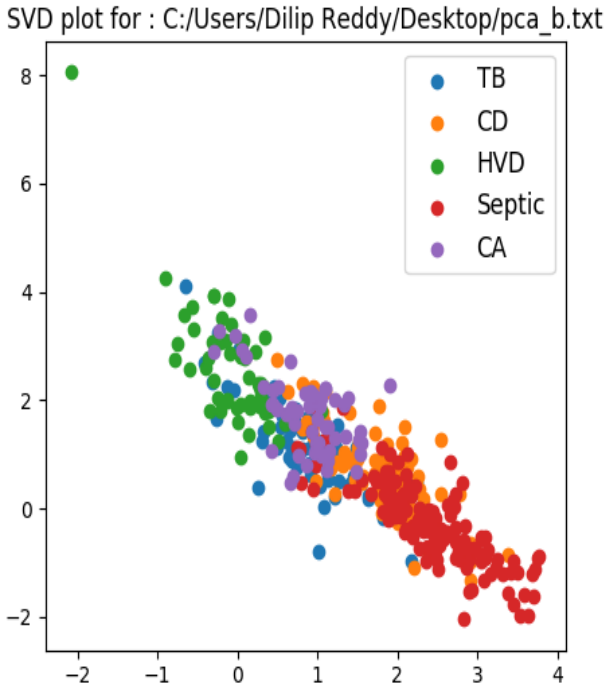


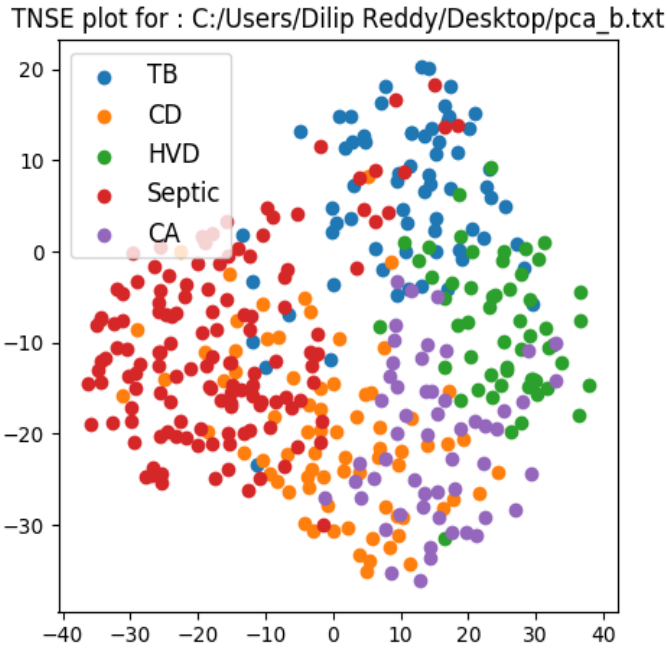




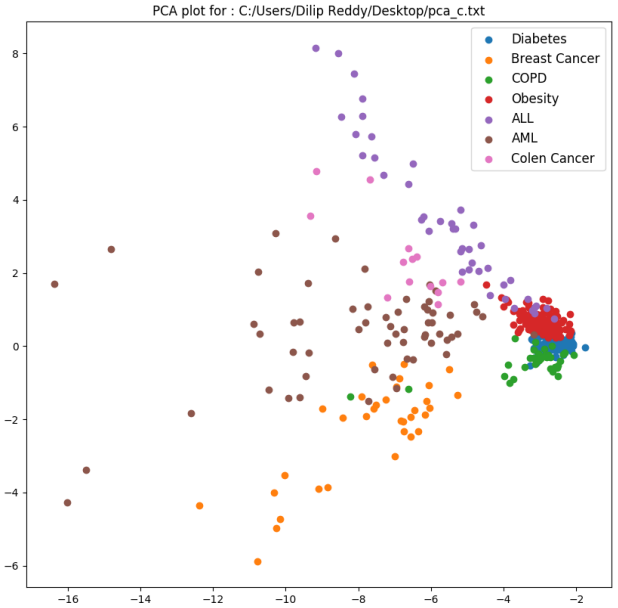
Plots for pca_b.txt:

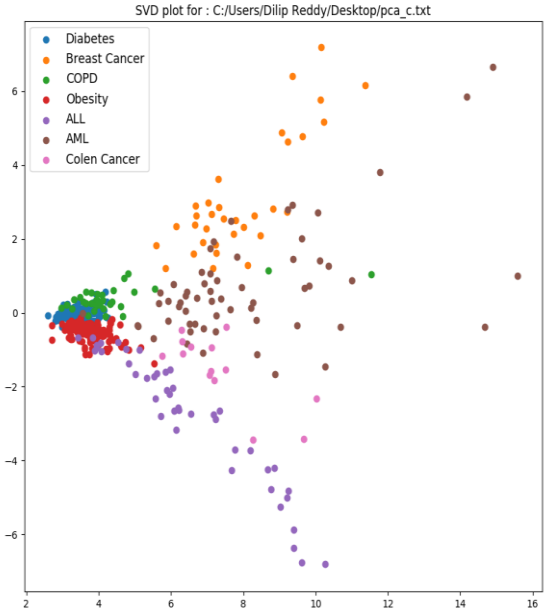


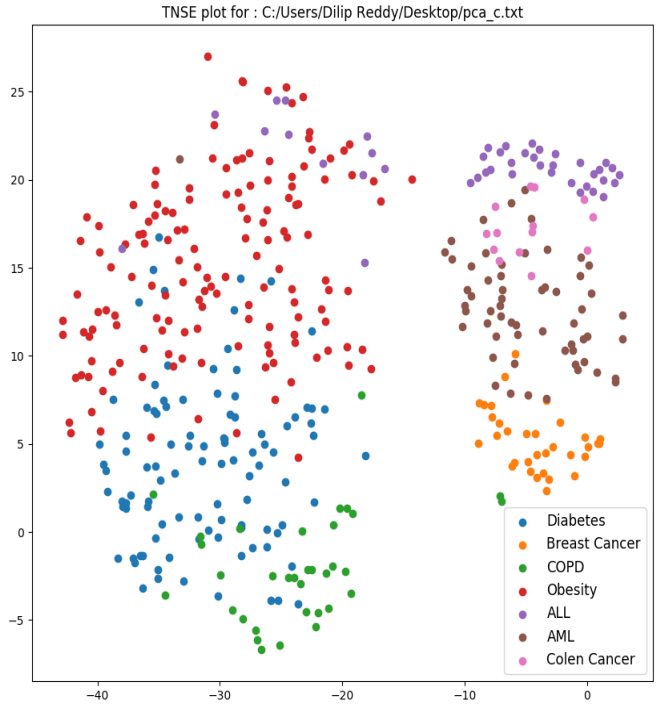




Plots for pca_c.txt:







Comparing the Results:

1. In General Principal Component Analysis is carried out using Eigen value decomposition of covariance matrix. PCA can also done Singular Value Decomposition(SVD) and t-Distributed Stochastic Neighbor Embedding (t-SNE).
2. Both PCA and SVD uses linear methods for dimensionality reduction. So, the results of PCA and SVD looks similar.
3. t-Distributed Stochastic Neighbor Embedding (t-SNE) uses non-linear methods for dimensionality reduction. t-SNE leads to huge computations on large datasets. So it is generally used for data visualization.
4. Due to above differences the plots for SVD and PCA looks similar where plots for t-SNE looks different.