Principal Component Analysis :

Implemented Principal Component Analysis (PCA) for the three datasets. The first method to do PCA was the eigen method as discussed in class.

Steps:

- Find the mean of each feature (column) and subtract that from data to get the mean centered data X:

  $X = D - D_{mean}$ where D is data, Dmean is the mean vector of features .
- Find the covariance matrix S using the formula:

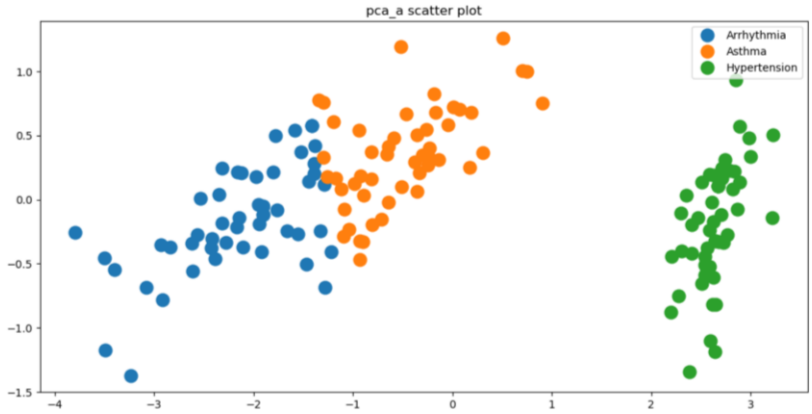
  $$S = \frac{1}{n} XX^T$$ where n is the number of data points.
- The covariance matrix is given as input to the function np.linalg.eigen. It returns the eigen value and corresponding eigen vector.
- Sort the eigen value in descending order and filter the top 2 eigen values along with their eigen vectors.
- Find the dot product of the mean centered data and the eigen vector.
- Append the resultant matrix with the disease feature column and do the scatter plot.

Np.linalg.svd funciton was used to do the Single Value Decomposition and sklearn's TSNE function was used for the non-linear dimensionality reduction. TSNE uses probabilistic approach and hence the plots are different with respect to the other methods.

You can see that the scatter plot of svd is a mirror image of the eigen method, if the input is the mean-centered data. But it is different when the input without being mean centered is used.

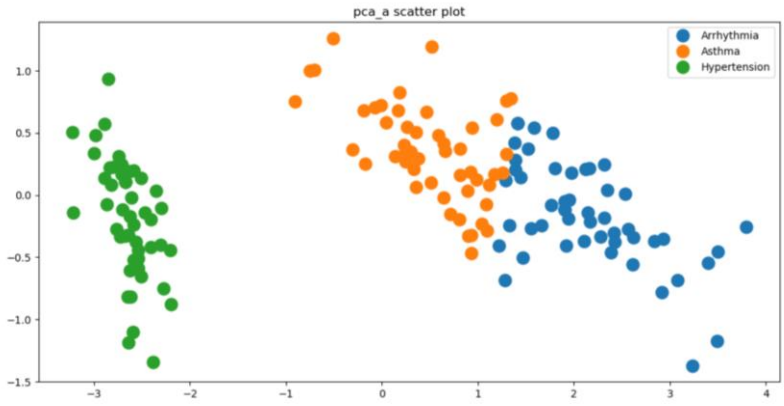
Pca_a.txt

Eigen value method :

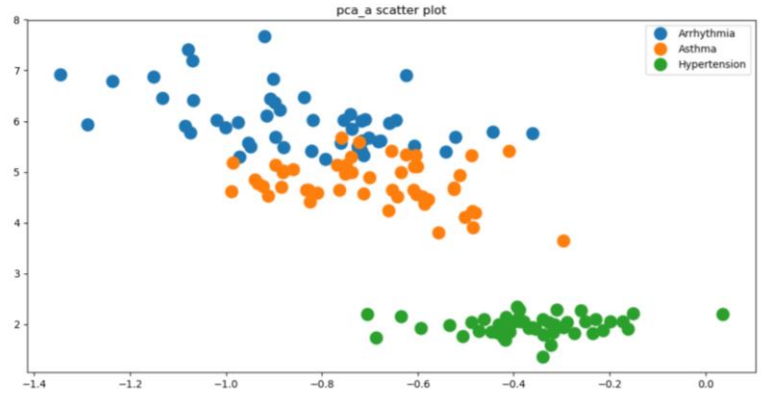


SVD method :

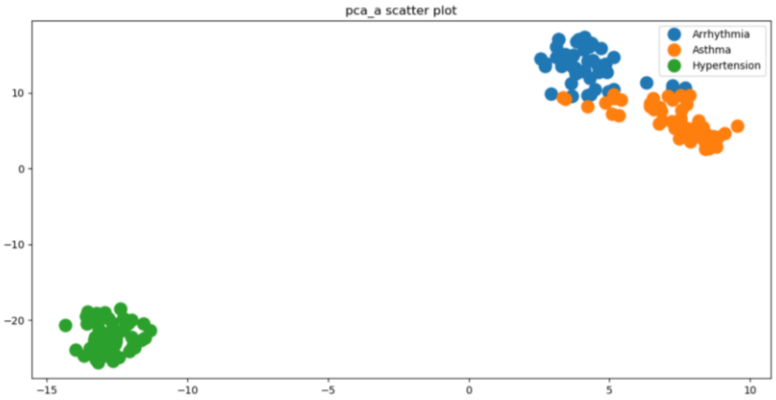
Using normalized data:



Using non-normalized data:

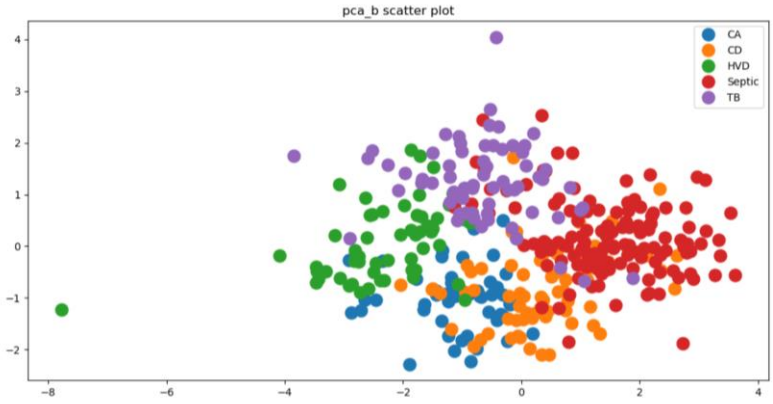


t-sne method :



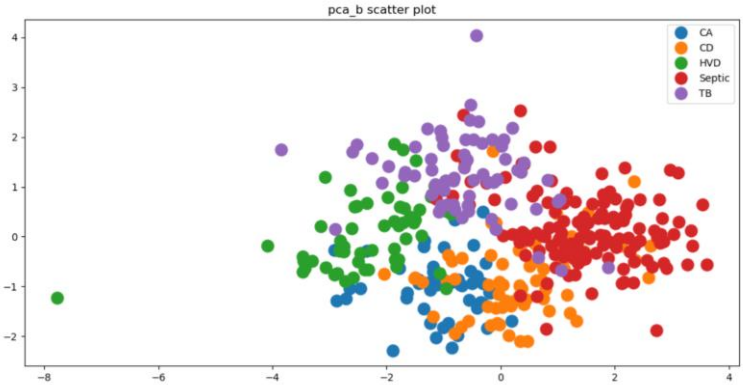
Pca_b.txt

Eigen value method :

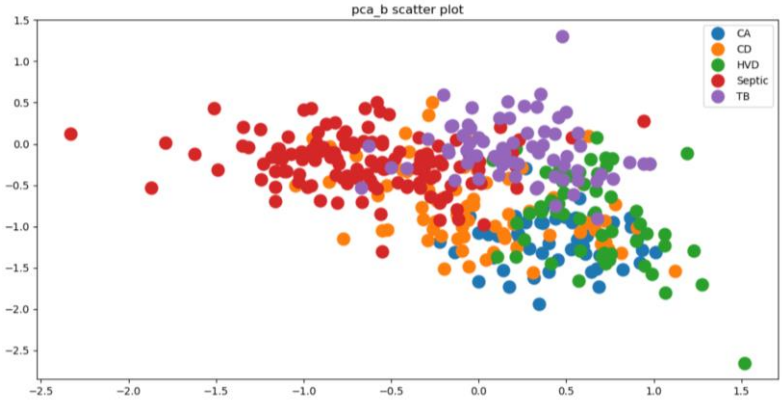


SVD method :

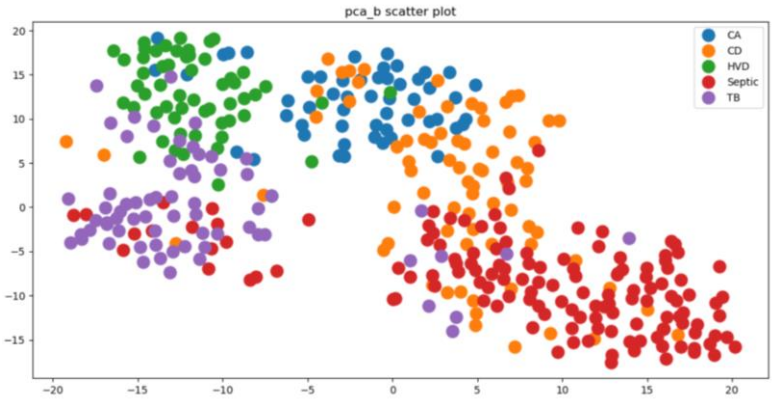
Using normalized data:



Using non-normalized data:

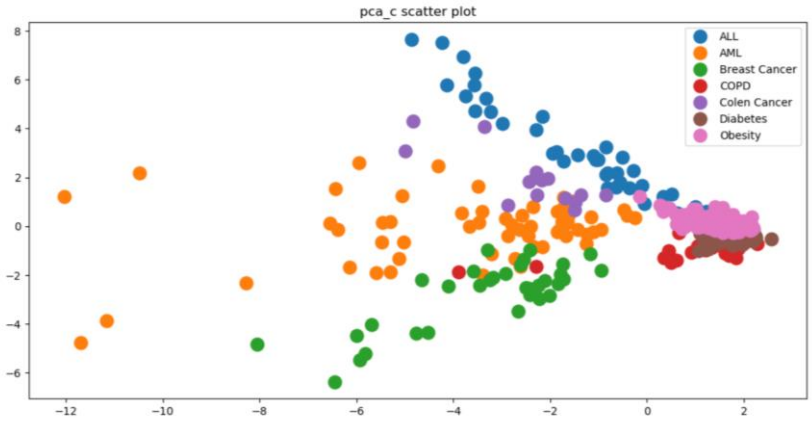


t-sne method :



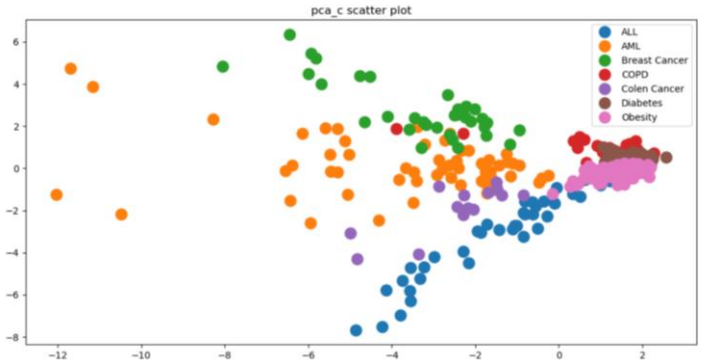
Pca_c.txt

Eigen value method :

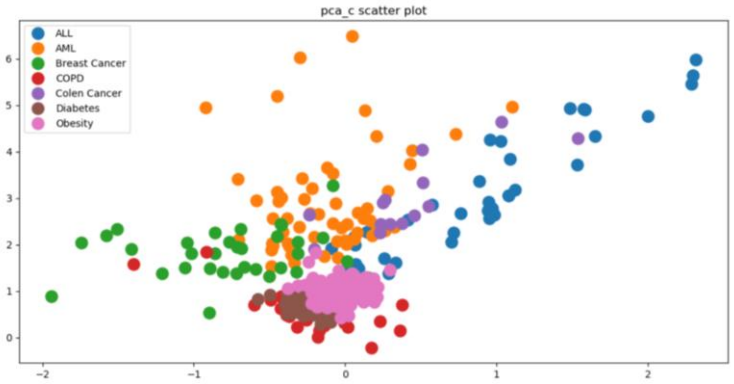


SVD method :

Using normalized data:



Using non-normalized data:



t-sne method :