CSE601 Data Mining and Bioinformatics

Project1 part1(PCA)

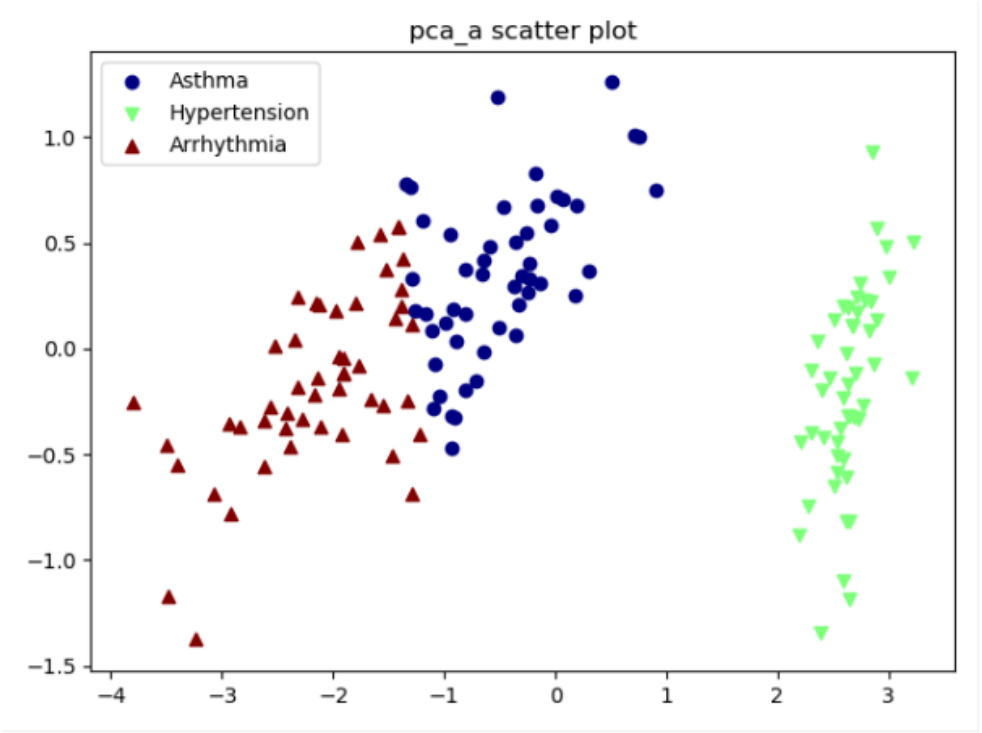
Kaige Gao 50320916
Junyang Li 50320301
Jingsong Li 50322345

PCA Flow:

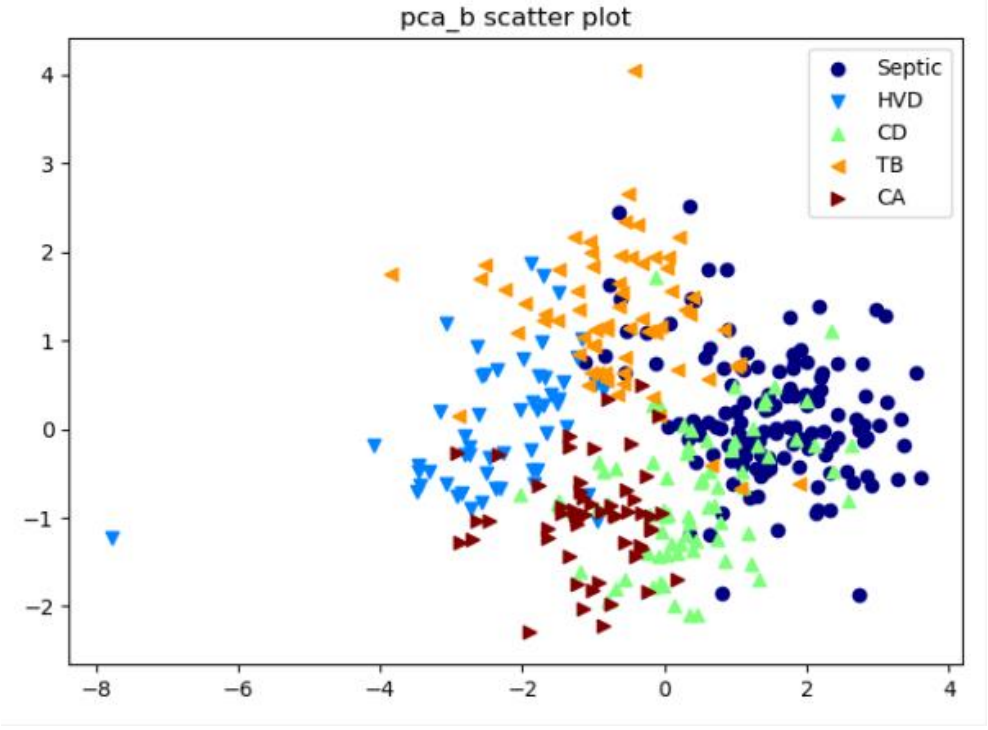
1.The file was load into program as an input matrix named **InputMatrix.**
2.Then call the method named **get_pcm(input_matrix)** that returns the principal components named **principle_components_matrix.**
3.In the **get_pcm,** an adjusted matrix named **adjusted_matrix** is calculated by subtracting column mean from value in each tuple in the input matrix
4.Co-variance matrix named **cov_matrix** is obtained by taking the product of adjusted matrix with its transpose and dividing it by total number of records.
5. A list of Eigen values named **eig_val** and list of eigen vectors named **eig_vec** are obtained from the covariance matrix.
6. Eigen vectors corresponding to top 2 eigen values are selected which form the principal components.
7. Each unique disease is assigned a unique integer value in a list named **DiseaseEncoded** obtained from a dictionary named **d.**
8. Further each unique integer value corresponds to a single color and shape in the plot.

Results:

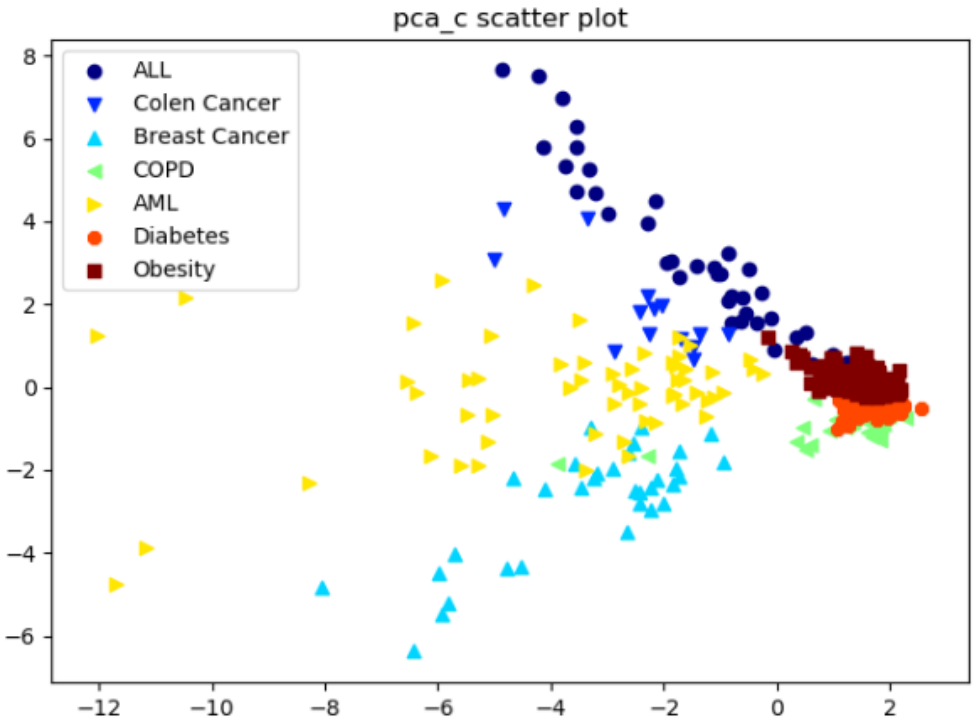
PCA:



“pac_a.txt” using PCA

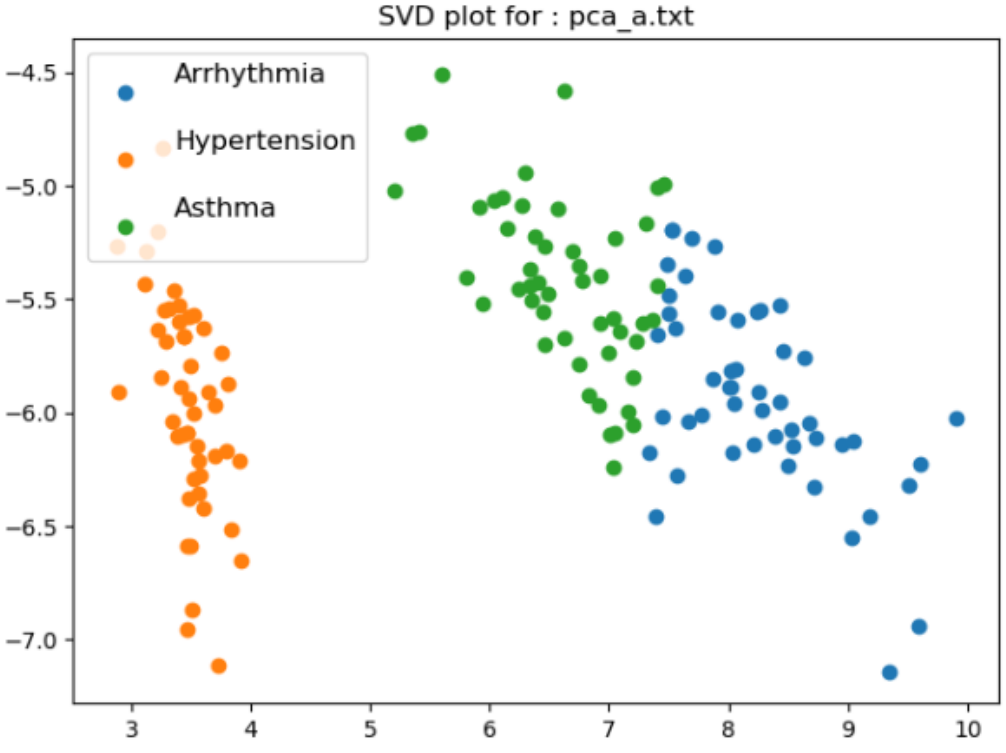


“pac_b.txt” using PCA

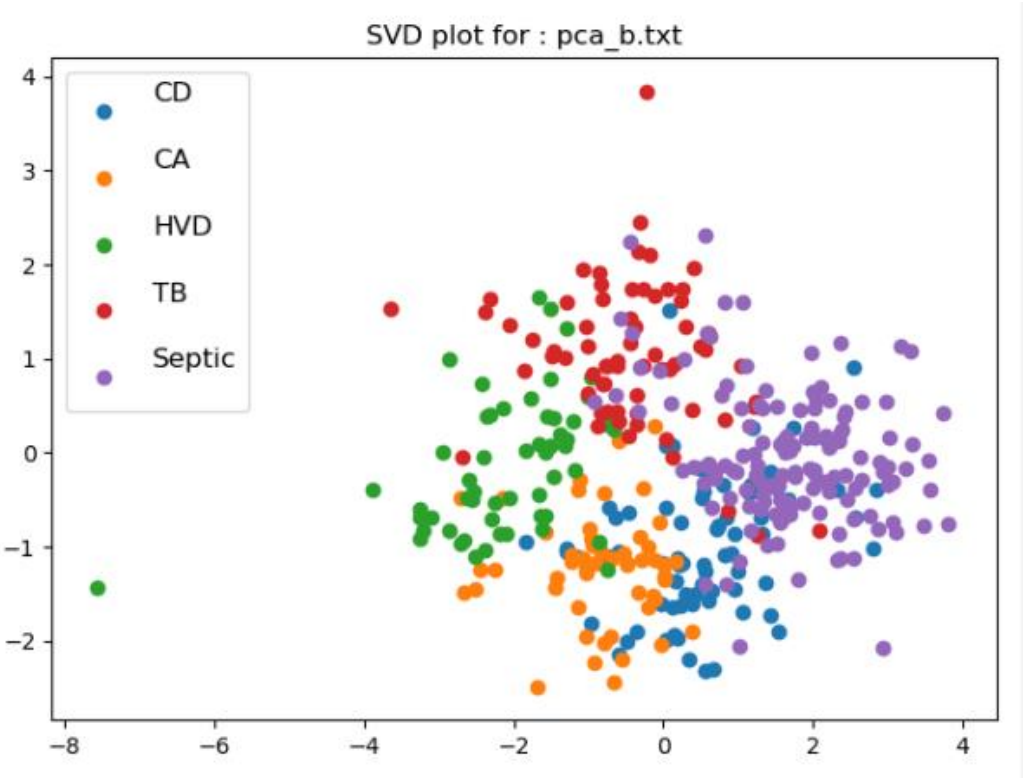


“pac_c.txt” using PCA

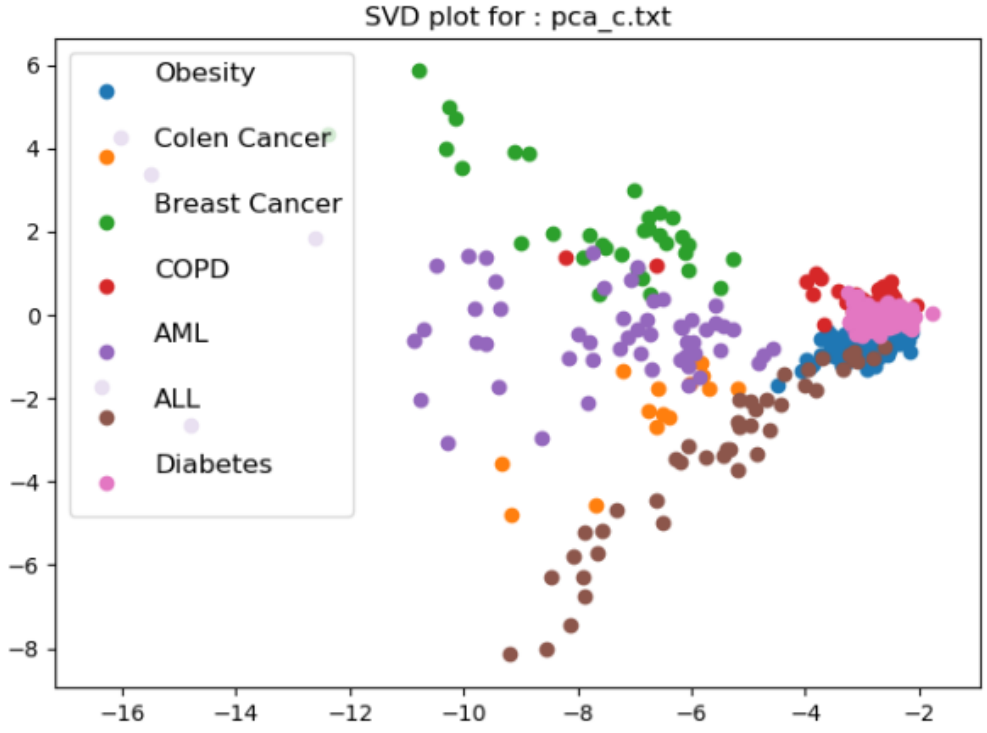
SVD:



“pac_a.txt” using SVD

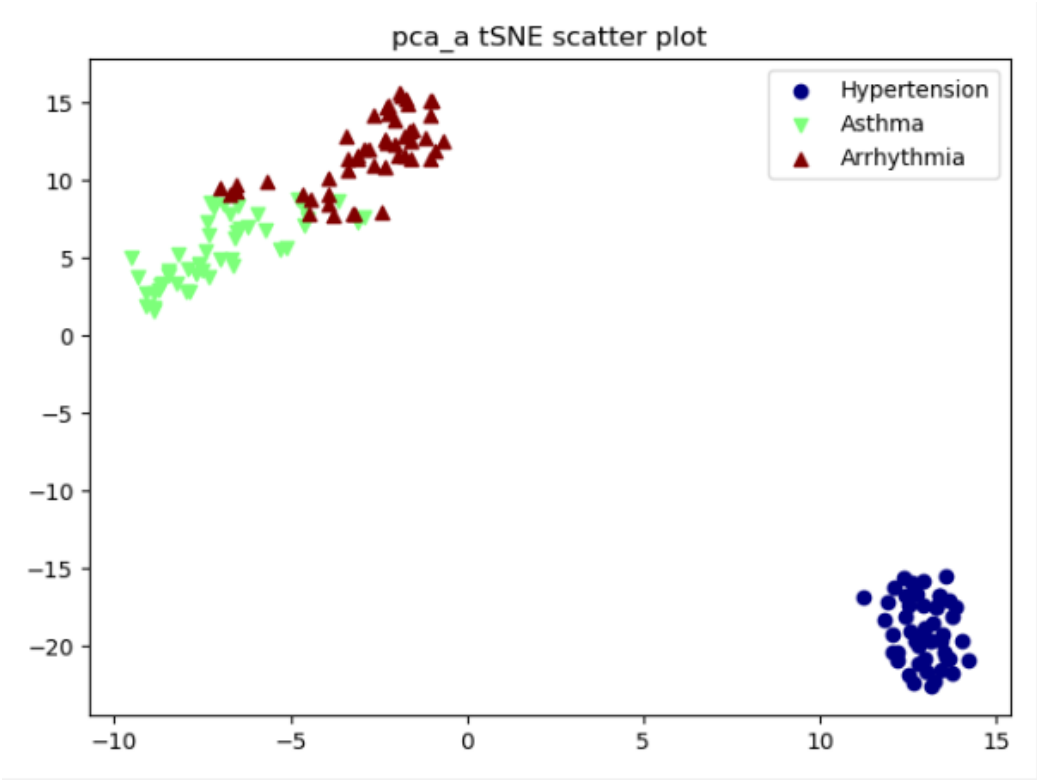


“pac_b.txt” using SVD

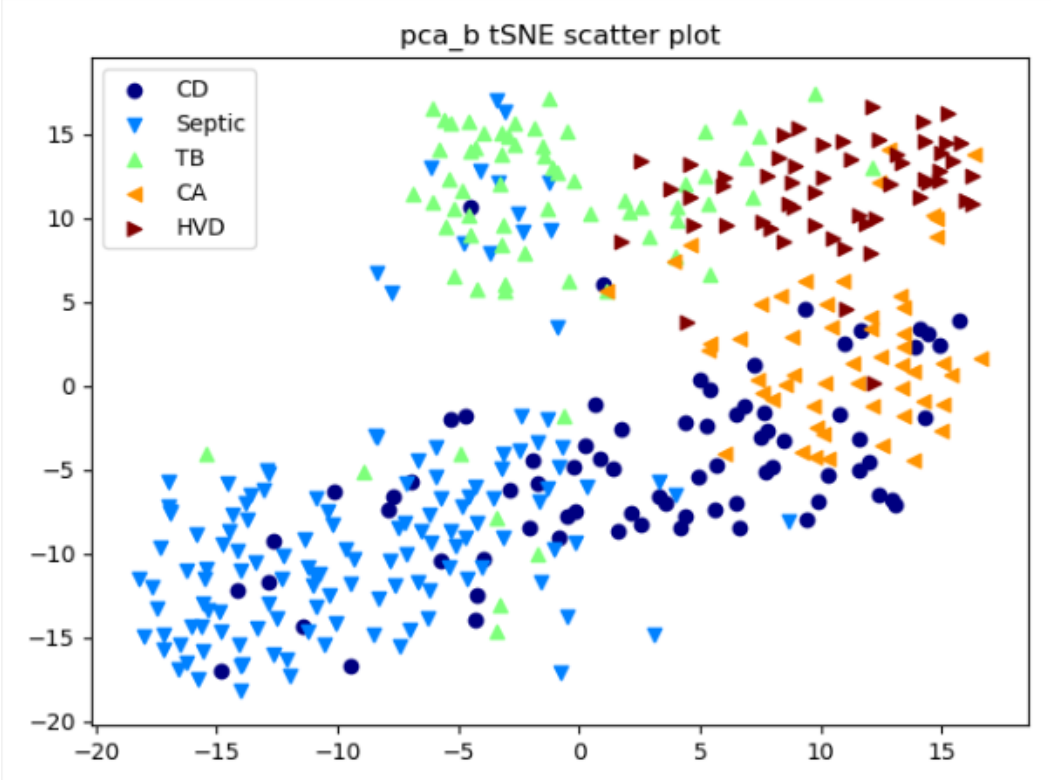


“pac_c.txt” using SVD

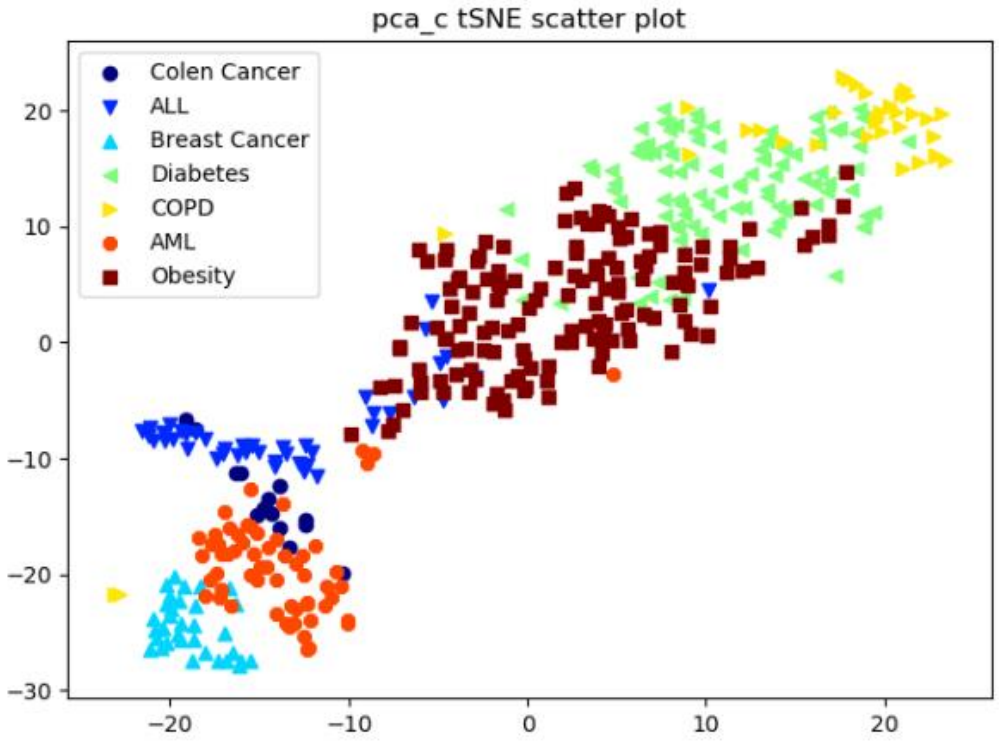
t-SNE:



“pac_a.txt” using t-SNE



“pac_b.txt” using t-SNE



“pac_c.txt” using t-SNE

Discuss:

SVD and PCA dimensionality reduction methods which are based on computing the eigenvalues and eigenvectors, while retaining the important information. The formulas which are used to calculate the covariance matrix for the two methods are also similar. Therefore, the end results for the two methods are also similar.

t-SNE is an another technique for dimensionality reduction and is particularly well suited for the visualization of high-dimensional datasets. Contrary to PCA it is not a mathematical technique but a probabilistic one. This method minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a
distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding.