CSE 601
Principal Component Analysis Report

Submitted by

Paritosh Kumar Velalam (pvelalam) (50295537)

Sampreeth Boddi Reddy (sboddire) (50298591)

i. Scatter Plots:

1. **Data Set : pca_a.txt**
   Algorithm: PCA

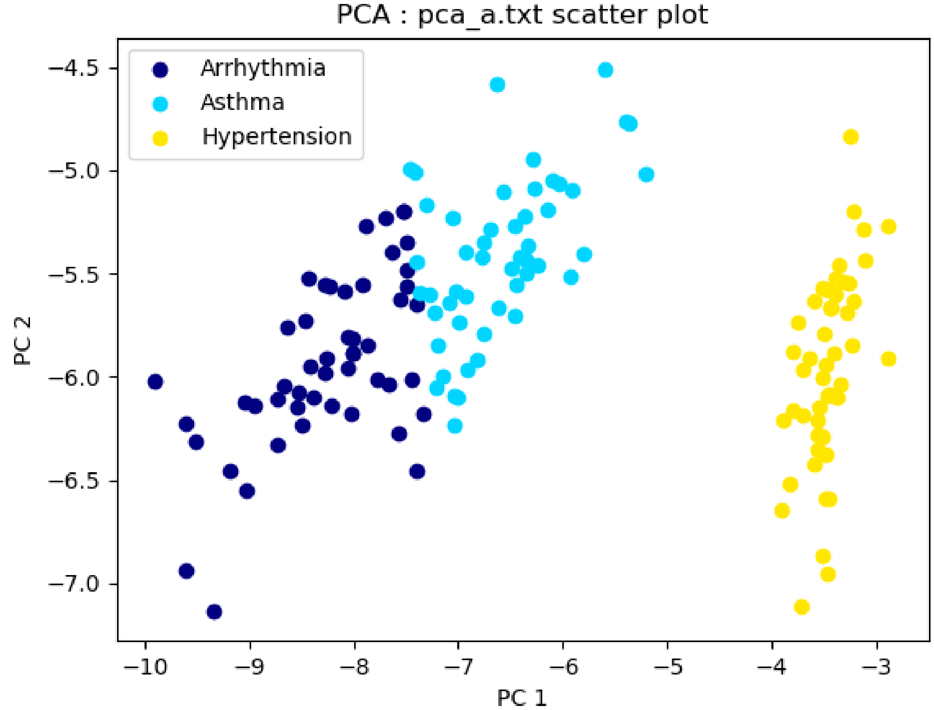

Figure1. Scatter Plot for Dataset: pca_a.txt and Algorithm: PCA

2. **Data Set : pca_a.txt**
   Algorithm: SVD

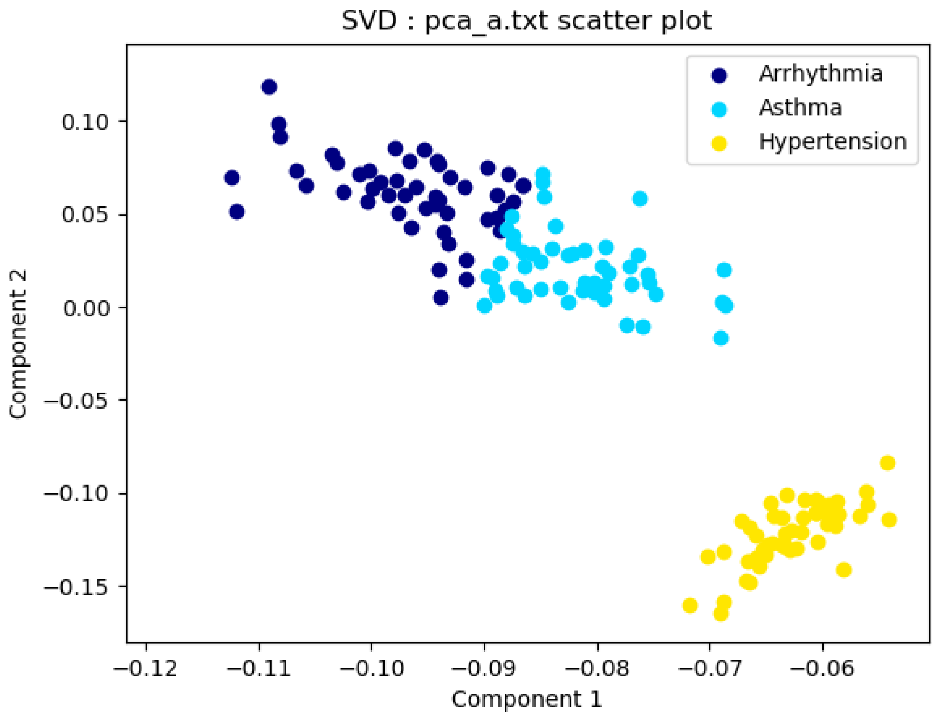

Figure2. Scatter plot for Dataset: pca_a.txt and Algorithm: SVD

3. Data Set: pca_a.txt
Algorithm: t-SNE

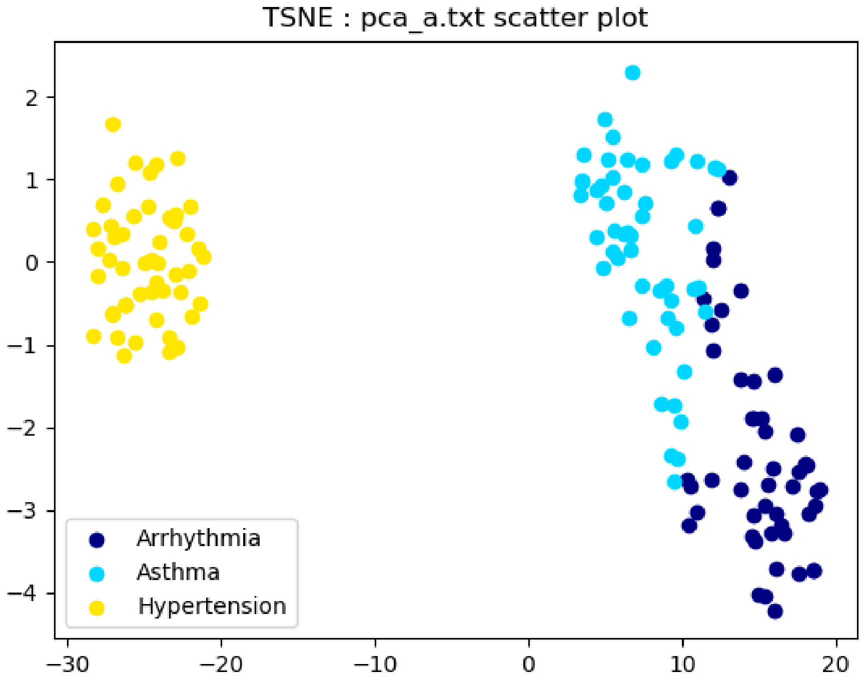

Figure 3. Scatter plot for Dataset: pca_a.txt and Algorithm: t-SNE

4. Data Set: pca_b.txt
Algorithm: PCA

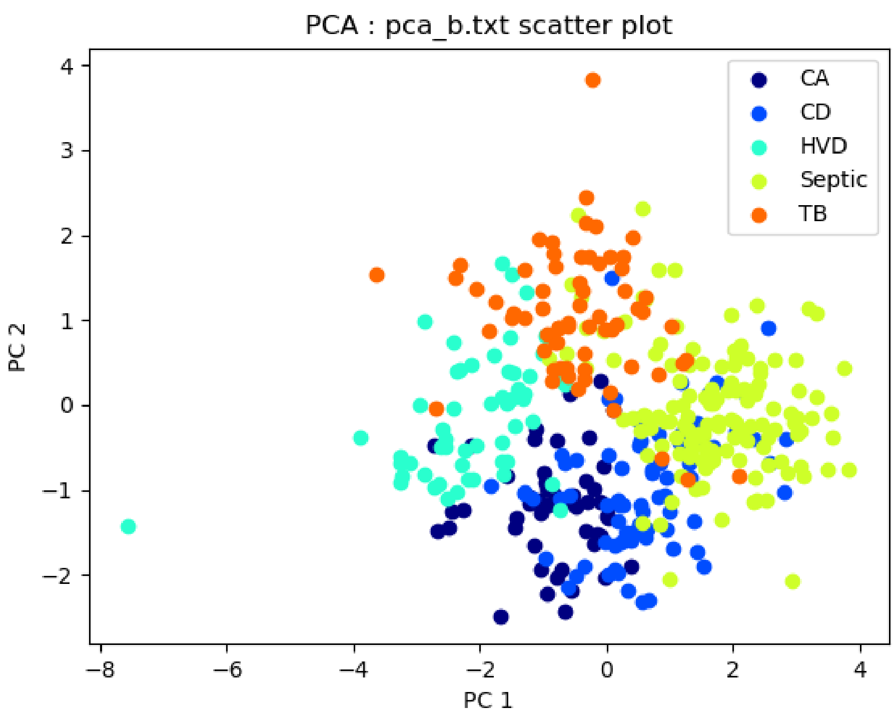

Figure 4. Scatter plot for Dataset: pca_b.txt and Algorithm: PCA

5. **Data Set: pca_b.txt**
   Algorithm: SVD

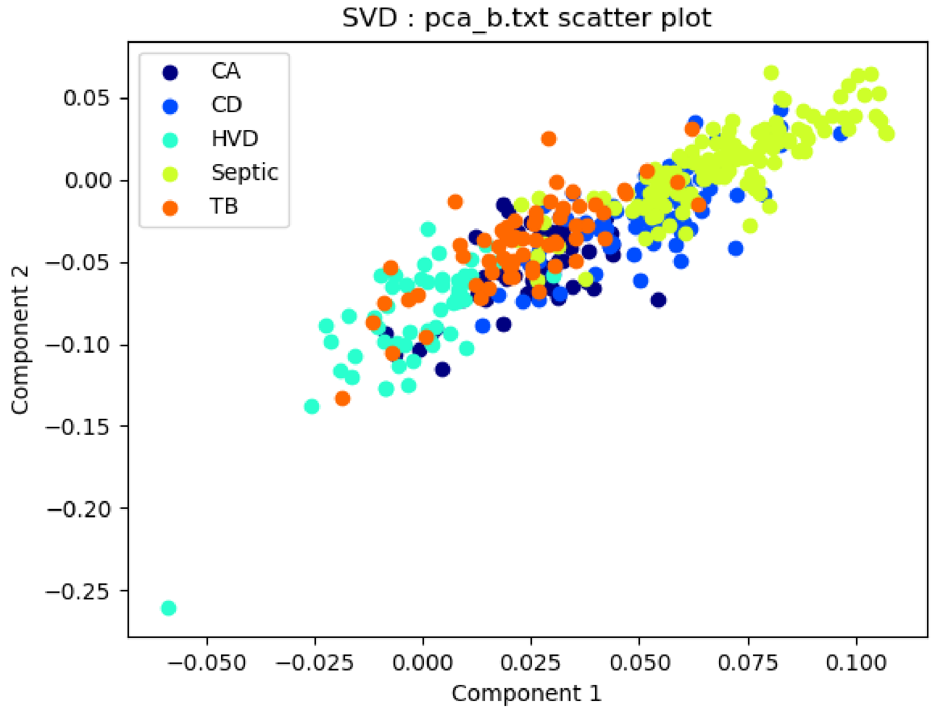

Figure 5: Scatter plot for Dataset: pca_b.txt and Algorithm: SVD

6. **Dataset: pca_b.txt**
   Algorithm: t-SNE

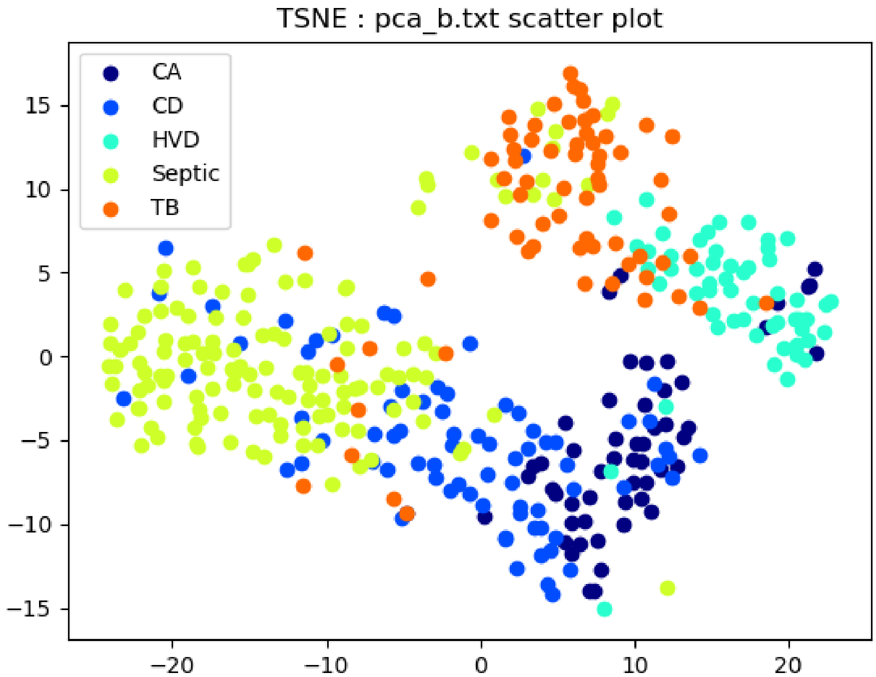

Figure 6. Scatter plot for Dataset: pca_b.txt and Algorithm: t-SNE

7. **Dataset: pca_c.txt**
   Algorithm: PCA

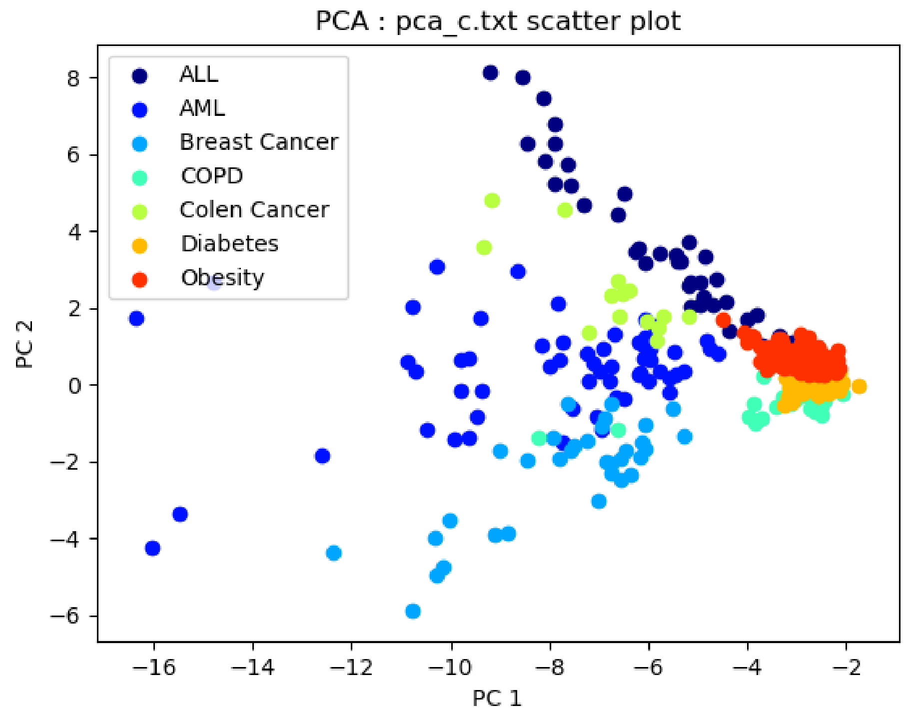

Figure 7. Scatter plot for Dataset: pca_c.txt and Algorithm: PCA

8. **Dataset: pca_c.txt**
   Algorithm: SVD

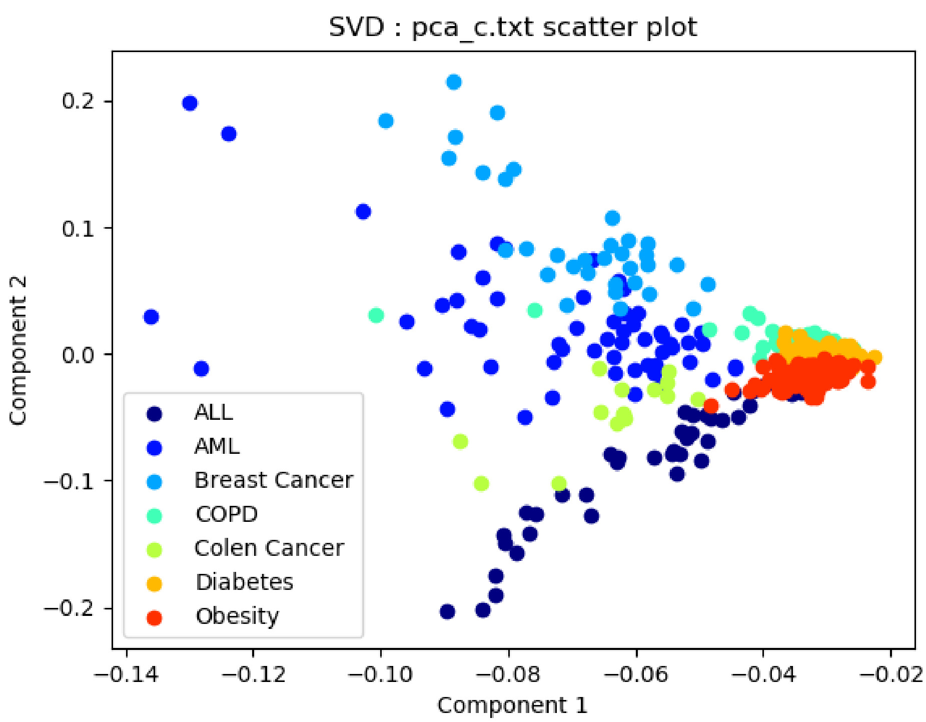

Figure 8. Scatter plot for Dataset: pca_c.txt and Algorithm: SVD

9. **Dataset: pca_c.txt**
   Algorithm: t-SNE

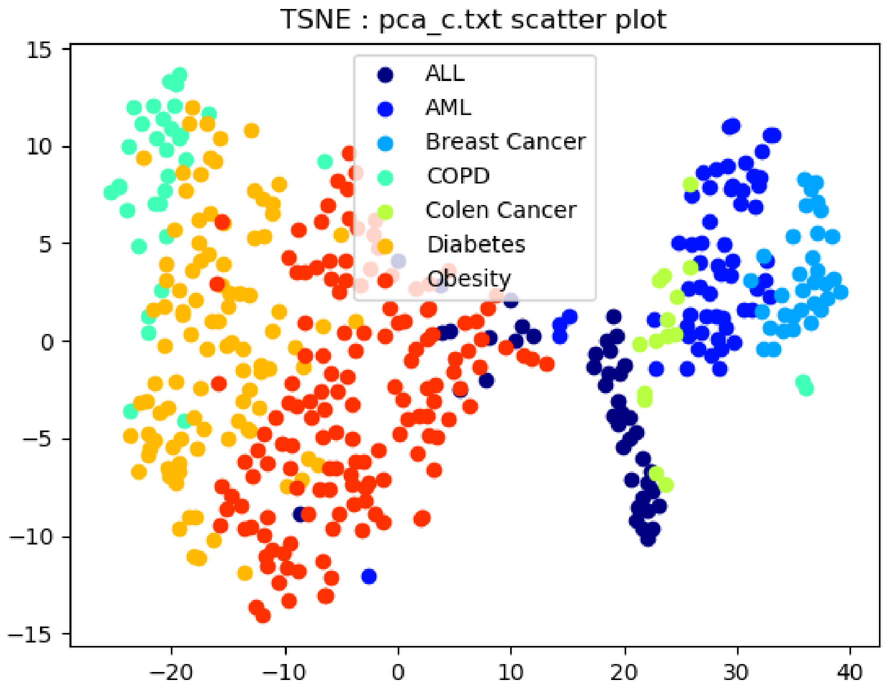

Figure 9. Scatter plot for Dataset: pca_c.txt and Algorithm: t-SNE

PCA:

- Each point is color attributed to the disease it corresponds to in the given dataset.
- Scatter plot shows the variation along two principal components.

SVD:

- Each point is color attributed to the disease it corresponds to in the given dataset.
- Numpy Linear Algebra package is used for Singular Value Decomposition of given dataset.
- Components 1 and 2 signify the largest variations.

TSNE:

- Each point is color attributed to the disease it corresponds to in the given dataset.
- TSNE from sklearn.manifold package is used for implementing TSNE algorithm.
- Number of components is set to 2 for obtaining representation in two dimensions.
- The scatter plot shows that the same disease points form clusters.

ii.

Flow of PCA Algorithm:

a. Calculate mean X_m of input feature vector X.
b. Calculate $X_n = X - X_m$
c. Calculate covariance $S = (1/n-1)X_n^TX$. "n" is the number of samples of X
d. Calculate Eigen values and corresponding Eigen vectors of S.
e. Form a matrix E with first column as Eigen vector of highest Eigen value and second column as Eigen vector of second highest Eigen value for reduction of X to two dimensions.
f. Calculate $X_d = X.E$. X_d is the representation of X in two dimensions.

Steps followed for implementation of PCA Algorithm:

a. Read the input data and form the input feature matrix X by removing the last column from the input data.
b. Extract the labels from the last column of input data.
c. Calculate the mean of each row of X (X_m) using np.mean.
d. Subtract Xm from X (Xn = X - Xm).
e. Calculate covariance $S = (1/n-1)X_n^TX$. "n" is the number of rows of X.
f. Calculate Eigen values and Eigen vectors of S using np.linalg.eig.
g. Select the first two columns of Eigen vector matrix returned from step e and form a matrix E. The Eigen vectors are returned in the decreasing order of Eigen values in step e.
h. Calculate Xd = X.E using np.dot.

Existing packages used:

a. numpy for converting dataset to array.
b. np.linalg from numpy to compute Eigen values, Eigen vectors and SVD computation.
c. TSNE from sklearn to implement TSNE algorithm.
d. Matplotlib.pyplot for plotting scatter plots.

Results discussion:

- PCA centers the data and then rotates the centered data to obtain points with maximum variance as the top principal components.
- SVD corresponds to compactly summarizing the data and the way it deviates from zero.
- The results of PCA and SVD will be same when mean centered data is used for SVD computation. Also the plots will be similar in this case.
- Contrary to PCA and SVD, TSNE is a probabilistic technique. TSNE represents the data in lesser dimensions by matching the distributions in the data.
- TSNE shows more clear variations in data as compared to PCA but is computationally heavy.