Data Mining

Project 1: Dimensionality Reduction & Association Analysis

Team No. 39

Arvind Srinivass Ramanathan arvindsr 50205659
Naveen Muralidhar Prakash naveenmu 50208032
Senthil Kumar Laguduva Yadindra Kumar laguduva 50207553

Scatter Plots:

Observation:

Whilst using the mean centered matrix for PCA and SVD we get mirror images of each other in graphs. The graph for t-SNE is observed to be new and we cannot guarantee reproducibility of the graph as it uses a probabilistic approach. Also, in t-SNE we directly move the data through dimensions thus not generating eigenvectors to form new data.

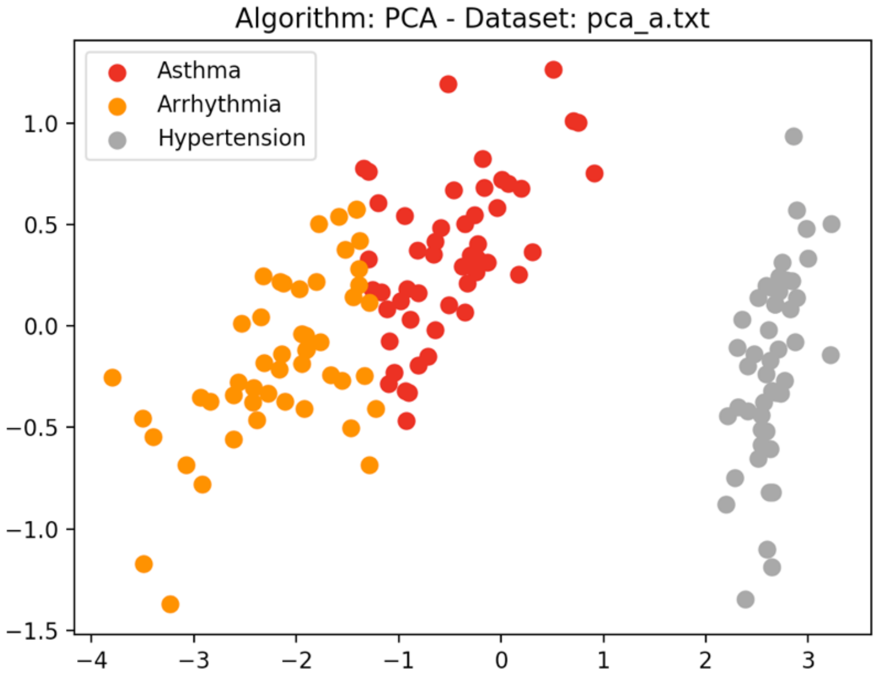
The graphs plotted are for the implementation of PCA algorithm, SVD algorithm and t-SNE algorithm. In the implementation of PCA we begin by mean centering the feature matrix. However, we use both the feature matrix and the mean centered matrix for SVD and t-SNE matrix.

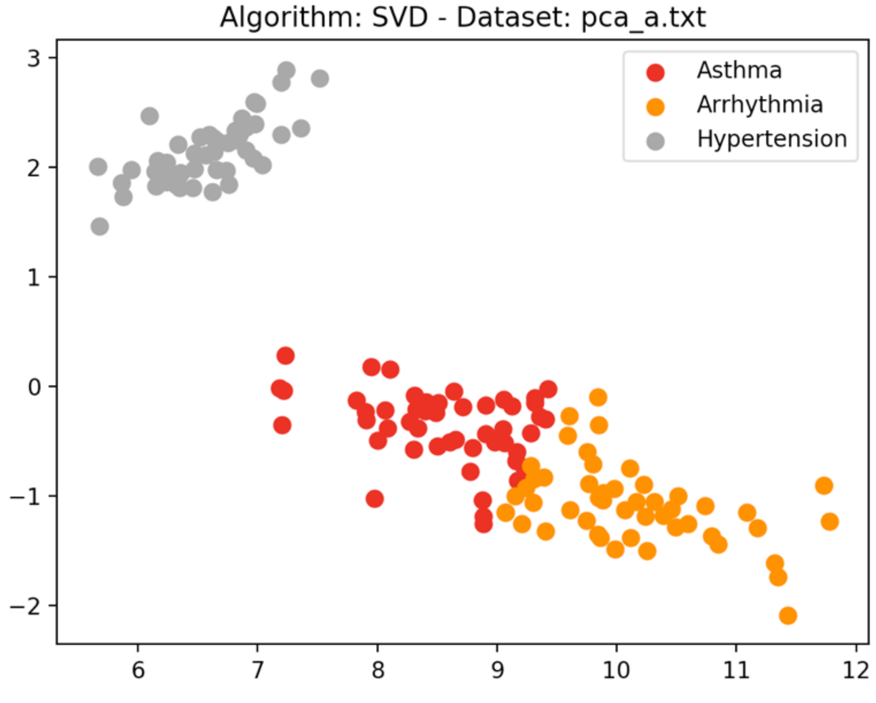
PCA algorithm is performed by eigenvalue decomposition of the data covariance matrix. PCA begins with mean centering of the feature matrix. Upon performing the PCA algorithm we generate the maximum difference along any number of dimensions which is less than the number of dimensions in the original data. PCA is an analysis approach.

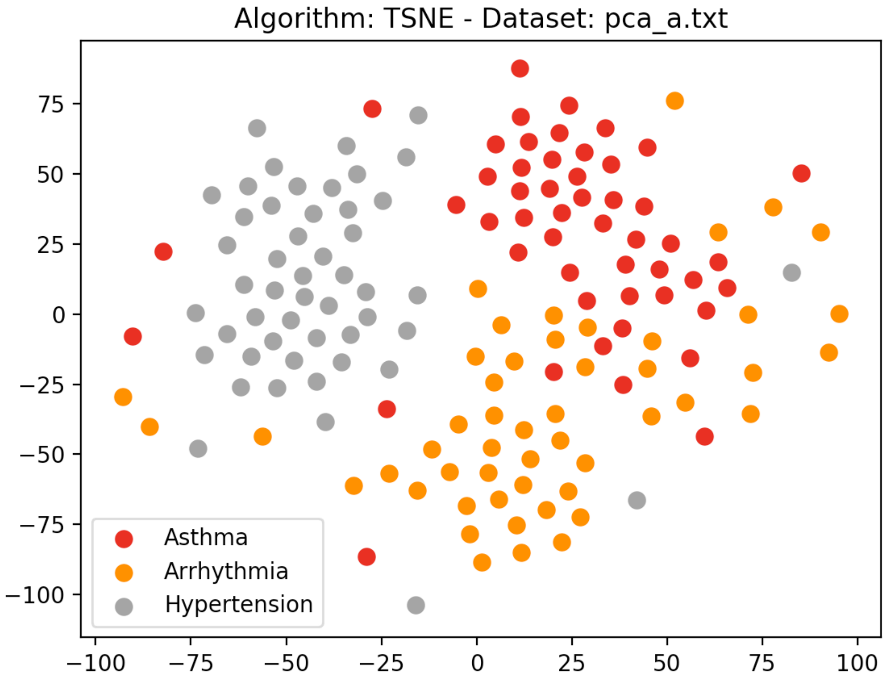
The clustering of PCA and SVD is different because in SVD implementation we use the feature matrix directly without mean centering the features. The SVD is a numerical method but it is said to have higher mathematical accuracy and is hence preferred.

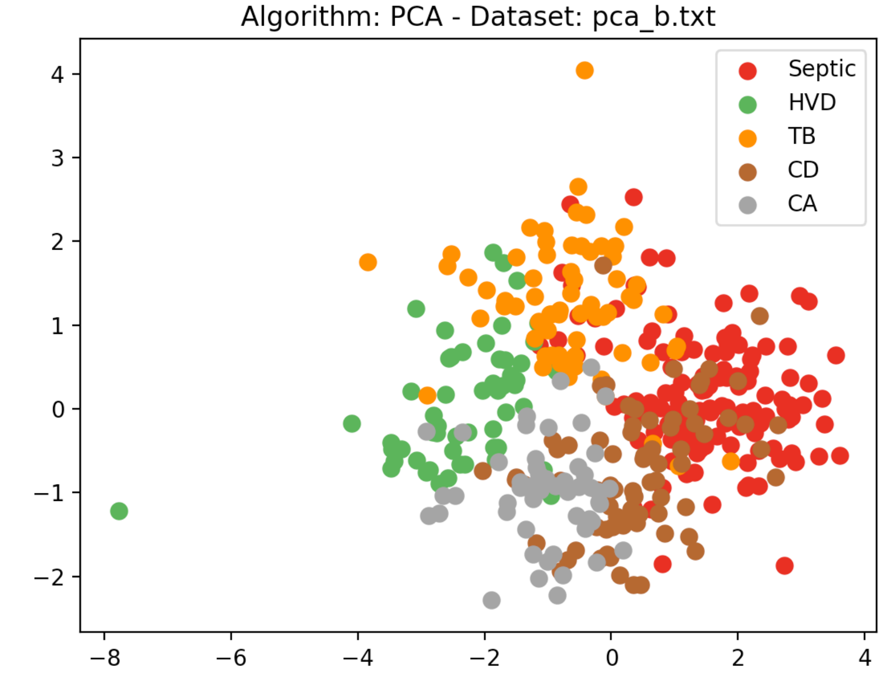
In conclusion, the implementation of PCA will differ based on the application. Different applications will require us to perform PCA on either the feature matrix or the mean centered feature matrix. In some applications if the data is not sensitive to the mean of the distribution we can center the matrix. In other cases where data is not sensitive to higher values we can directly use the feature matrix. Similarly, the same applies for SVD.

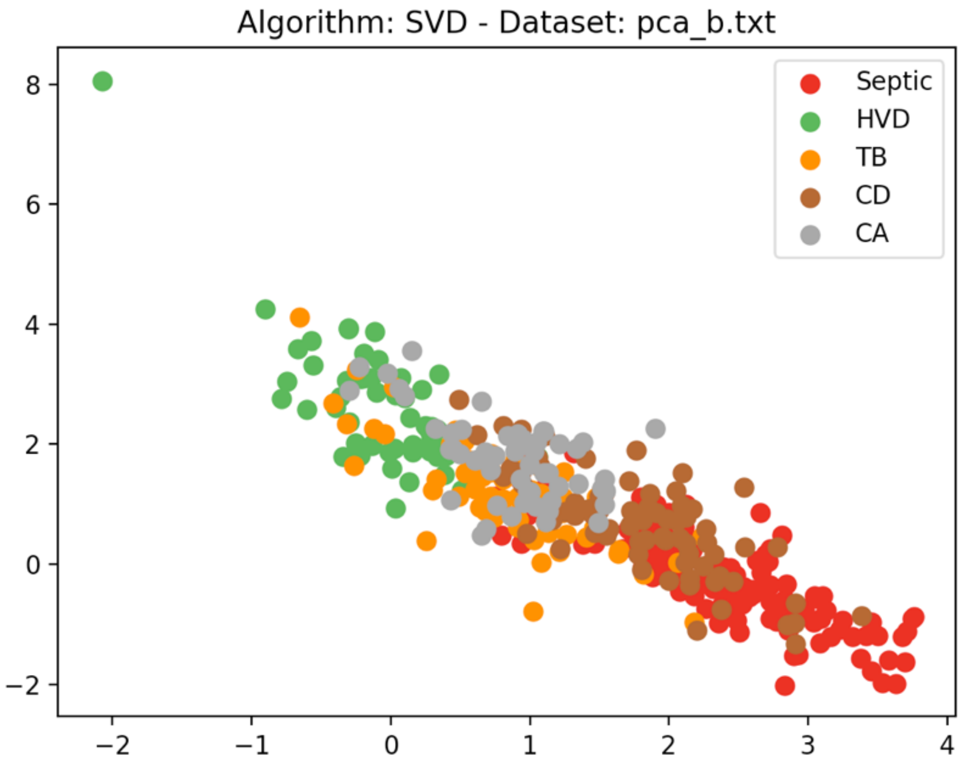
It is a non-linear dimensionality reduction technique for mapping higher dimensionality data into lesser dimensions. This technique brings similar points together and moves away dissimilar points. t-SNE is performed by first constructing a probability distribution over pairs of high dimensional objects in such a way that similar objects have higher probability of being picked while the dissimilar points have lesser probability of being picked. This is followed by defining a similar probability distribution over the paints in a low-dimensional map and it minimizes the divergence between the two distributions with respect to the location of the points in the map. PCA can still be favored over t-SNE because of its deterministic nature. Also, using PCA will generate a new axes system which can be used to project new data.

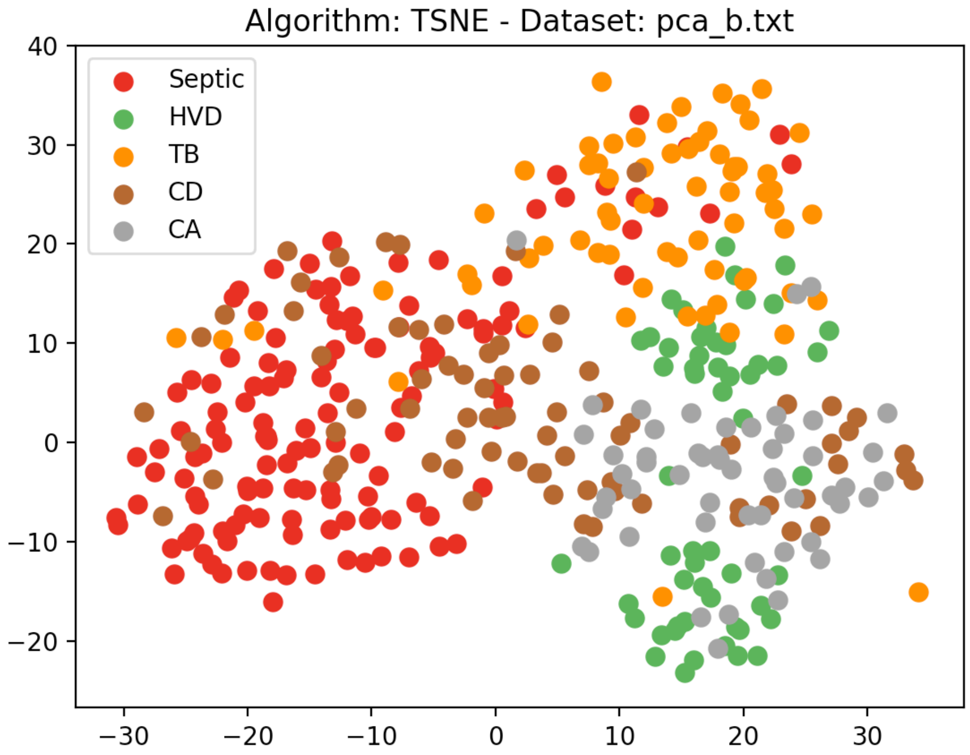

Algorithm: PCA - Dataset: pca_a.txt
Asthma
Arrhythmia
Hypertension
1.0
0.5
0.0
−0.5
−1.0
−1.5
−4
−3
−2
−1
0
1
2
3

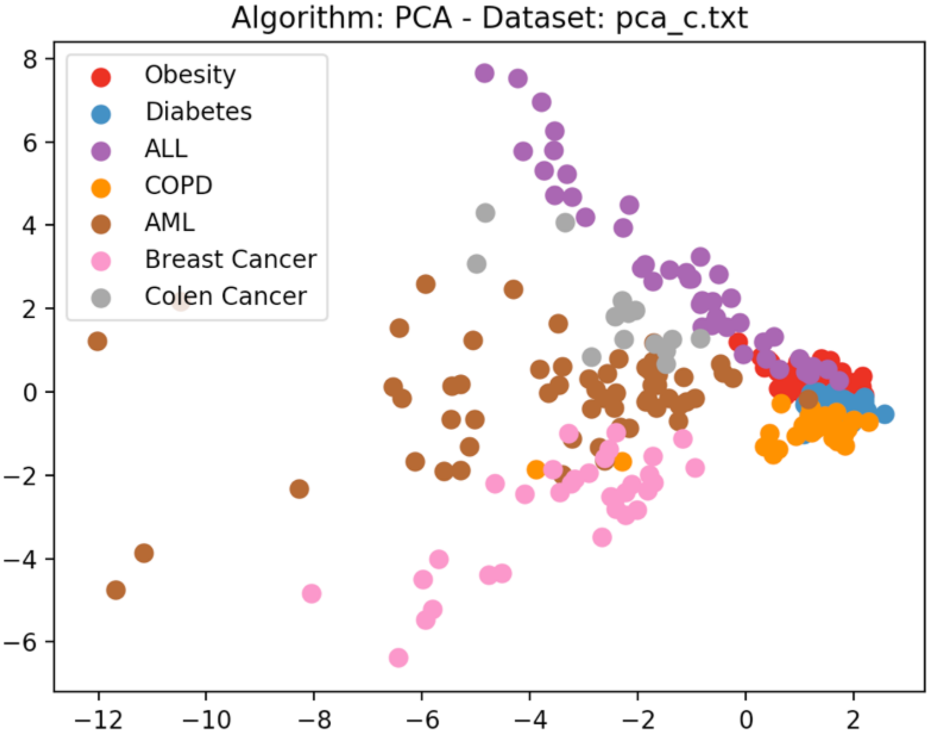

Algorithm: SVD - Dataset: pca_a.txt
Asthma
Arrhythmia
Hypertension
3
2
1
0
−1
−2
6
7
8
9
10
11
12

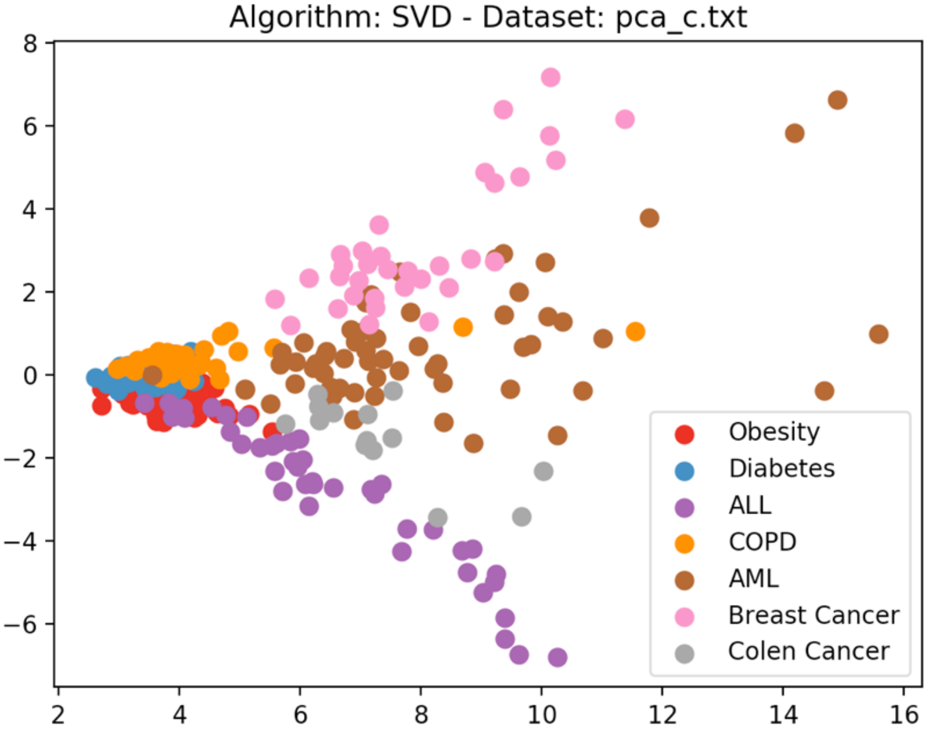

Algorithm: TSNE - Dataset: pca_a.txt

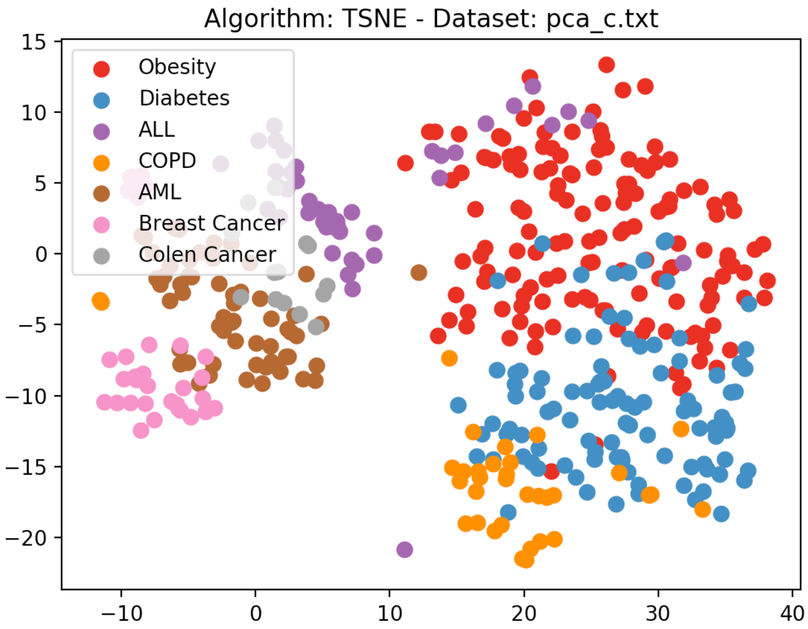

Algorithm: PCA - Dataset: pca_b.txt


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

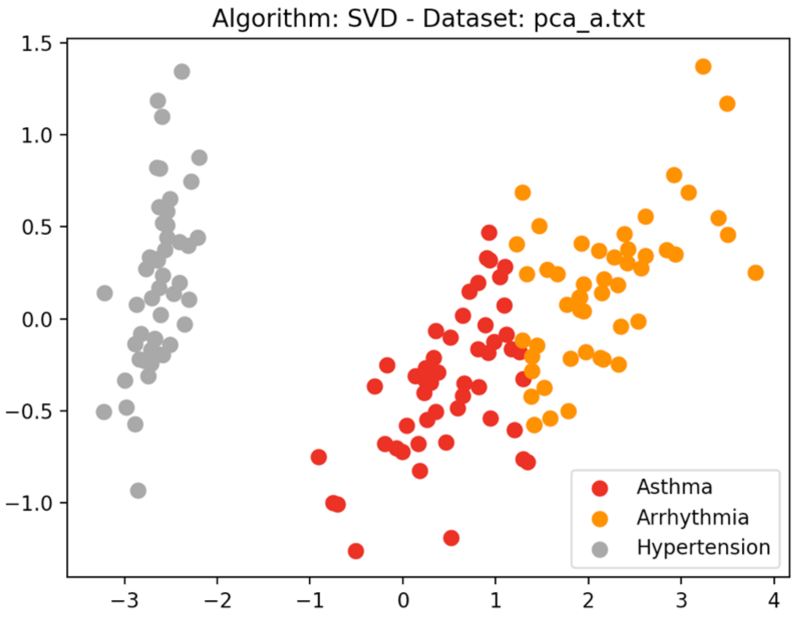

Algorithm: TSNE - Dataset: pca_b.txt
Septic
HVD
TB
CD
CA
40
30
20
10
0
−10
−20
−30
−20
−10
0
10
20
30

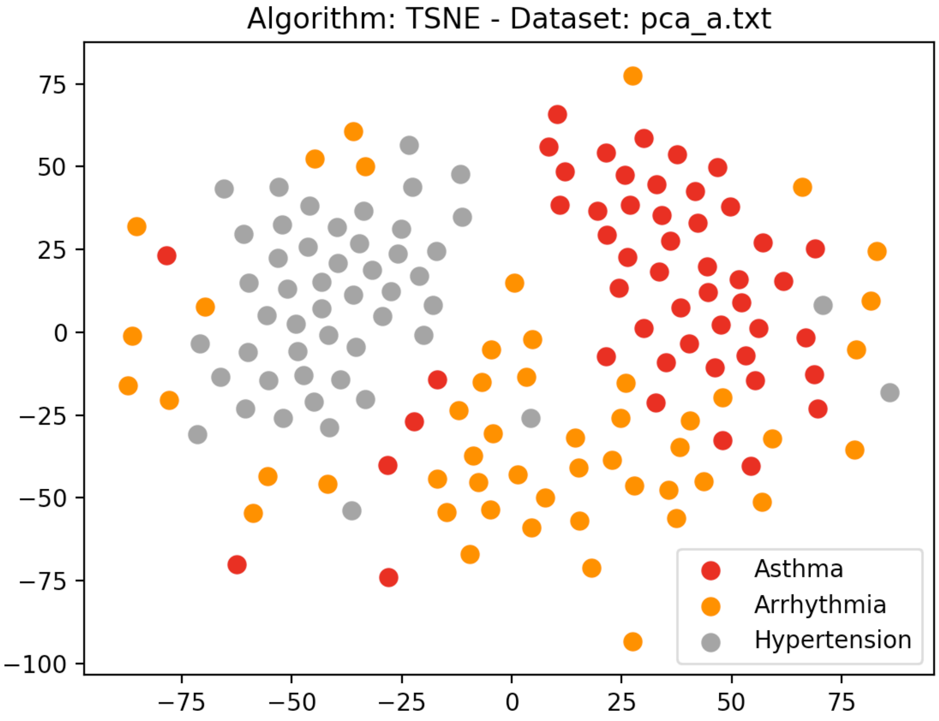

Algorithm: PCA - Dataset: pca_c.txt
Obesity
Diabetes
ALL
COPD
AML
Breast Cancer
Colen Cancer
8
6
4
2
0
−2
−4
−6
−12
−10
−8
−6
−4
−2
0
2

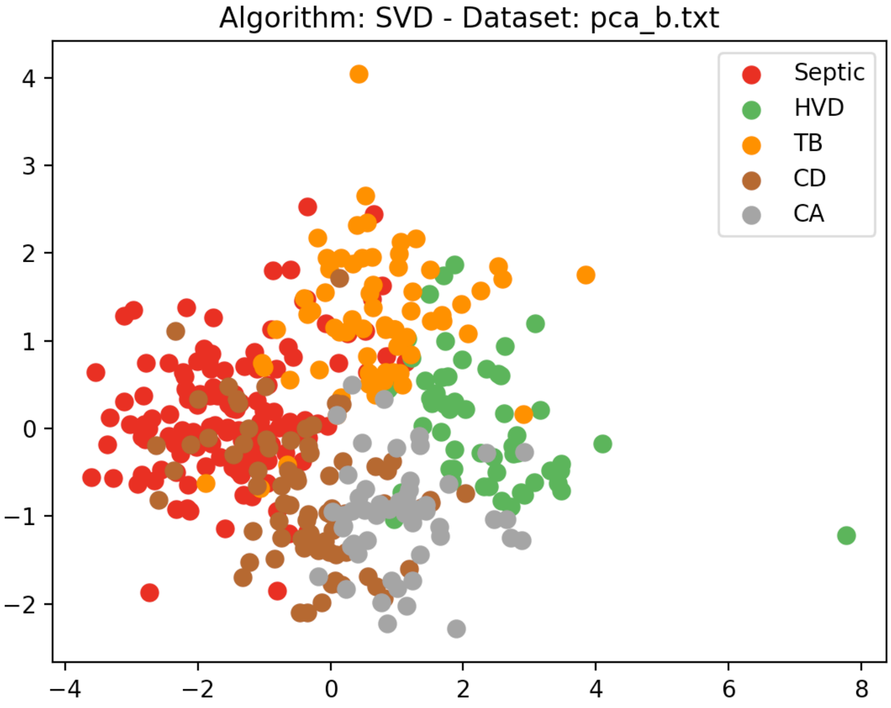

Algorithm: SVD - Dataset: pca_c.txt
Obesity
Diabetes
ALL
COPD
AML
Breast Cancer
Colen Cancer
8
6
4
2
0
−2
−4
−6
2
4
6
8
10
12
14
16

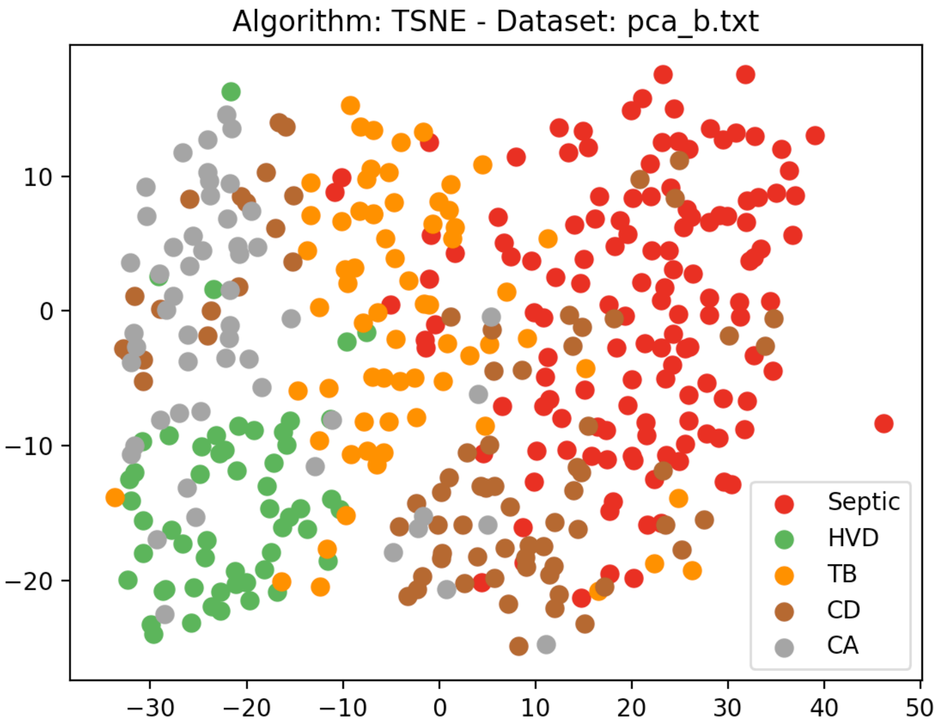


When we run SVD and t-SNE on the feature matrix after mean centering the feature matrix we get the following graphs for pca_a.txt, pca_b.txt and pca_c.txt.

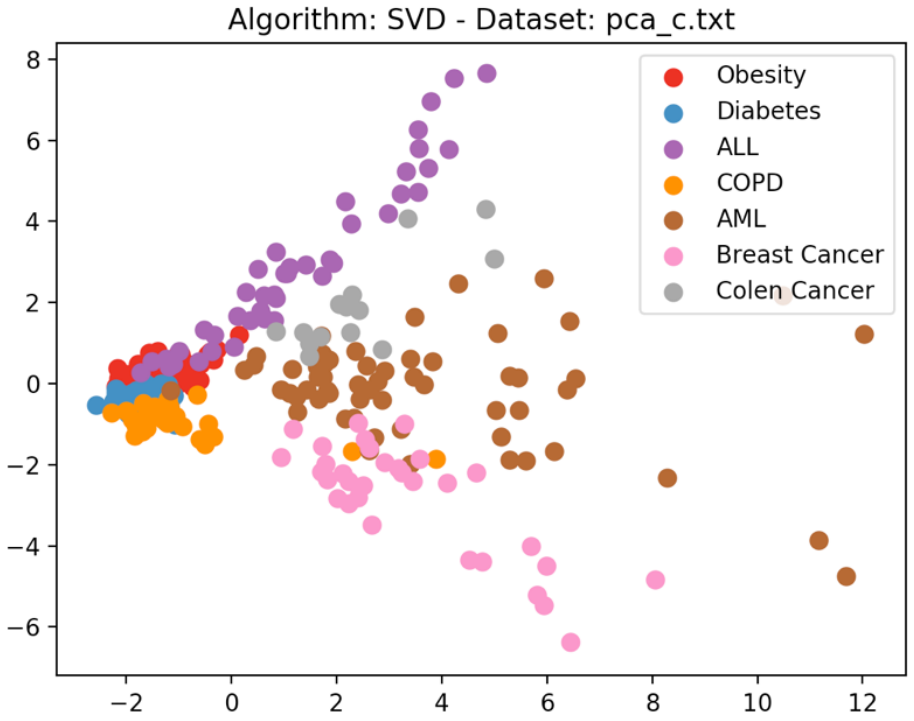



Algorithm: TSNE - Dataset: pca_a.txt

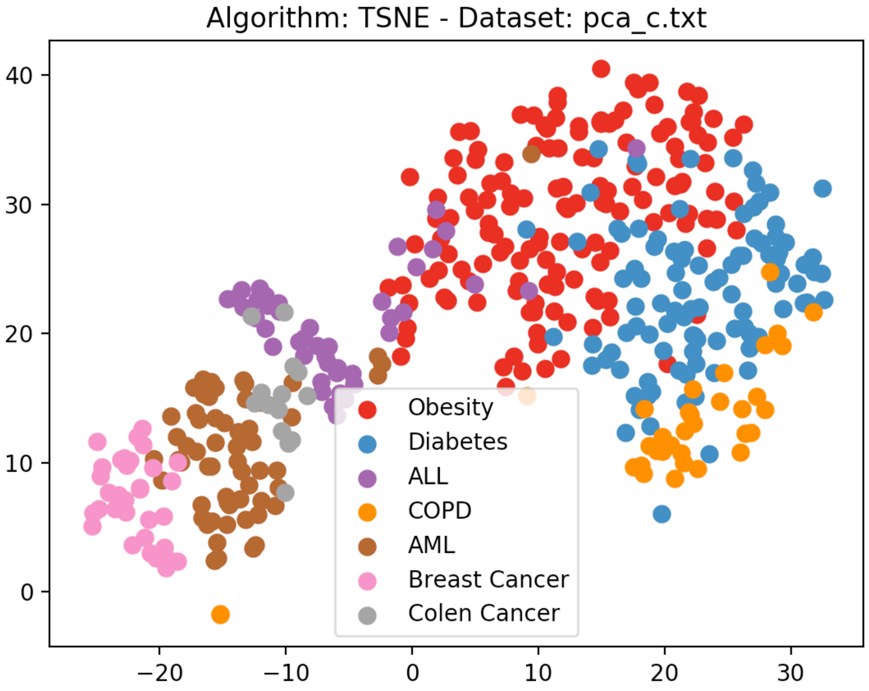

Algorithm: SVD - Dataset: pca_b.txt



Algorithm: TSNE - Dataset: pca_b.txt



Algorithm: SVD - Dataset: pca_c.txt



PCA Algorithm Flow

- We input all the features except the class label into a matrix '*matrix*' to begin with.
- We mean-center the feature matrix in '*mean*'. Thus, we have the *adjustedMatrix* as a difference between *matrix* and *mean.*
- After mean centering we generate the covariance matrix of the mean centered feature matrix and store it in *covarianceMatrix.*
- All eigenvalues and eigenvectors of the covariance matrix are generated. The top two dimensions with highest variance are selected based on a sorted vector of eigenvalues. Corresponding eigenvectors are also selected. They were stored in *eigenValueFS* and *eigenVectorsFS* respectively.
- The dot product of *adjustedMatrix* against each of the eigen vectors is taken and stored into a two-dimensional array *arrayMultiple* for plotting. This combines the related dimensions and lets us focus on the uncorrelated dimensions with maximum variance.
- This *arrayMultiple* holds the principle components of the feature matrix and is used to plot the graph.
- SVD is performed using truncatedSVD() of the sklearn.decomposition package by setting *n_components* parameter to 2. We then fit the SVD into the *adjustedMatrix*. Here n is 2 as we require only two dimensions.
- t-SNE is performed using TSNE() of the sklearn.manifold package by setting *n_components* parameter to 2. We then fit the SVD into the *adjustedMatrix*. Here n is 2 as we require only two dimensions.