CSE 601

Data Mining Project 1

Dimensionality Reduction

Anjana Guruprasad (50205233)

Apoorva Hejib (50206516)

Shreya Ravi Hegde (50208485)

Introduction:

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

The number of principal components is less than or equal to the number of original attributes or features.

This transformation is such that the first principal component has the largest possible variance (accounts for as much of the variability in the original data as possible) and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

Implementation:

We implement Principle Component Analysis (PCA) algorithm to perform dimensionality reduction on the given biomedical data files. We perform dimensionality reduction to better represent and analyze data. The basic concept is to combine dimensions that are highly correlated or dependent and focus on the independent ones. We end up with a smaller set of dimensions that retains the variance in original data.

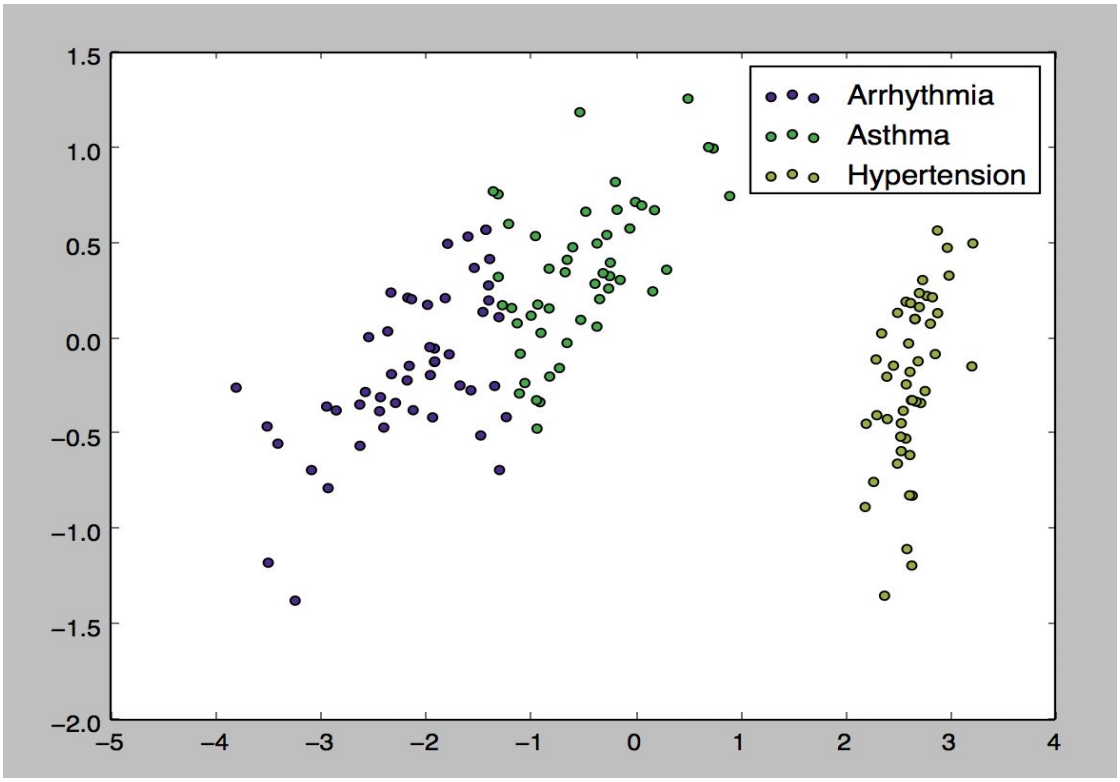
- Follow the steps of the PCA Algorithm on the data files given:
  - Convert the original data set to a matrix with all the attributes and along each row center the data along its respective mean.
  - Compute the covariance matrix of this mean centered matrix and obtain its eigenvalues and eigenvectors.
  - Since we are reducing the dimensions to 2, we require 2 Principal components.
  - Obtain the two highest eigenvalues and their corresponding eigenvectors since we are reducing the dimensions to 2.
  - Plot both these vectors on each axis.

- Project the high-dimensional data to 2 dimensions and plot the 2-dimensional data points with a scatter plot.

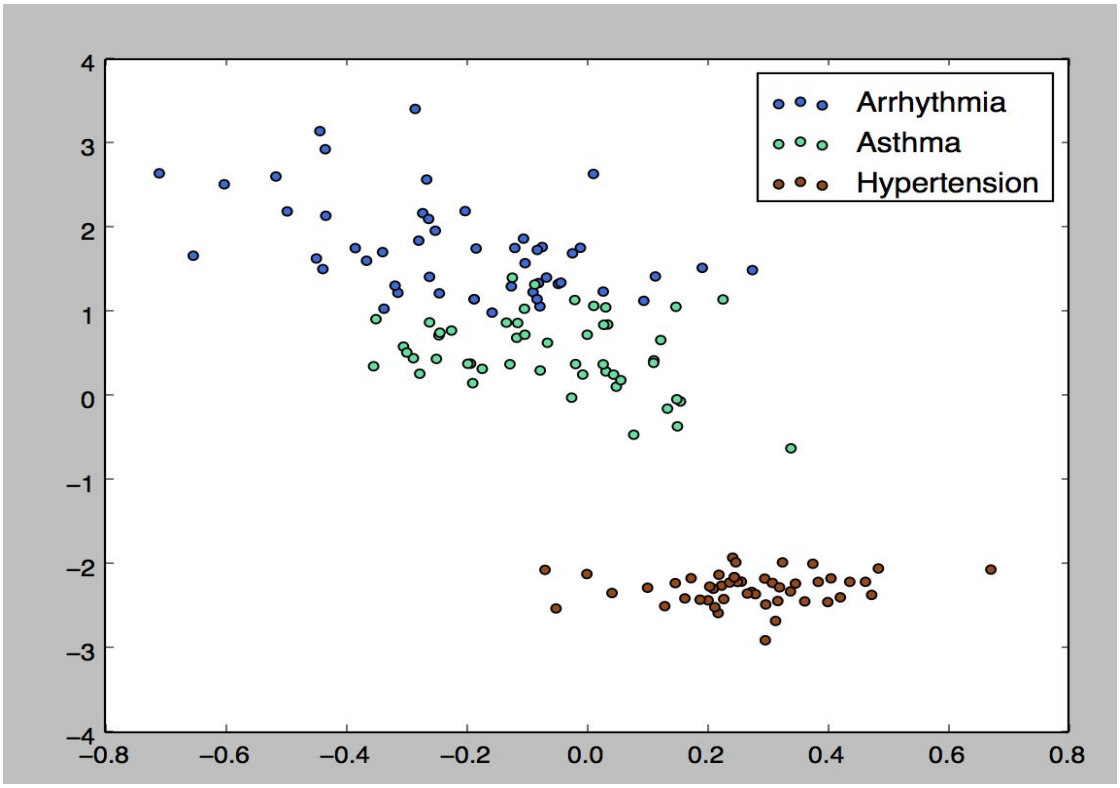
- Apply existing packages of SVD and t-SNE algorithms to obtain 2-dimensional data points. Similarly plot these values with the help of scatter plots.

Dataset Results (Dimensionality Reduction):

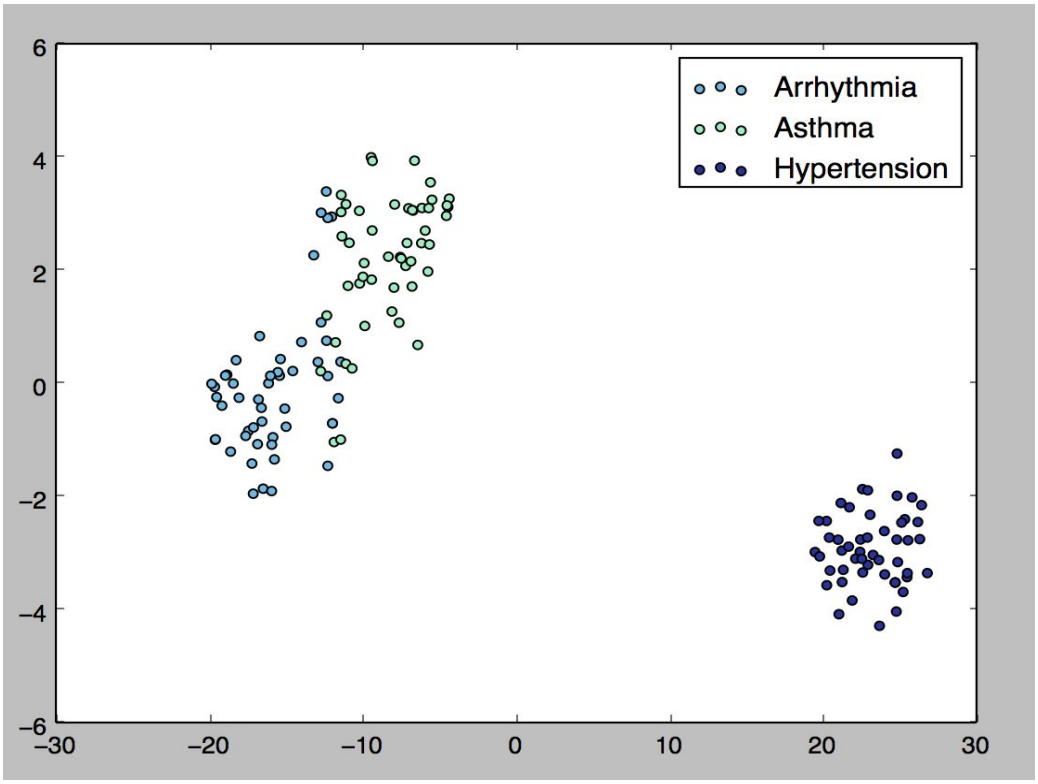
Dataset : **“pca_a”**



PCA_A : PCA

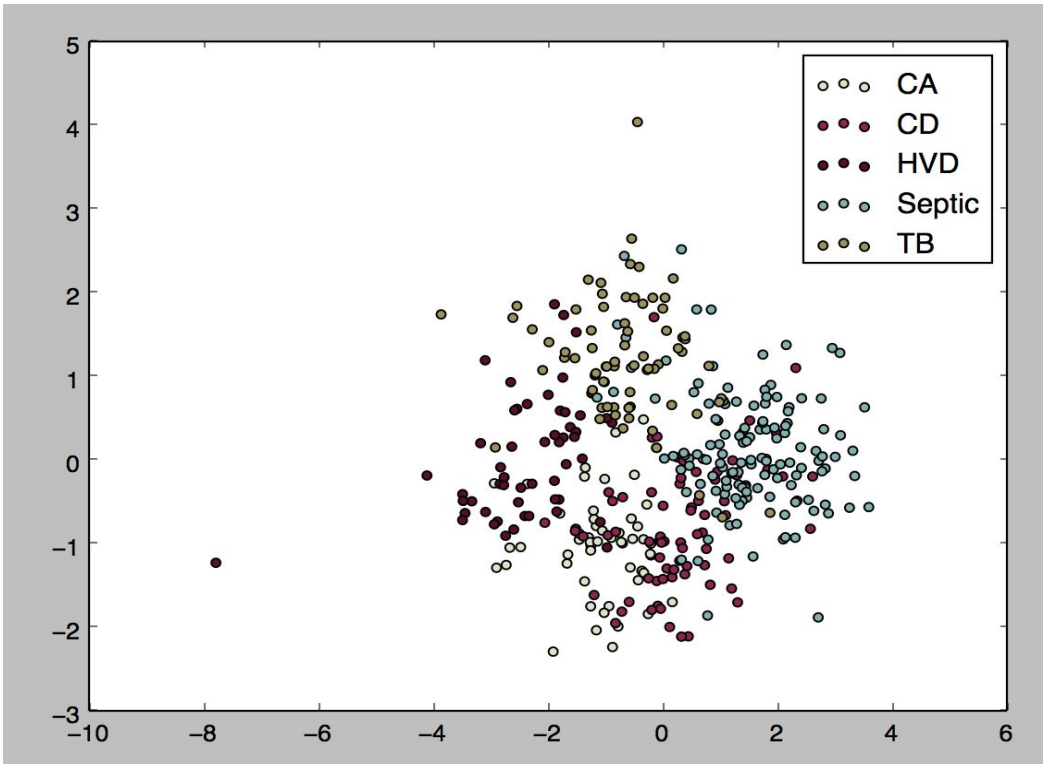


PCA_A : SVD

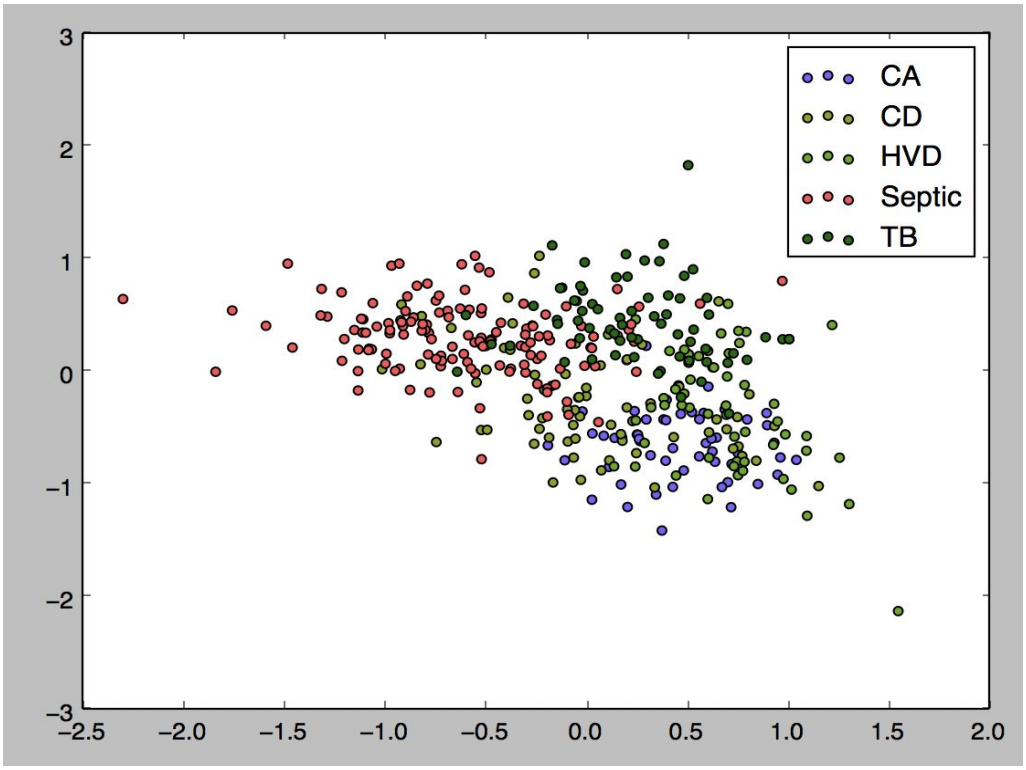


PCA_A : T-SNE

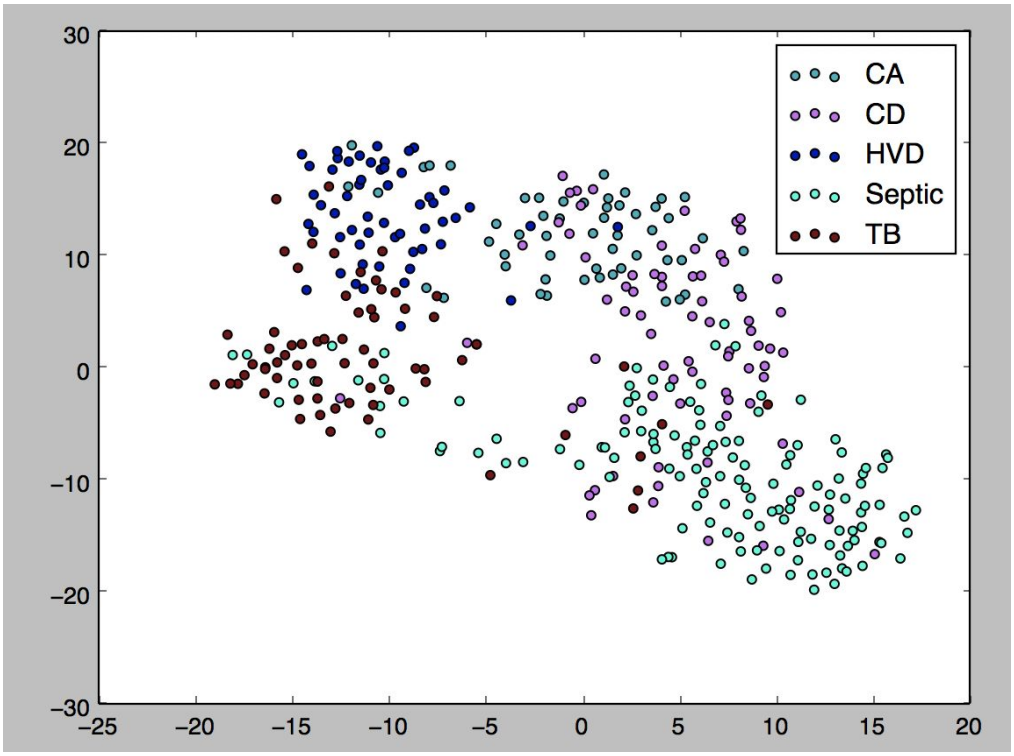
Dataset : **“pca_b”**



PCA_B : PCA

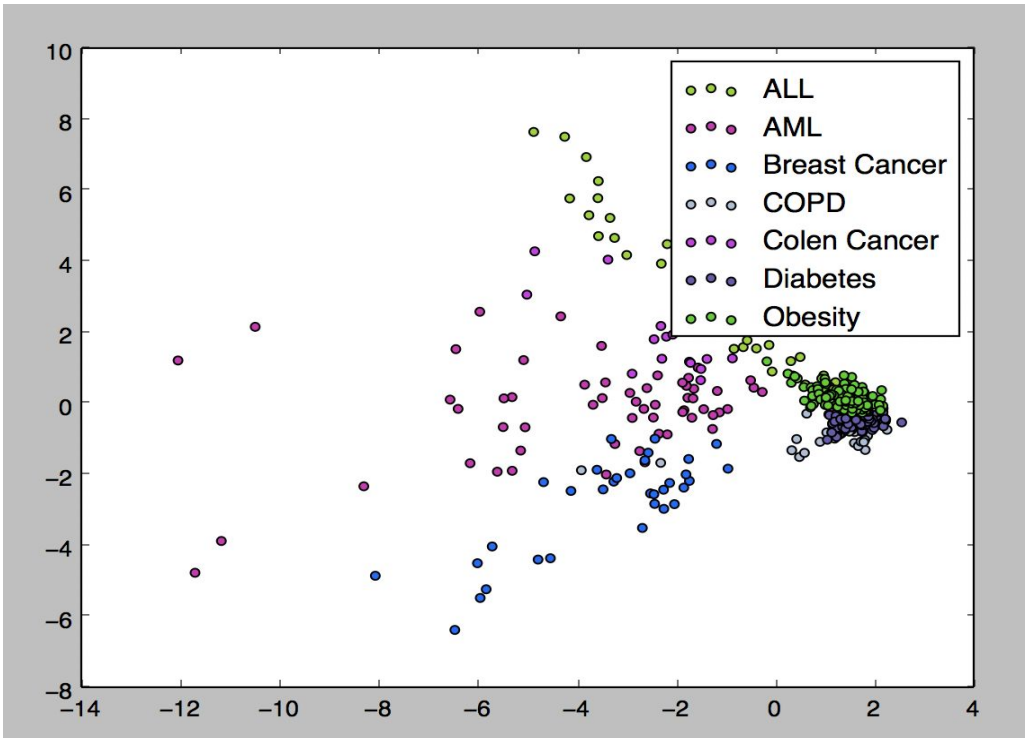


PCA_B : SVD

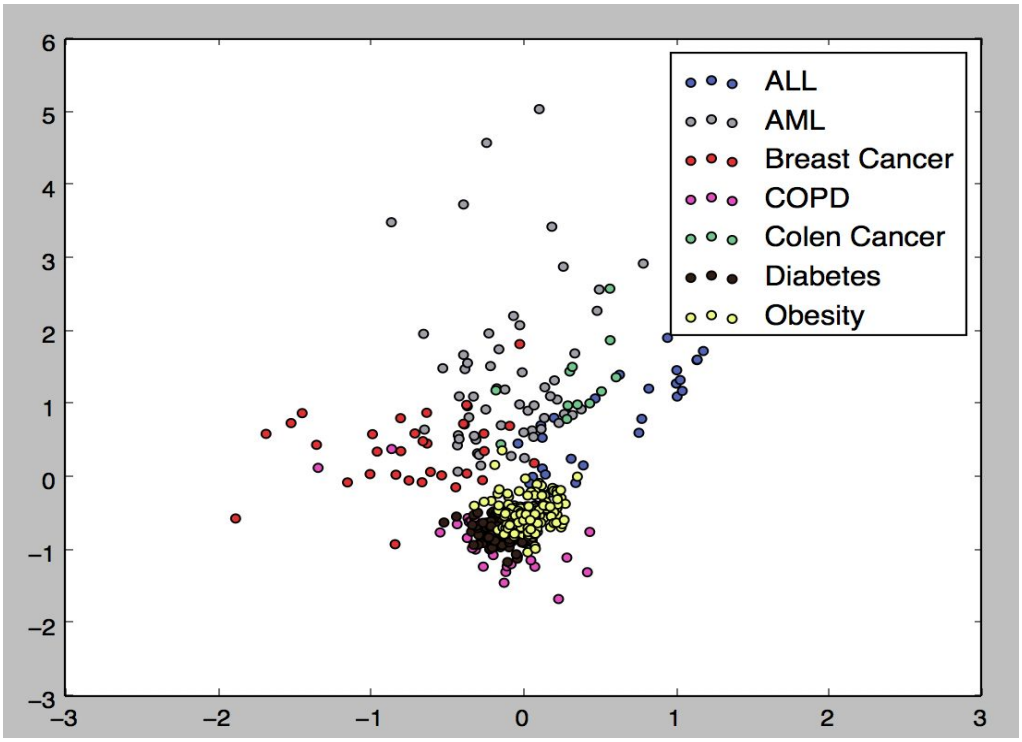


PCA_B : T-SNE

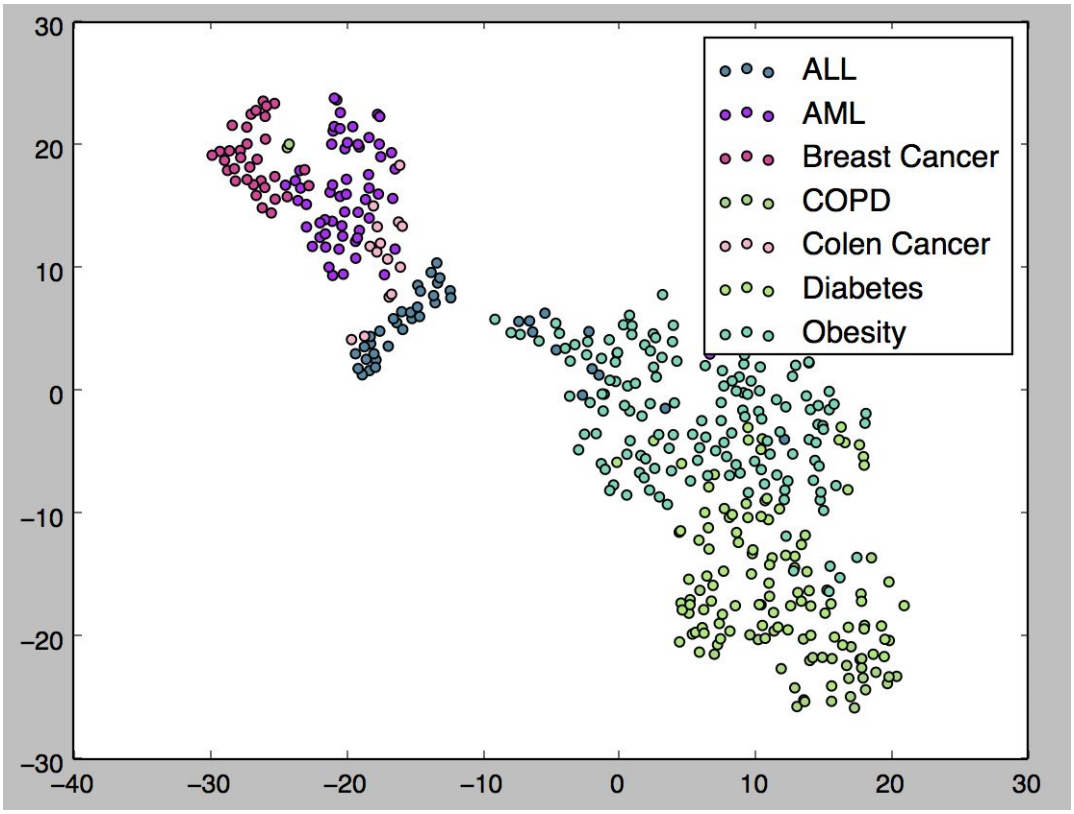
Dataset : **"pca_c"**



PCA_C : PCA



PCA_C : SVD



PCA_C : T-SNE

Conclusions:

For pca_a.txt:

- PCA - We can see three almost distinct clusters, one for each disease present in the file.
- SVD - SVD also distinctly clusters the given data into three different clusters for each disease very similar to PCA.
- t-SNE - t-SNE also clusters the data similarly.
- One common observation is that Arrhythmia and Asthma seem correlated.
- With these methods we can easily distinguish the diseases and it helps us classify a particular disease.
- We also observe that hypertension would be easy to classify as none of these data points overlap with the other two diseases.

For pca_b.txt:

- In PCA and SVD we observe well defined, distinct clusters for each disease but the clusters tend to overlap with each other.
- These data points could be correlated as there is a huge overlap.
- In the t-SNE plot, the clusters are not as well defined and the data points are more spread out.

For file pca_c.txt

- The PCA and SVD plots are very similar. We observe that the clusters formed for the diseases COPD, Diabetes and Obesity are well defined and close to each other such that there is a huge overlap. These diseases could be highly correlated.
- We see clusters for each disease formed and distinguishable from each other with only a few data points overlapping.

References:

-> Lecture notes

-> https://en.wikipedia.org/wiki/Principal_component_analysis