DATA MINING & BIOINFORMATICS

Dimensionality Reduction(PCA)

Himal Dwarakanath	himaldwa	50207594
MuthuPalaniappan Karuppayya	muthupal	50208484
Neeharika Nelaturu	nnelatur	50207062

Principal Component Analysis

Principle

- A set of correlated dimensions is transferred to a set of uncorrelated dimensions.
- Data is mapped to a space of lower dimensionality.

Implementation of PCA Algorithm

Our implementation of the PCA algorithm is done basically by following below steps,

- Existing axes in n-dimensional space is reduced to 2-dimensional space and can be viewed as a rotation of the existing axes to new positions in the space defined by original dimensions.
- The data is plotted in the new 2-dimensional space in which the axes are orthogonal and represent directions with maximum variability.

PCA Steps

- Step 1 – Standardize:
  The scale of data is standardized. The original data is centered around the mean by adjusting original data by the mean.

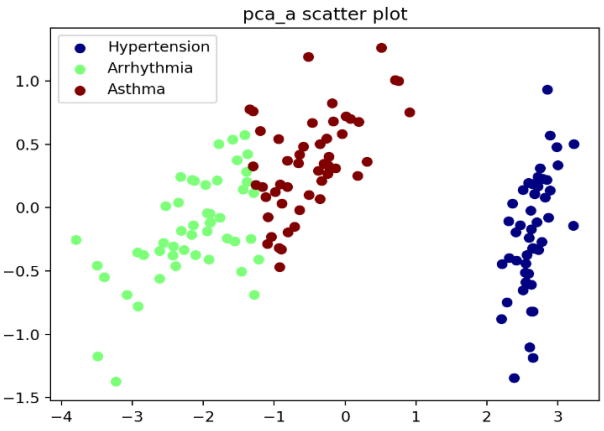
  $X = X - \bar{X}$

- Step 2 – Calculate covariance:
  $$S = \frac{1}{n} XX^T$$

- Step 3 – Deduce eigens:
  Find the eigenvectors and eigenvalues of S.
  $$Sa = \lambda a$$
  Select the top 2 eigen values and their corresponding eigen vectors which form the principal components.

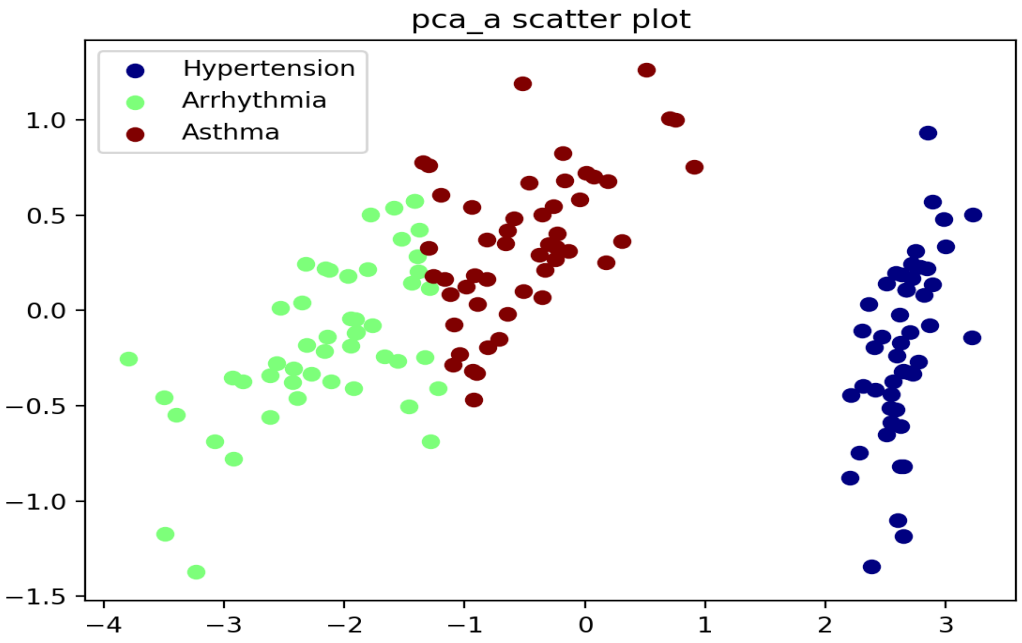
- Step 4 – Re-orient data:
  Data is re-oriented from a n-dimensional space to a 2-dimensional space.

- Step 5 – Plot re-oriented data:

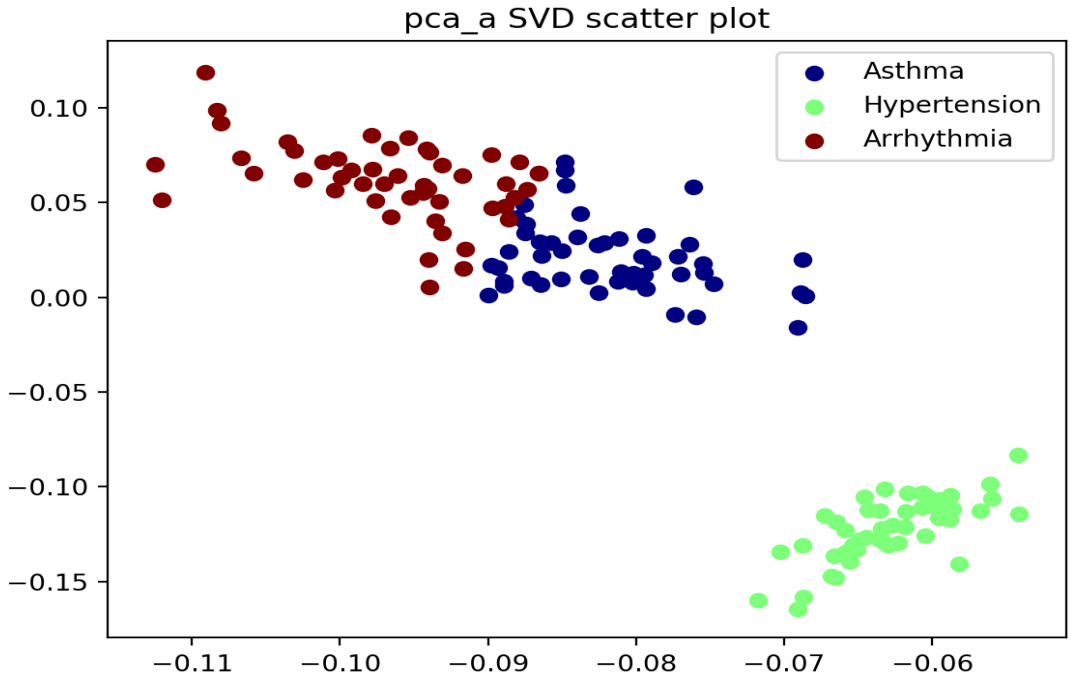


PCA Algorithm Flow

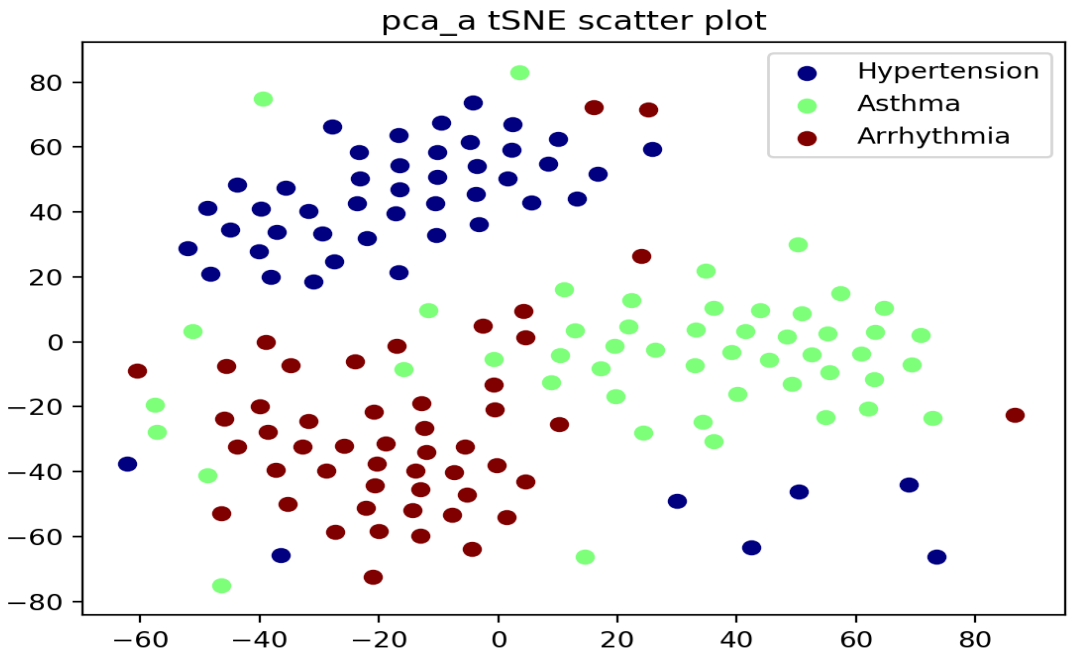
- The file was read into an input matrix (input_matrix).
- An adjusted matrix (adjusted_matrix) is calculated by subtracting column mean from value in each tuple in the input matrix.
- Co-variance matrix (cov_matrix) is obtained by taking the product of adjusted matrix with its transpose and dividing it by total number of records.
- A list of Eigen values(eig_val) and list of eigen vectors(eig_vec) are obtained from the co-variance matrix
- Eigen vectors corresponding to top 2 eigen values are selected which form the principal components.
- Re-orient data into 2-D space using principal components.
- Each unique disease is assigned a unique integer value in a list (disease_list_encoded) obtained from dictionary(d).
- Further each unique integer value corresponds to a single color in the plot.
- Diseases are plotted based on color encoding.



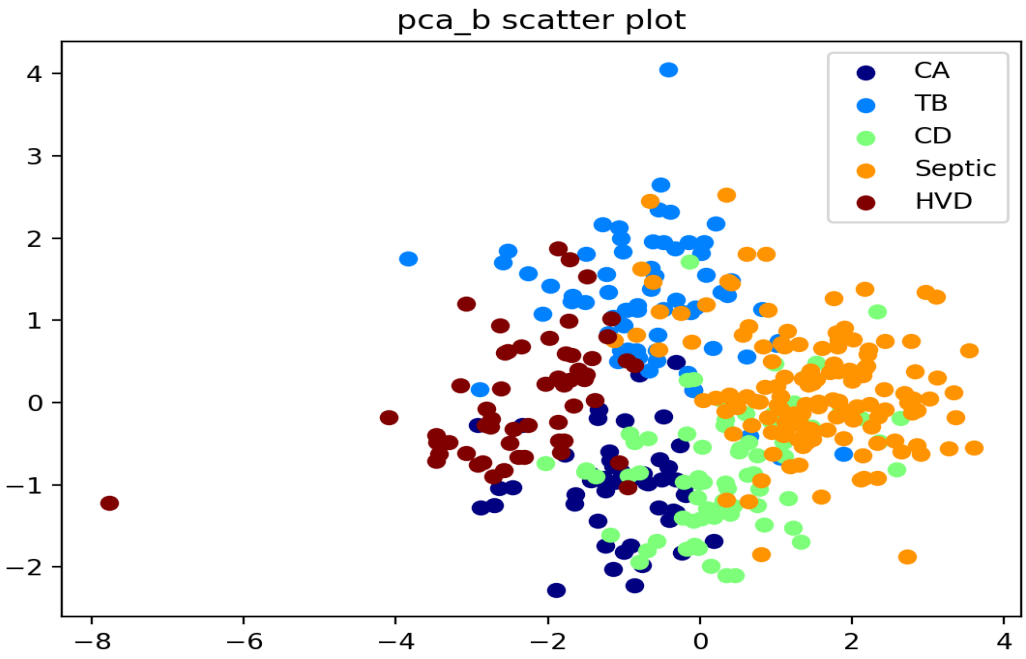
Filename: pca_a ; Algorithm: PCA



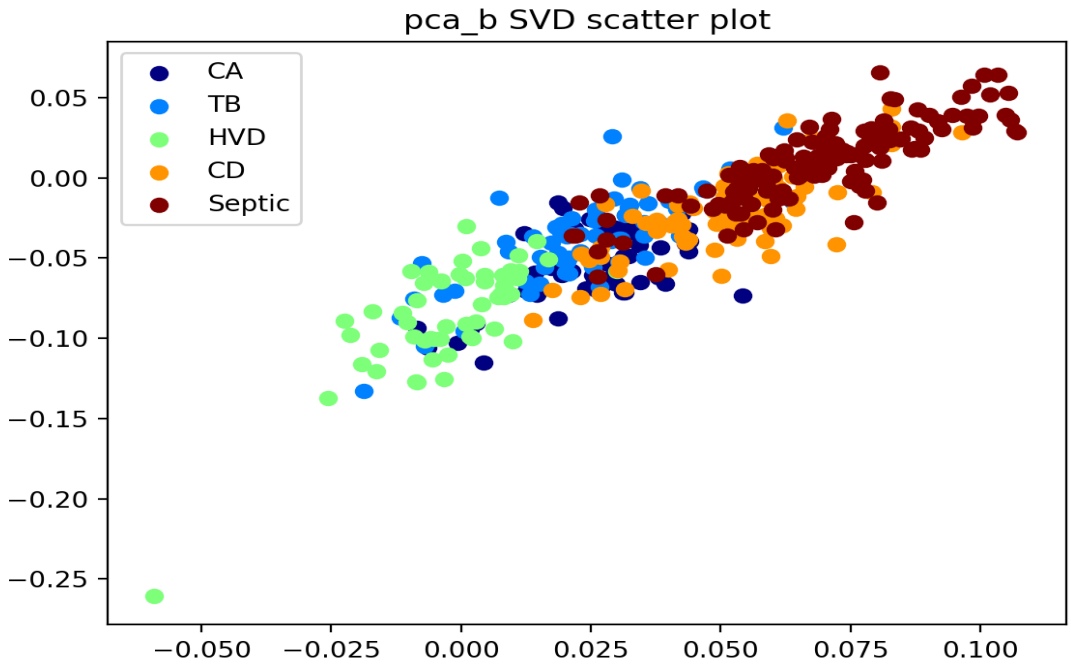
Filename: pca_a ; Algorithm: SVD (with input matrix)



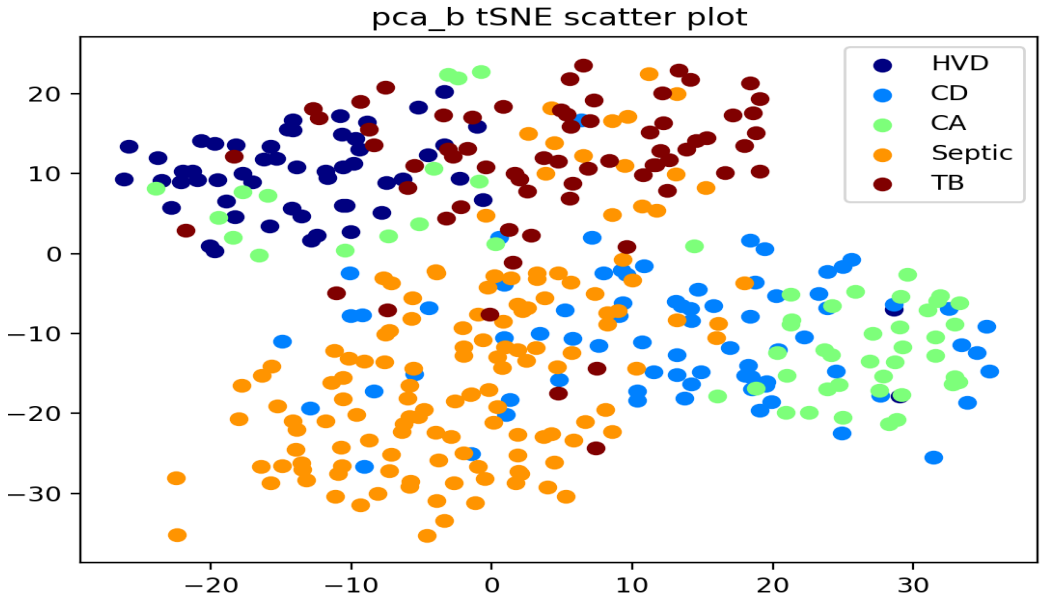
Filename: pca_a ; Algorithm: t-SNE



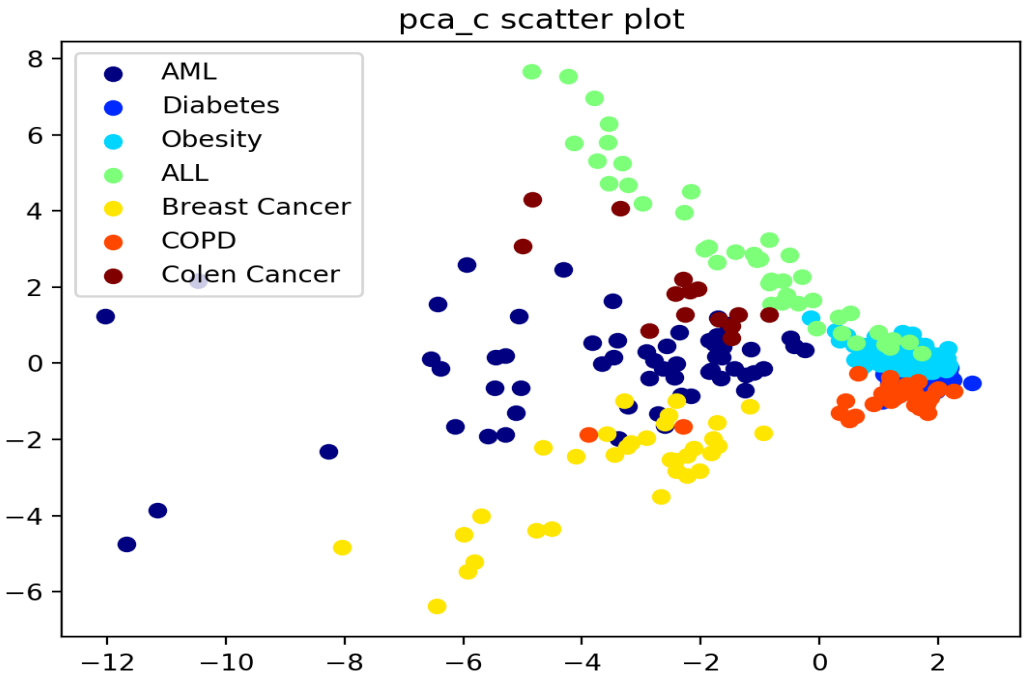
Filename: pca_b ; Algorithm: PCA



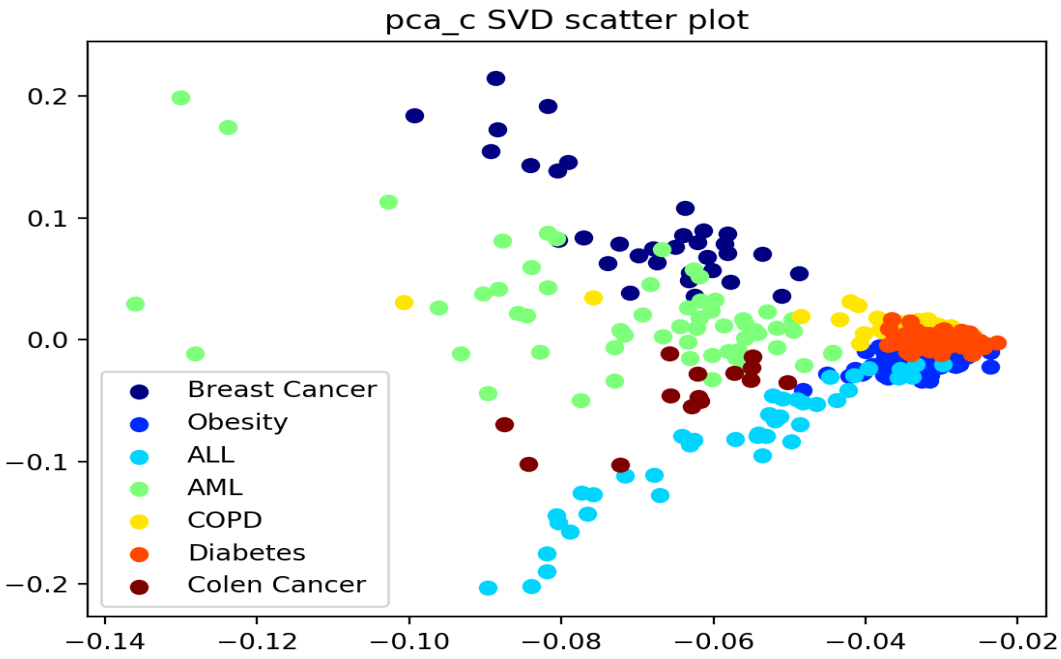
Filename: pca_b ; Algorithm: SVD (with input matrix)



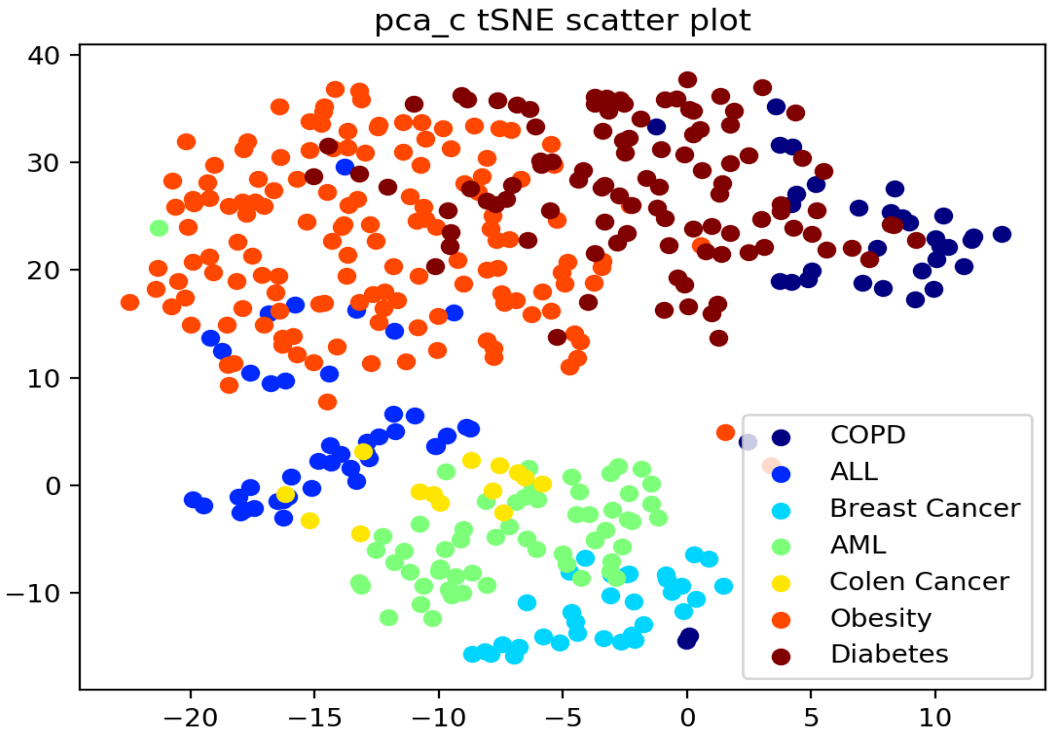
Filename: pca_b ; Algorithm: t-SNE



Filename: pca_c ; Algorithm: PCA



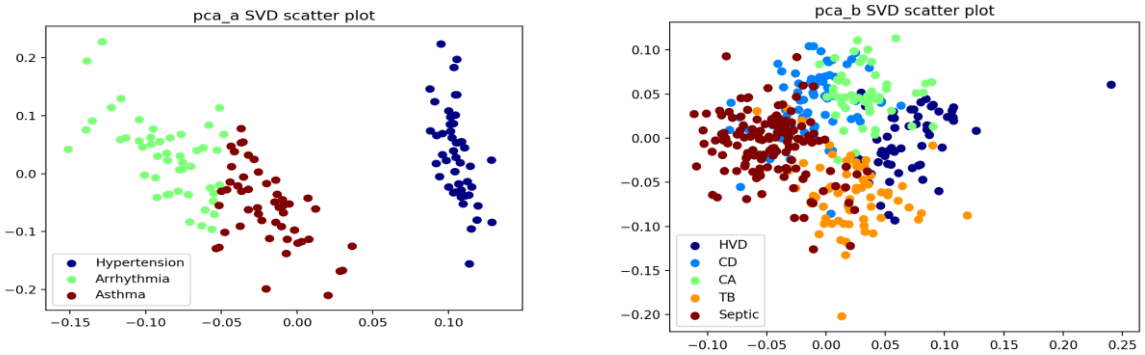
Filename: pca_c ; Algorithm: SVD (with input matrix)



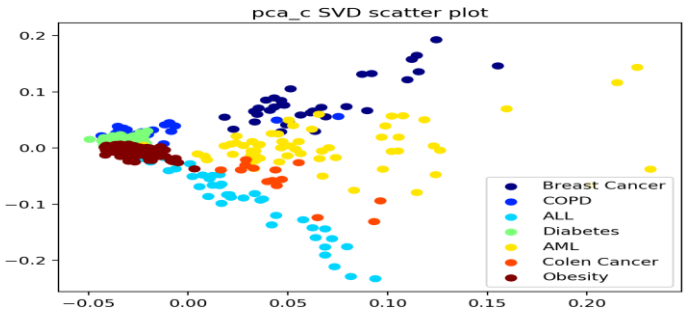
Filename: pca_c ; Algorithm: t-SNE

Inference:

- When input matrix is used for SVD, we observe that the plots differ from PCA. However, we observe that when mean-centered matrix is used, the PCA and SVD plots are mirror images of each other (similarity in direction is possibly a consequence of same eigen values and eigen vectors).



Filenames: pca_a, pca_b ; Algorithm: SVD (with mean-centered matrix)



Filename: pca_c ; Algorithm: SVD (with mean-centered matrix)

- In t-SNE, the plots are different from PCA and SVD. This can be explained by the fact that t-SNE uses probabilistic approach for reducing the components. Thus, we observe t-SNE plots change for every execution of the program.
- In t-SNE, dimensionality reduction is done using gradient descent whereas PCA uses eigen decomposition for dimensionality reduction.
- PCA uses the orthogonal transform to form uncorrelated variables and selects them in decreasing order of variance.