CSE 601 Project 1:

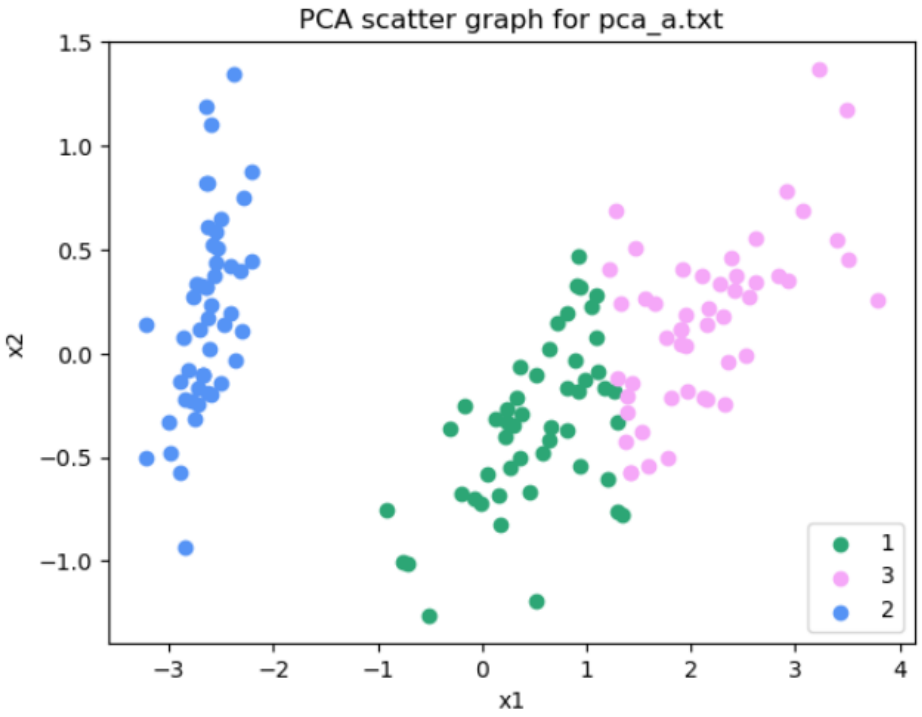
Dimensionality Reduction

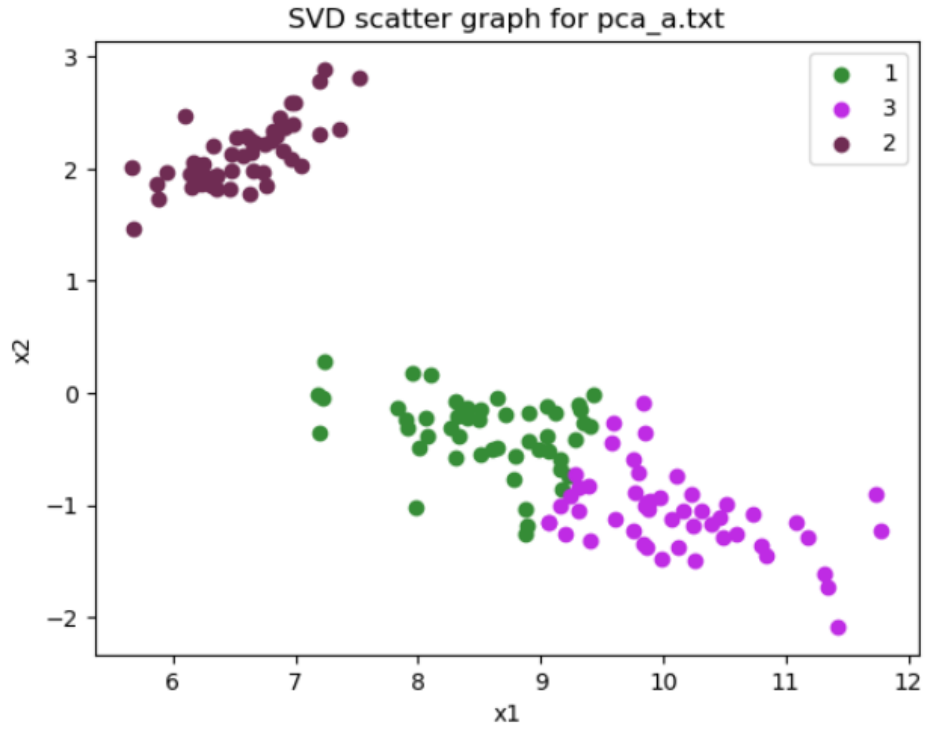
Team Member: Ninghui Jin, Bharath Chandra Reddy Ravula

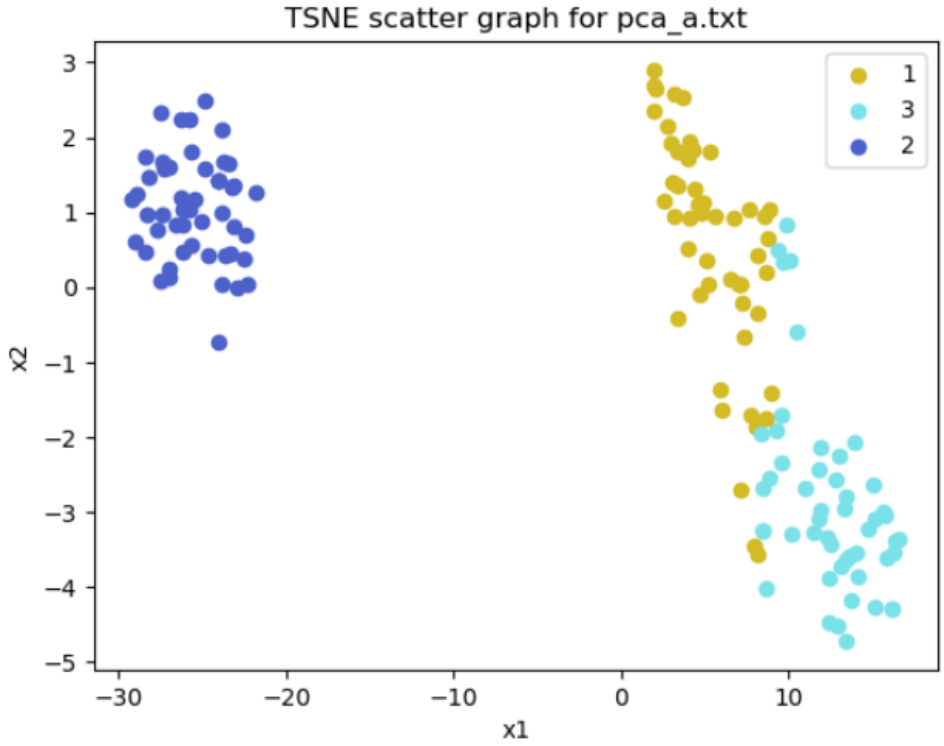
Instructor: Dr. A. Erdem Sariyuce

1. Nine scatter plots from three datasets and three algorithms. Label them properly by the dataset name and algorithm name in each plot.

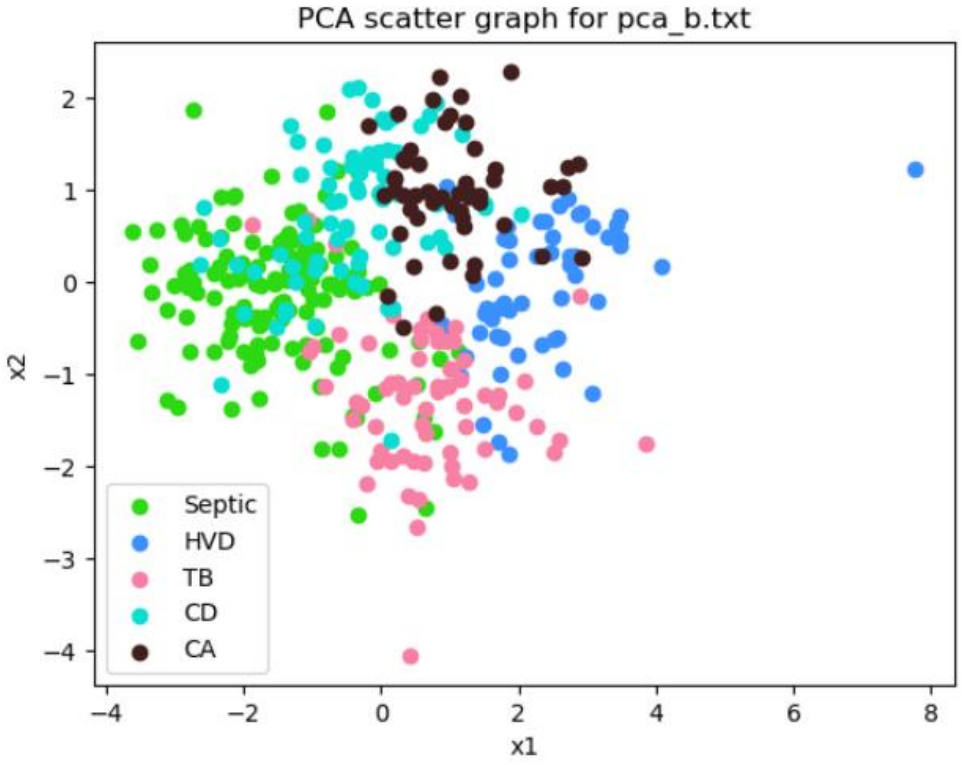
    - pca_a.txt

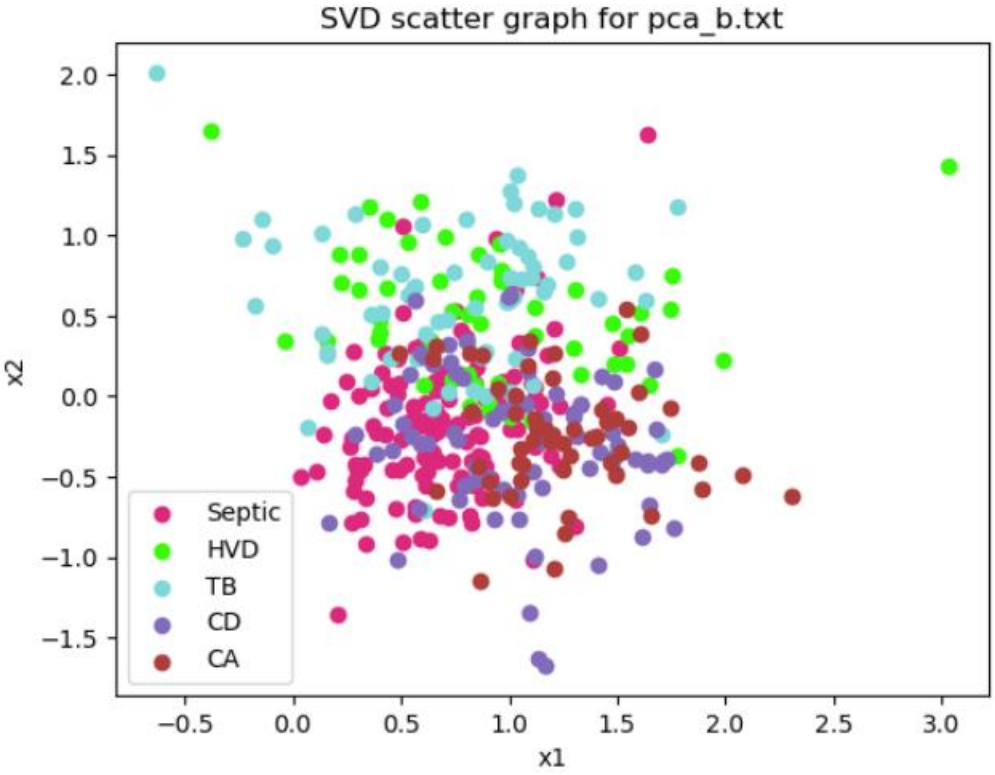


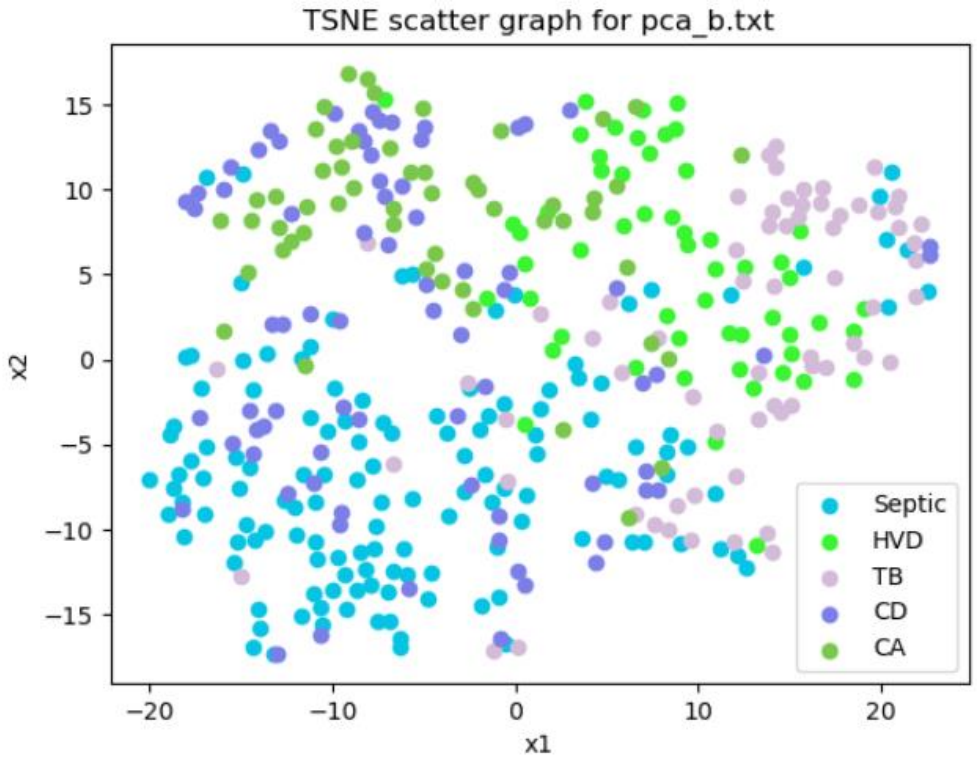

SVD scatter graph for pca_a.txt
1
3
2
x1
x2
3
2
1
0
−1
−2
6
7
8
9
10
11
12


TSNE scatter graph for pca_a.txt
1
3
2
x1
x2
3
2
1
0
−1
−2
−3
−4
−5
−30
−20
−10
0
10

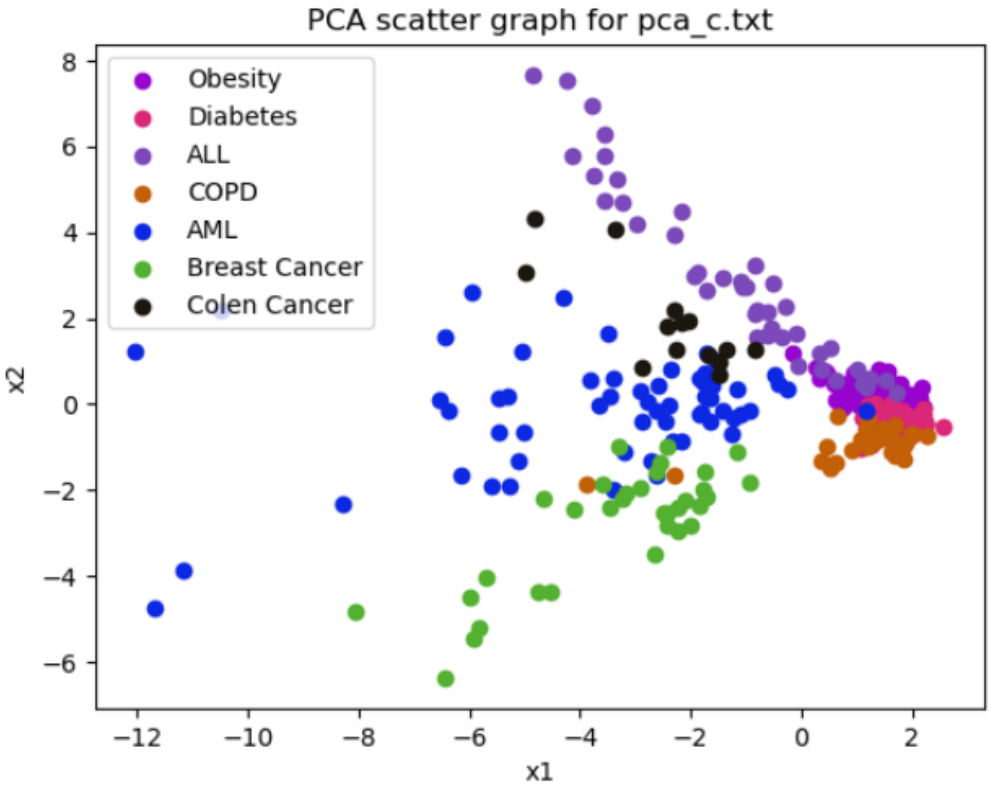
- pca_b.txt

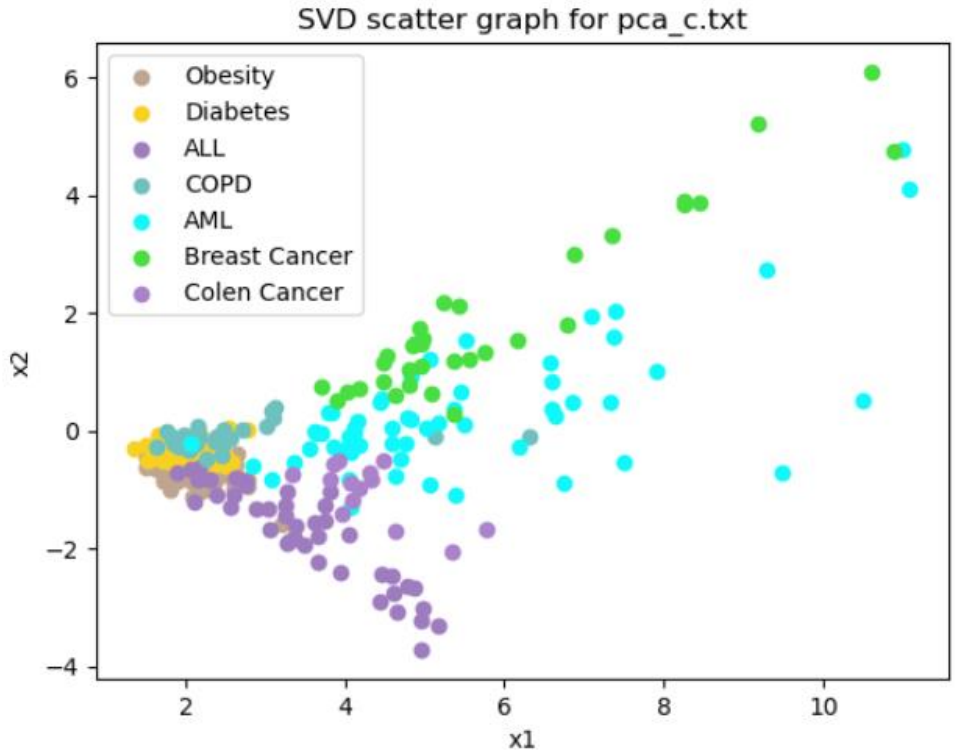

PCA scatter graph for pca_b.txt
x2
x1
Septic
HVD
TB
CD
CA

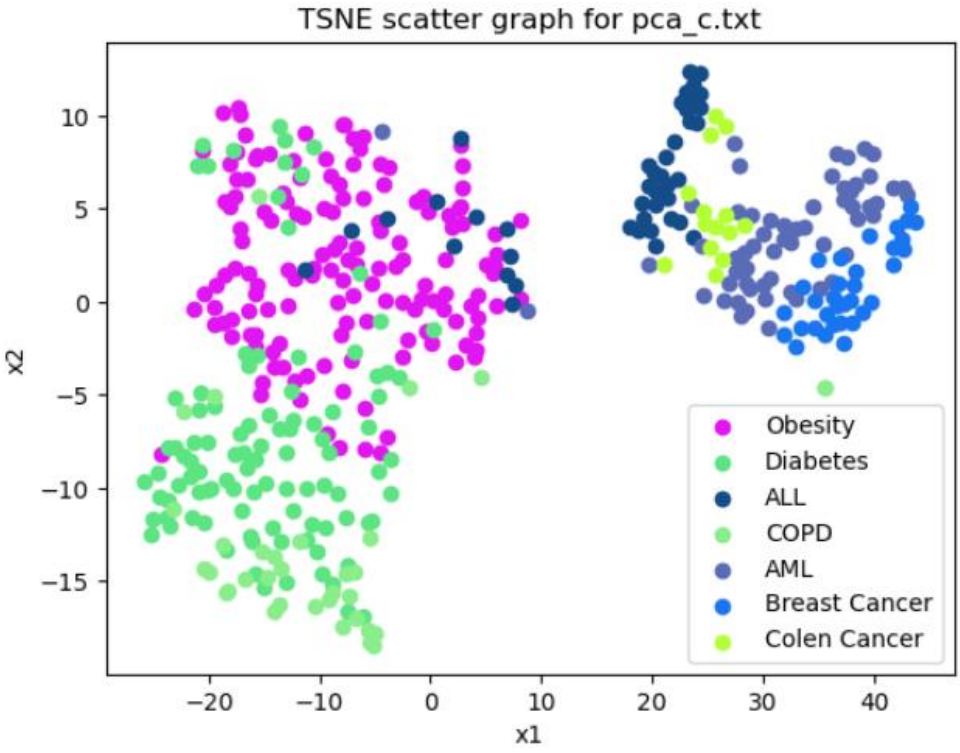

SVD scatter graph for pca_b.txt
x2
x1
Septic
HVD
TB
CD
CA


TSNE scatter graph for pca_b.txt
x2
x1
15
10
5
0
−5
−10
−15
−20
−10
0
10
20
Septic
HVD
TB
CD
CA

- pca_c.txt


PCA scatter graph for pca_c.txt
x2
x1
8
6
4
2
0
−2
−4
−6
−12
−10
−8
−6
−4
−2
0
2
Obesity
Diabetes
ALL
COPD
AML
Breast Cancer
Colen Cancer





2. Brief descriptions for the flow of your PCA implementation and discussion of the results obtained by different algorithms.

- Flow of PCA implementation.
    a. Find the mean value for each column in the dataset.
    b. Subtract the mean value of each column from the corresponding column, creating a mean-adjusted matrix.
    c. Compute the covariance matrix for the mean-adjusted matrix.
    d. Find the eigenvalues and eigenvectors of the covariance matrix.
    e. Choose the top 2 eigenvalues and their corresponding eigenvectors.
    f. The selected eigenvectors are the principal components.
    g. Project the mean-adjusted data onto the principal components by taking the dot product of the transposed principal components and the transposed mean-adjusted data.
- Discussion of the results obtained by different algorithms.

The result obtained from PCA aligns with the direction of maximum variance, showcasing the projection onto the principal component. In contrast, the scatter plot created by SVD exhibits a wider spread. The shapes of the scatter points between them are somewhat similar, unlike t-SNE. Points generated by t-SNE tend to have larger values compared to other algorithms. Consequently, it becomes easier to discern the boundaries of different classes, as it preserves more of the relationships within the dataset.