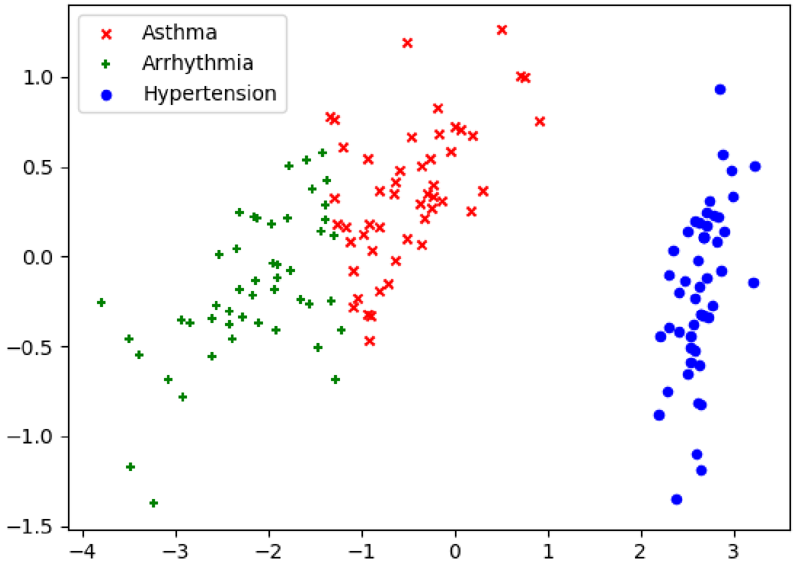
PCA ALGORITHM

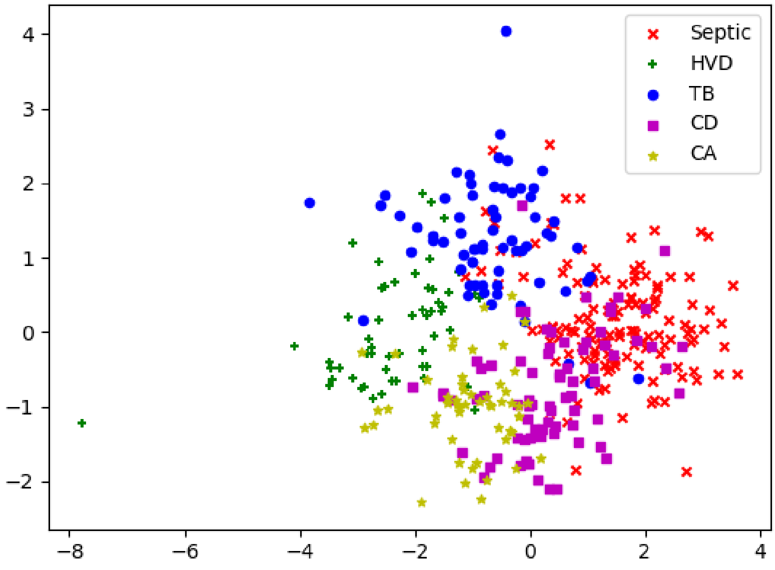
Group member: Fanzi Xiao Shaoming Xu Haowei Zhou

PLOTS on datasets

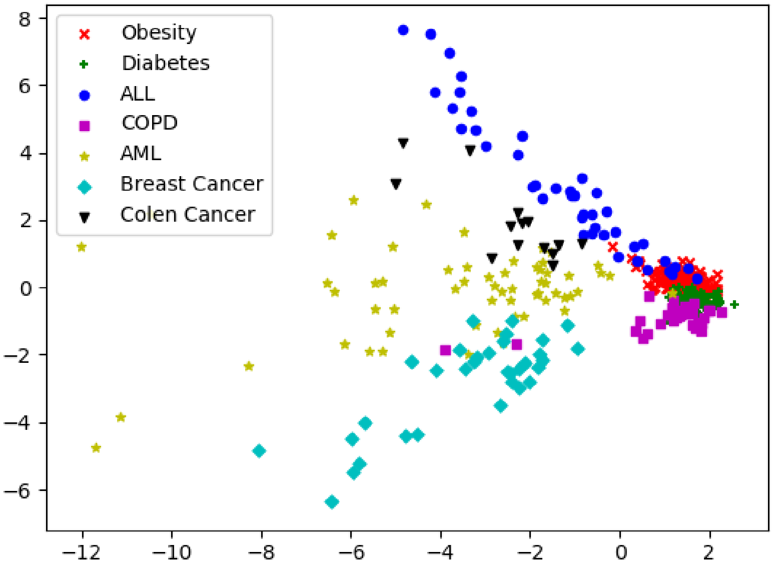
1. scatter plots on PCA



1.1 PCA_a

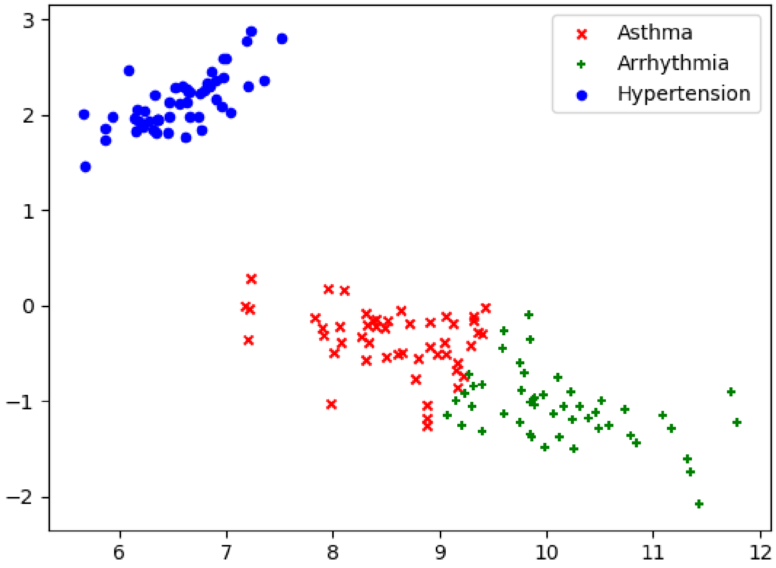


1.2 PCA_b

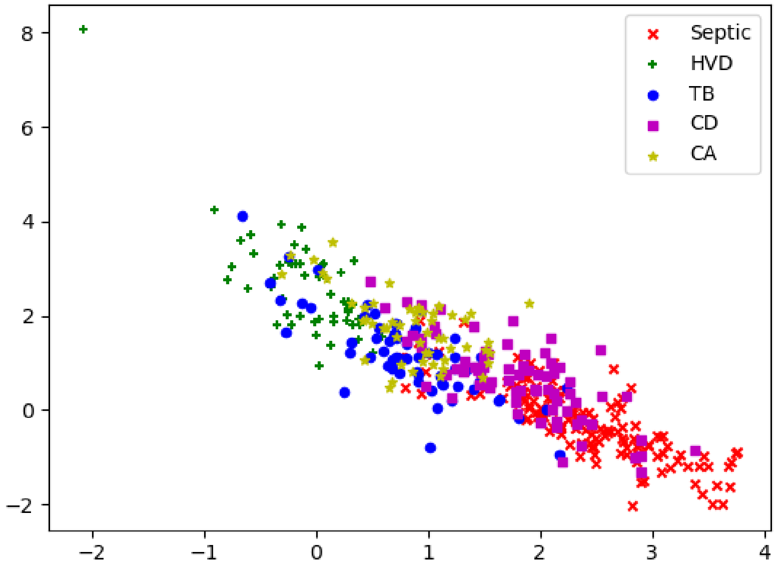


1.3 PCA_c

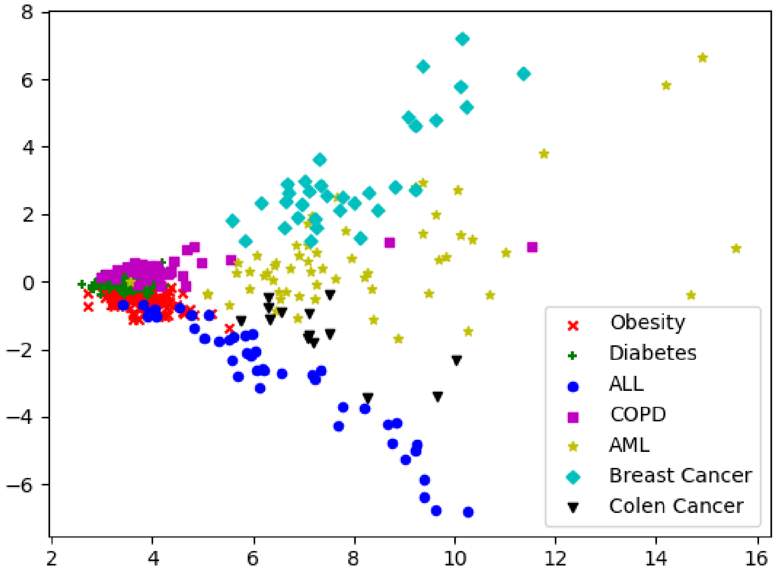
2 scatter plots on SVD



2.1 SVD_a

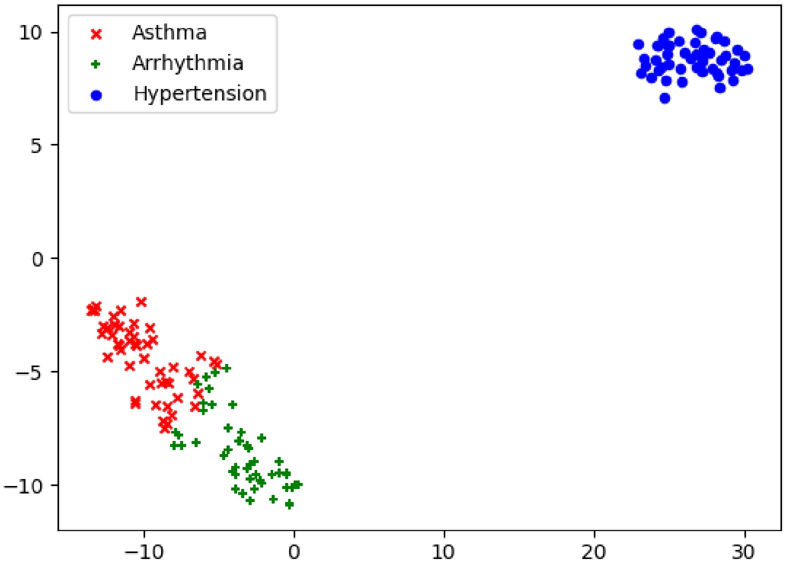


2.2 SVD_b

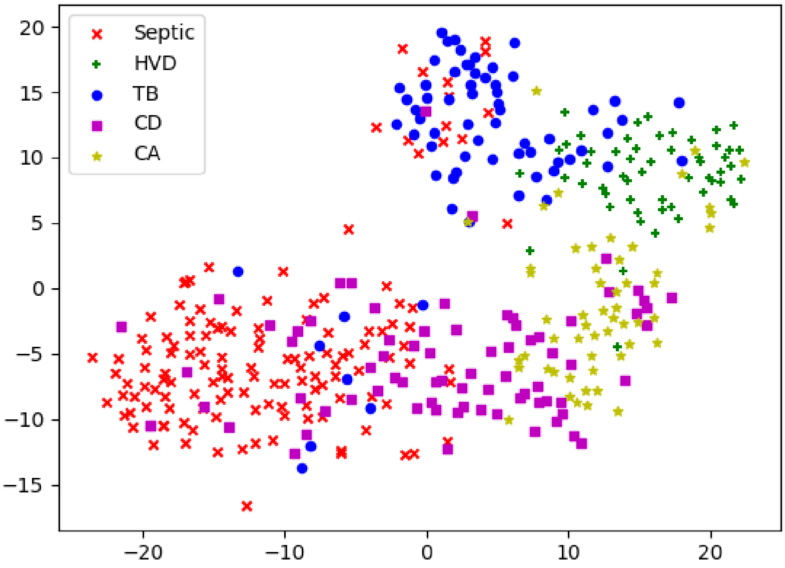


2.3 SVD_c

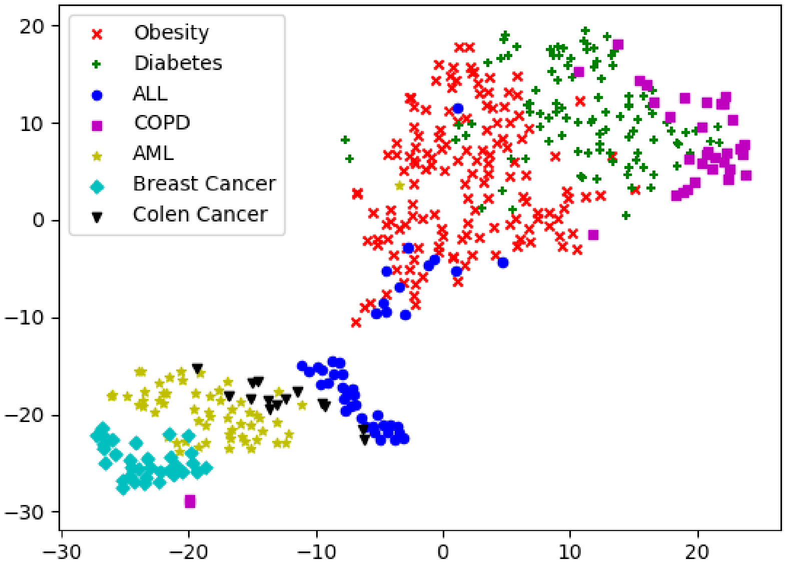
3 scatter plots on T-SNE



3.1 T-SNE_a



3.2 T-SNE_b



3.3 T-SNE_c

PCA implementation detail

We implement the PCA algorithm by these steps.

(1) Load the dataset and separate it into get a data matrix and a label matrix.

(2) Calculate the mean of each column of the data matrix, then adjust the original data matrix by the mean vector. Here we use the broadcasting idiom of Numpy to simplify the code.

(3) Compute the covariance matrix of the adjusted matrix, then calculate eigenvectors and eigenvalues decomposition on the covariance matrix.

(4) Using the indexes of first two largest eigenvalues to get the corresponding eigenvectors, pack them to matrix and use it to do the dot product on original matrix.

Followed these steps, finally we can get a two-dimension matrix.

Results discussion

From the scatter plots, we find all three algorithms PCA, SVD, T-SNE separate the pca_a data pretty well. But for the pca_b dataset, we can see the boundary among classes is not so clear among all three classes. And for the pca_c dataset, the pattern of PCA and SVD on dimension reduction seems a bit similar, both of them gather the diabetes and obesity class tightly. But for T-SNE, although the distance of points between diabetes and obesity class is close, but compared to PCA and SVD, they can be separated more easily.