**Implementation of Dimensionality reduction in Healthcare data**

**Abstract**

Dimensionality is an ever-existing issue in digital health data. Due to the complex nature such as the presence of multiple features upon which one dependent variable exists. Especially in a dataset where the patient’s disease is estimated based on other attributes of the patient, it’s quite difficult to figure out the disease of the patient without taking into account all those features. For the purpose for better visualization without losing much information in our multidimensional data, we have implemented three dimensionality reduction techniques such as Principle Component Analysis(PCA), Single value Decomposition(SVD) and T-Distributed Stochastic Neighbourhood Embedding(t-SNE) algorithm which has significantly reduced processing time and mitigated the curse of dimensionality by taking the features in the form of row-based in various health datasets.

**Introduction**

The performance of any data science model is directly dependent on the size of the dataset. Higher the dimension, lower the performance of the model, also known as the “curse of dimensionality”. The theory behind this concept is that, when the features are increased without increasing the training data, the model is overfitted hence the performance is affected. Dimensionality reduction techniques perform feature selection and feature extraction.

The objective of feature selection is to choose a portion of the original features to get rid of redundant and outdated characteristics without losing much information in the model. Feature extraction generates new features by projecting the data from high dimensional space to a low dimension space.

**Implemented methods**

**Principle Component Analysis**(PCA) is a method for linear dimensionality reduction that can be used to condense a large number of variables into a smaller set that retains the majority of the original data. In order to extract the maximum variance, it looks for a linear combination of variables. The PCA algorithm then eliminates this variance and looks for a second linear combination that, iteratively, explains the greatest percentage of the variance still present. The principal axis approach, which generates orthogonal (uncorrelated) factors, is what it is termed. Additionally, it requires computing the eigenvalues and eigenvectors of covariance matrices, sorting those eigenvectors in descending order of their eigenvalues. We project the original data into the directions of the orthogonal eigenvectors.

Steps to apply PCA technique for a dataset (with features and a class label associated with it):

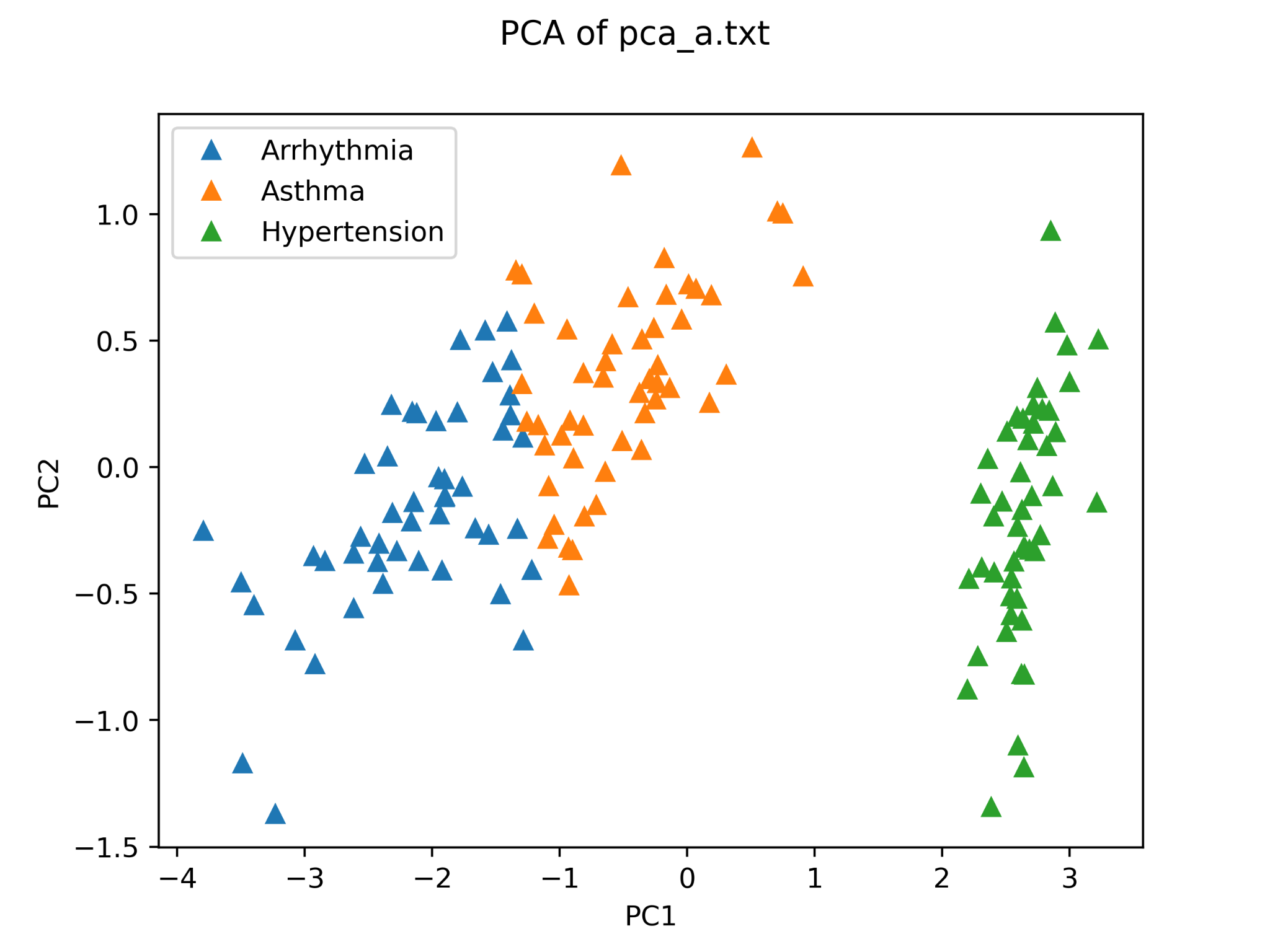
1. Split the features set and the class label of the dataset into separate matrices
2. Convert the features matrix into the mean centered data values.
3. Calculate the covariance matrix from the mean centered data of the features matrix
4. Compute the Eigen values and Eigen vectors from the covariance matrix
5. Sort the eigenvalues in descending order and likewise their eigenvectors.
6. Plot the points along the top two eigenvectors axes

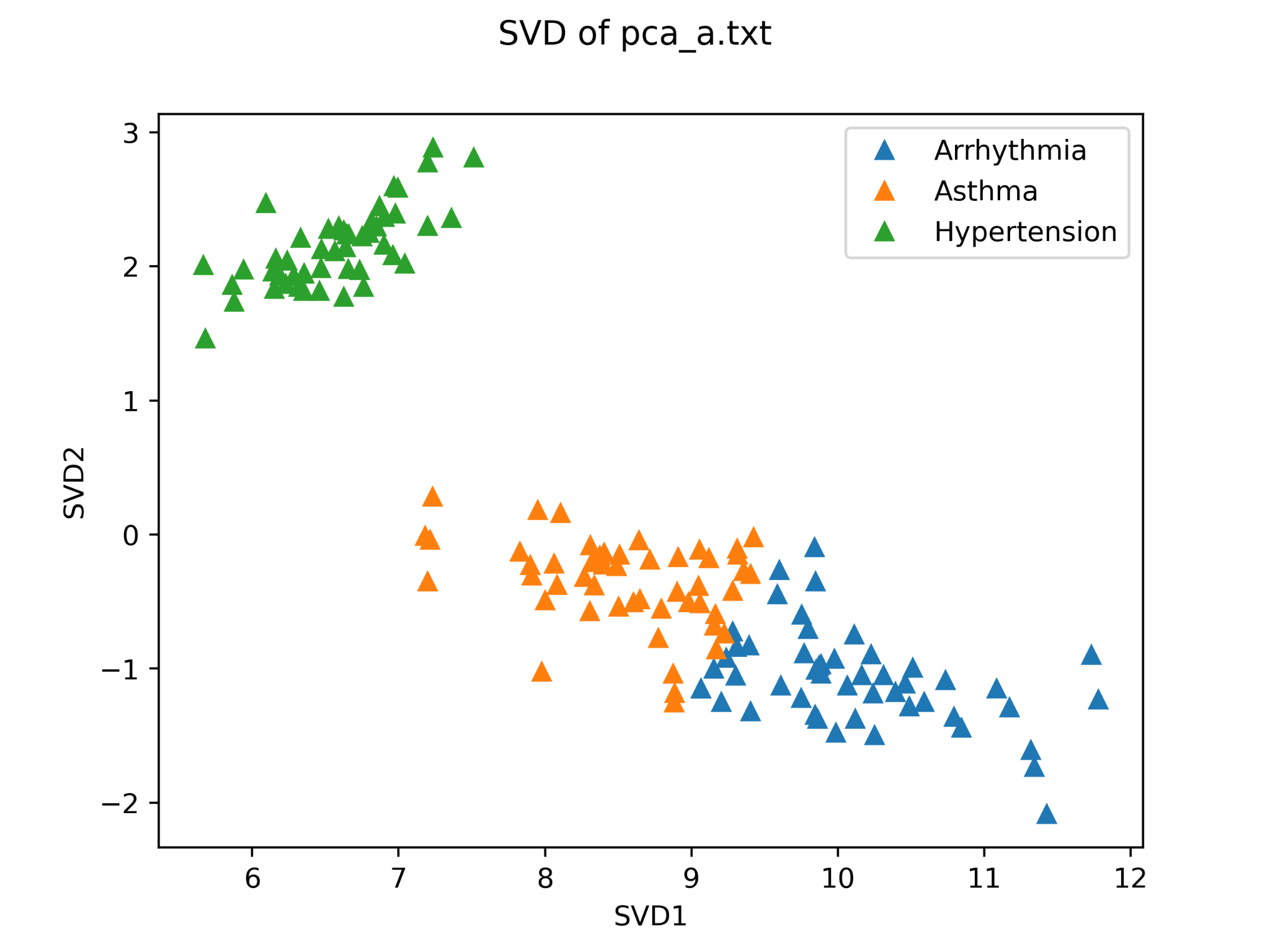
**Single Value Decomposition**(SVD) works on the principle of Rank of matrix. The maximum number of linearly independent row or column vectors in a matrix is defined as its rank. By performing decomposition, we can express our original matrix as a linear combination of low-rank matrices. Truncated SVD is more effective when used with sparse data because it does not centralize the data before computing the SVD. The n components parameter allows you to choose the number of features you wish to see in the outcome. The number of components, n must be strictly smaller than the number of features in the input matrix. In our dataset, the n\_component is 2 because we want to represent the data in a 2D plot.

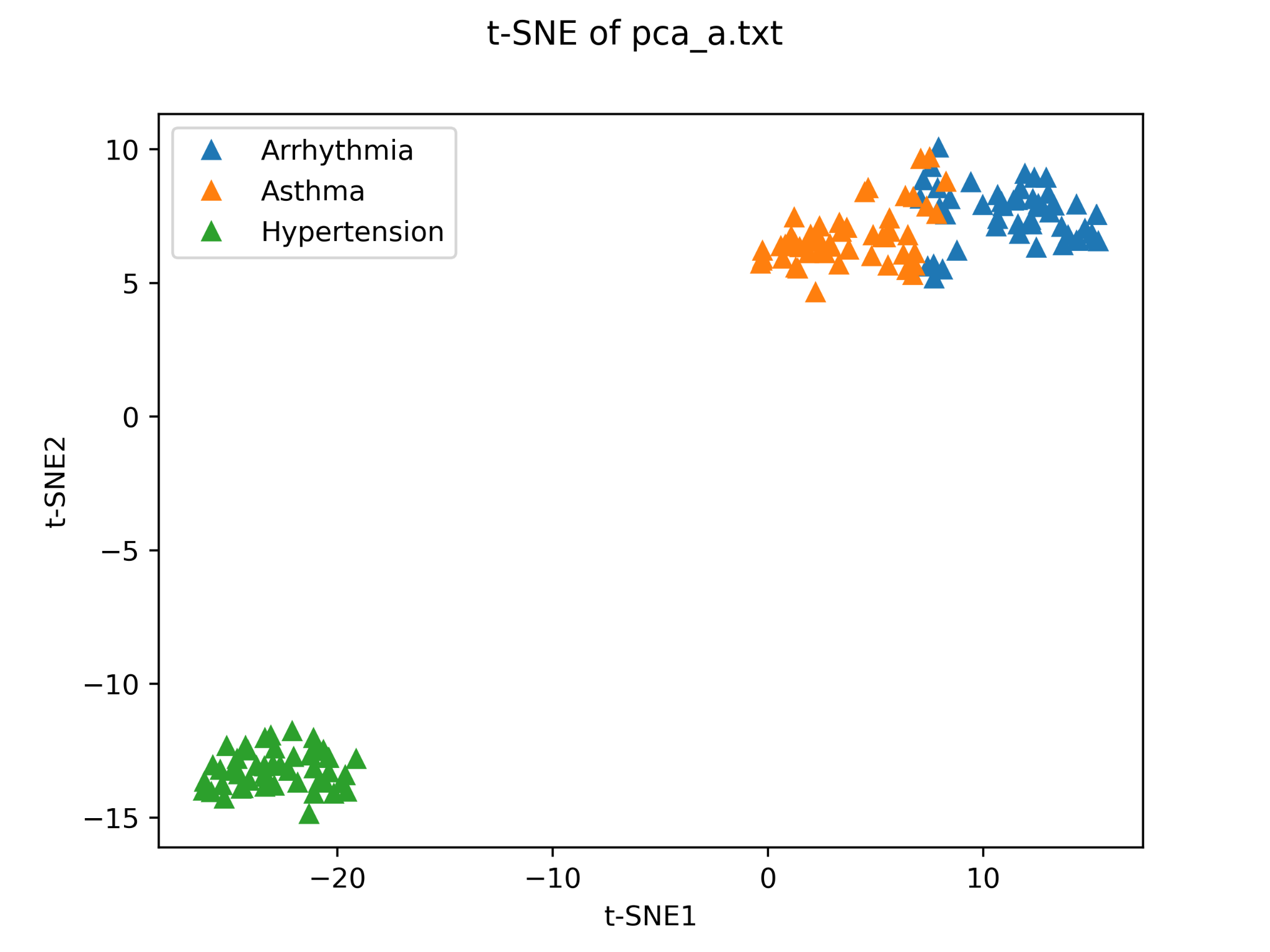
**T-Distributed Stochastic Neighbourhood Embedding** (t-SNE) is a nonlinear dimensionality reduction technique used for separating the data using a straight line and it majorly deals with linearly non-separable data. Although the t-SNE algorithm is advantageous in terms of flexibility, it’s quite tricky to interpret the data. The outcome is never the same because the process is iterative and generates distinct results for each runtime. The most important aspect of this algorithm is to preserve the distance between data points amongst each neighborhood but the distances between different neighborhoods are not preserved. Additionally, t-SNE is unable to maintain the neighborhood's overall size or shape.

**Dataset/ experiment results**

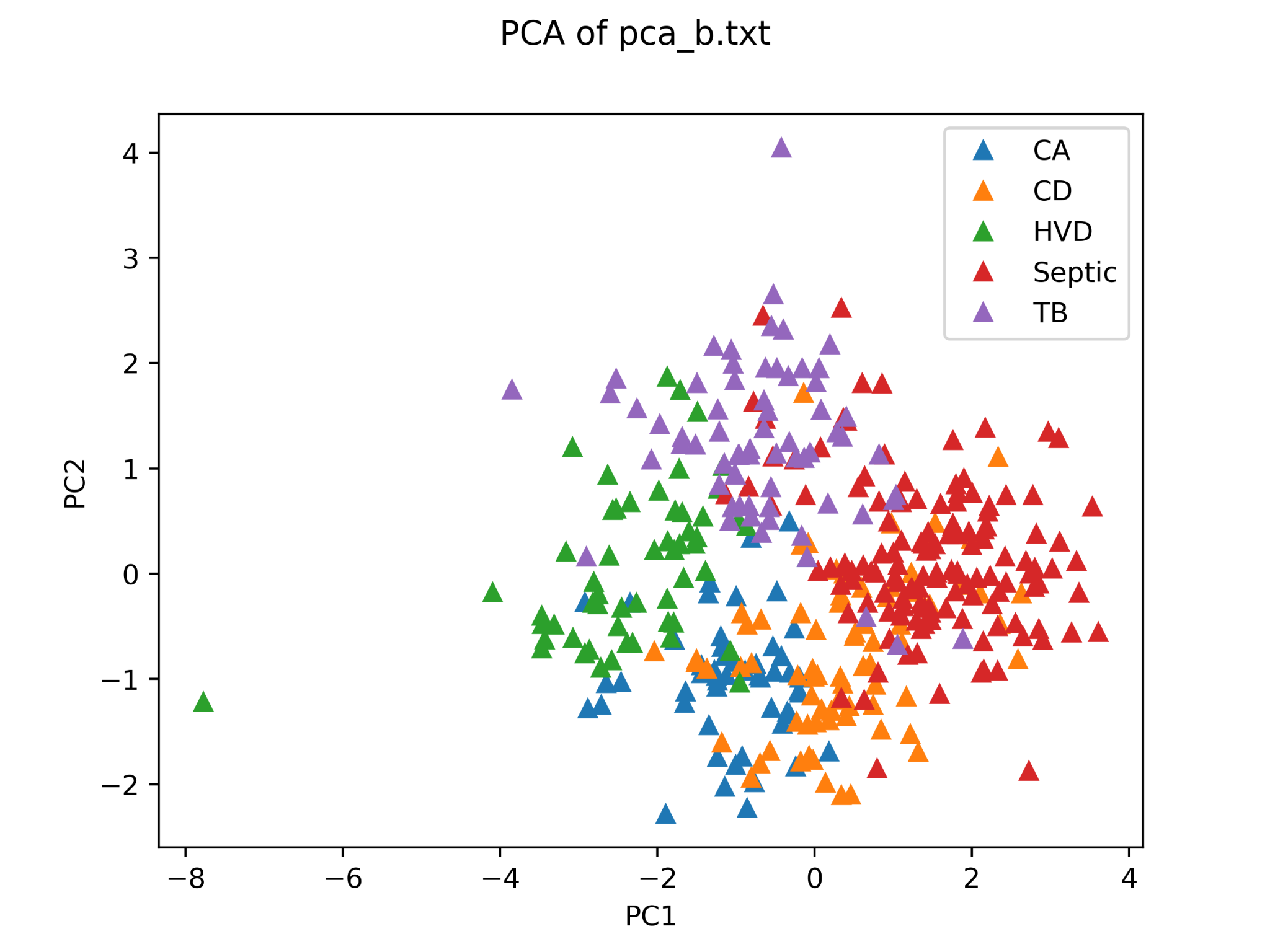
**Dataset pca\_a**: This dataset has 150 instances in total, 4 features and 3 disease classes Arrhythamia, Asthma and Hypertension with 50 instances each

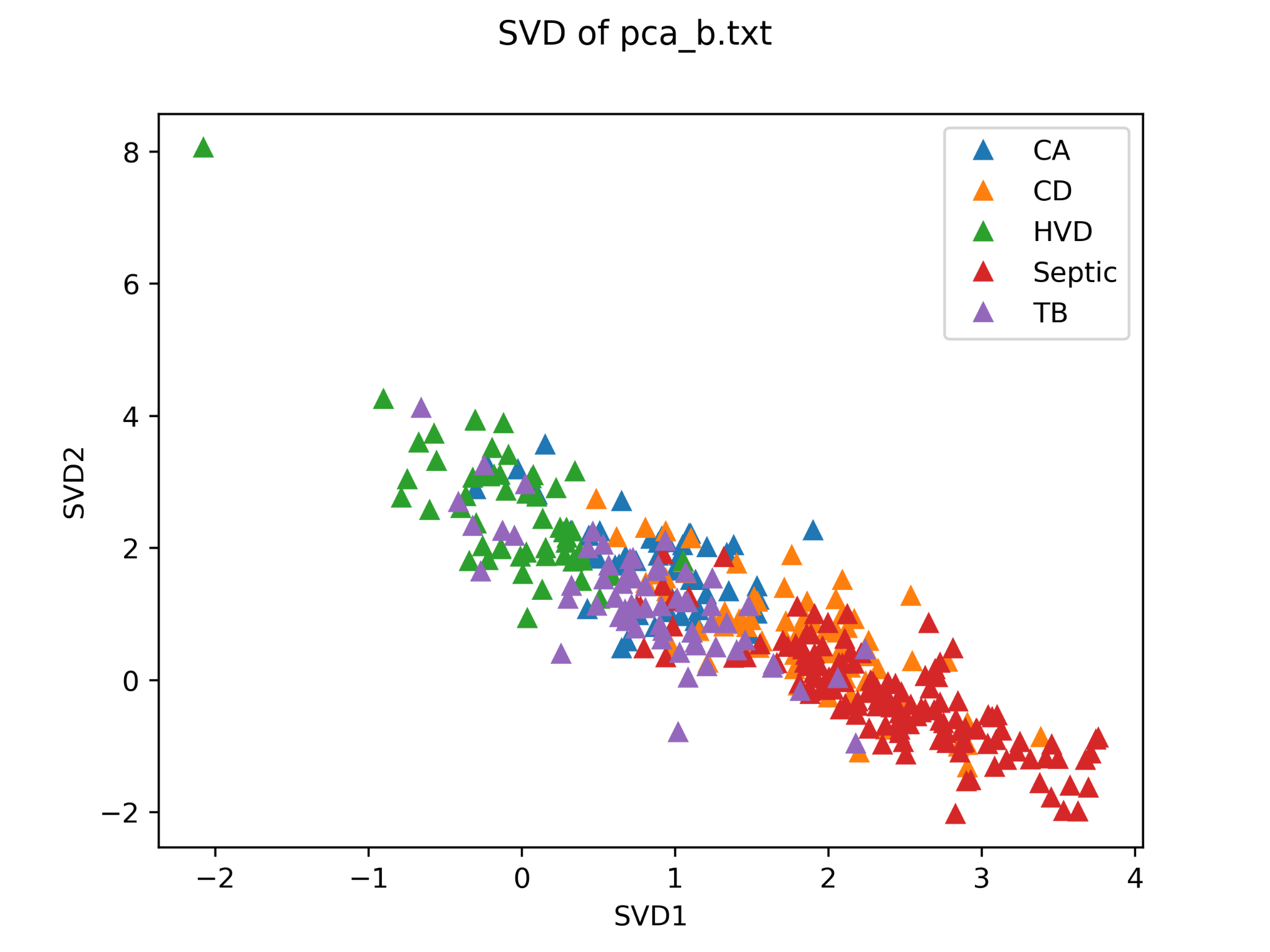


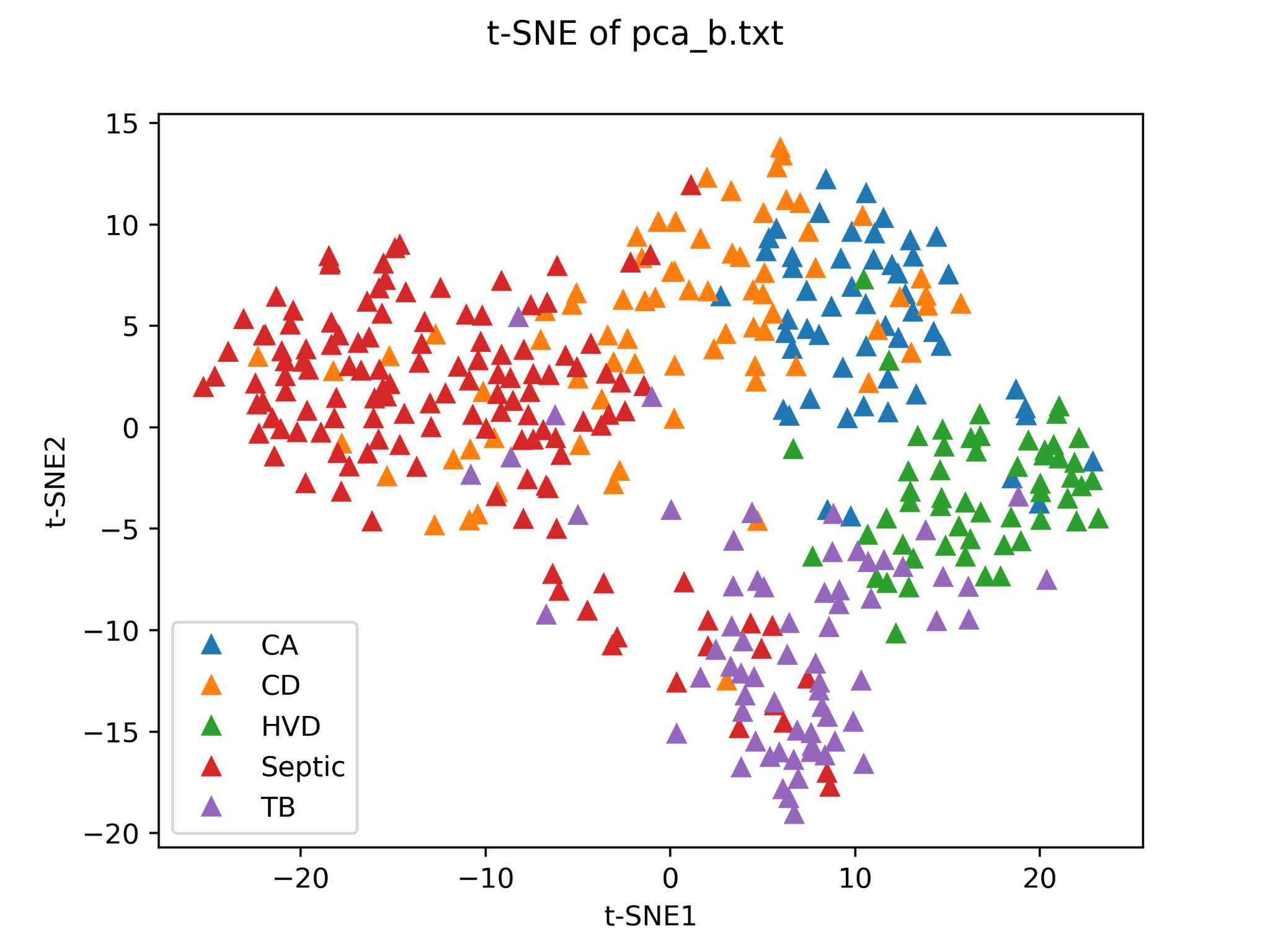




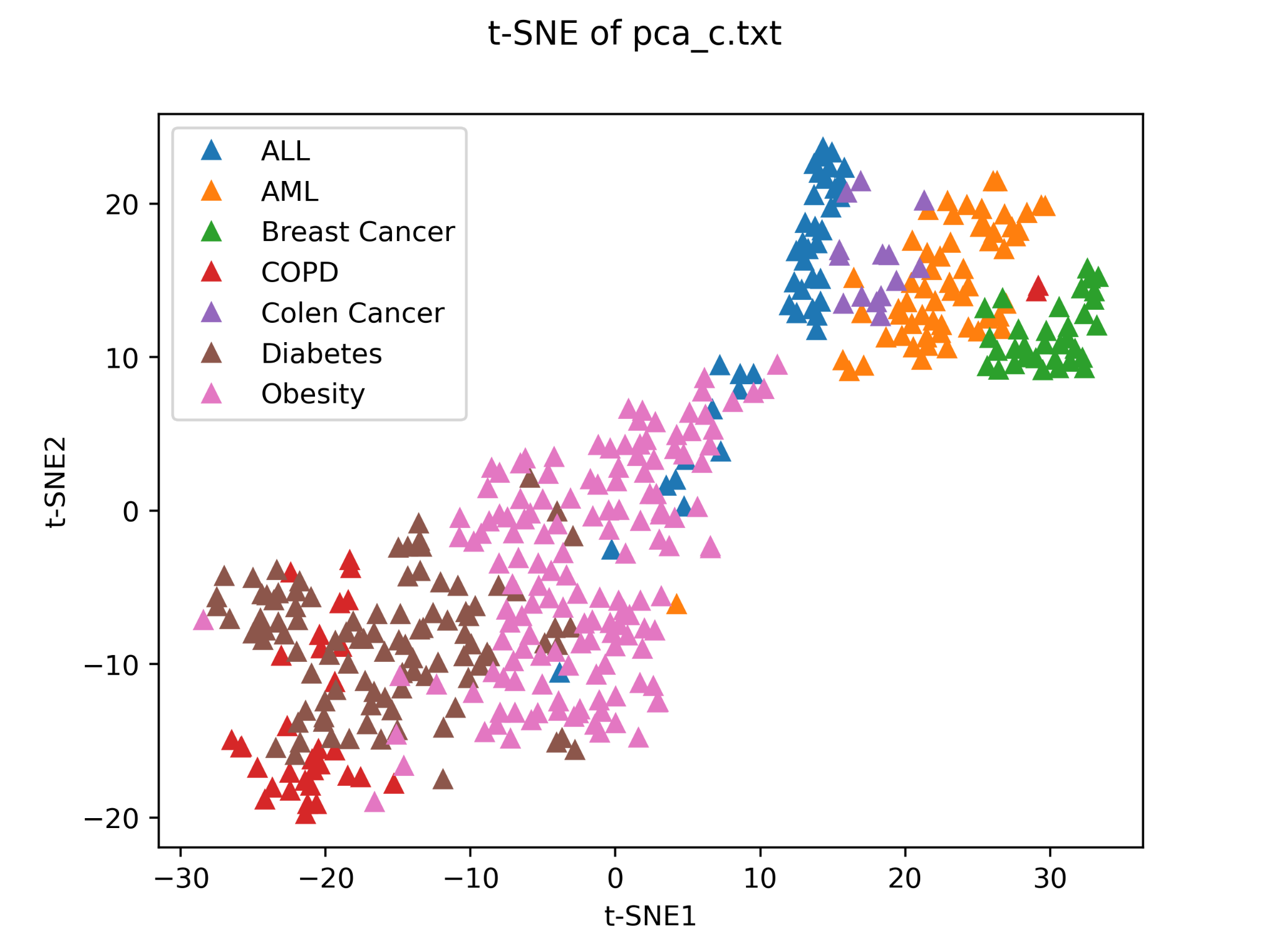
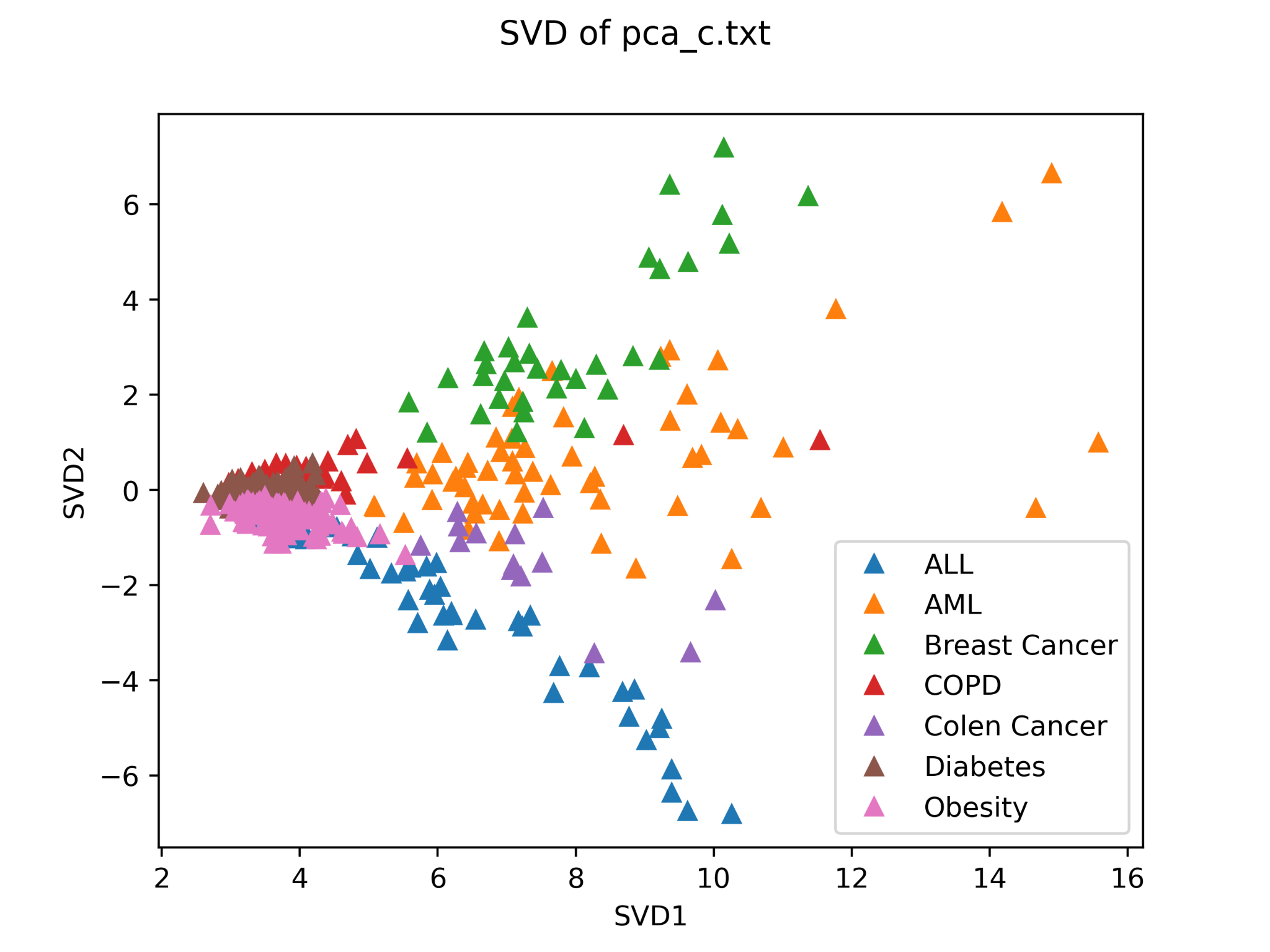
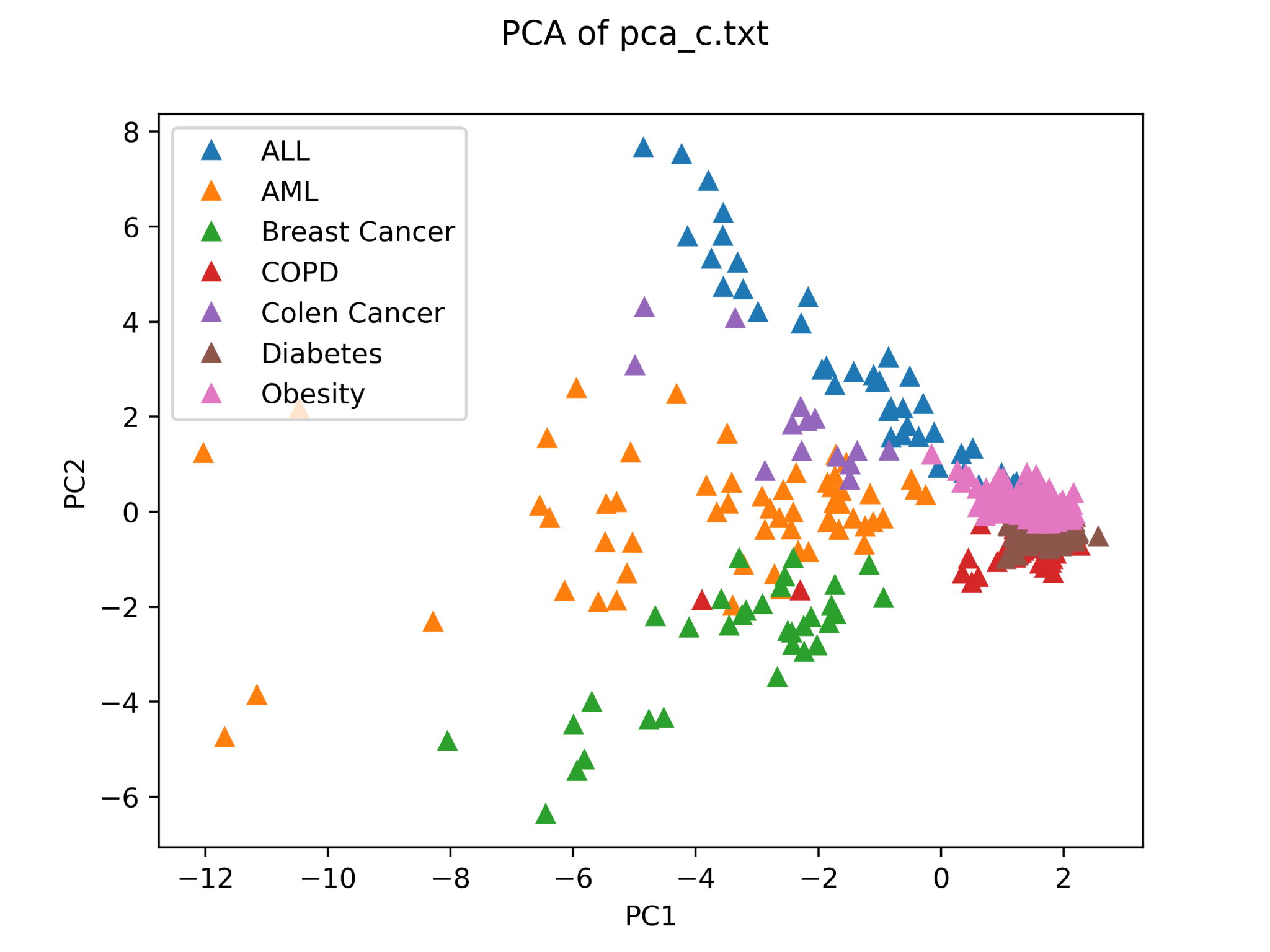
**Dataset pca\_b**: This dataset has 368 instances in total, 16 features and disease classes Septic with 135, HVD with 67, TB with 55, CD with 75 and CA with 54 instances







**Dataset pca\_c**: This dataset has 428 instances in total, 11 features and disease classes Obesity with 145, Diabetes with 100, ALL with 43, COPD with 34, AML with 59, Breast Cancer with 33 and Colen Cancer with 14 instances



**Conclusion**

In order to visualize a high-dimensional dataset in 2D and 3D plots, we explored different dimensionality reduction approaches. The PCA and truncate SVD are based on the notion that eigenvectors with lower eigenvalues should be eliminated, their differences are minimal. Results from t-SNE are superior to those from PCA and other linear dimensionality reduction techniques. This is because curved manifolds are difficult to represent using a linear approach, such as classical scaling. This t-SNE identifies clusters in the data and attempts to keep similar instances close together and different instances apart while reducing dimensionality. But t-SNE also has cons such as computational complexity, hyperparameter tuning. Thus, there is simply no method for transforming high-dimensional data into low dimensions while also maintaining the entire structure thus every approach will indeed have certain benefits when compared to others.