RESEARCH

Project-4

Sujan Darai (sujad96@zedat.fu-berlin.de), Matanat Mammadli (matanam94@zedat.fu-berlin.de), Samra Hamidovic (samrah96@zedat.fu-berlin.de)

Full list of author information is available at the end of the article

Abstract

Goal of the project: High-dimensional data visualization using the digits and breast cancer toy datasets involves applying dimensionality reduction techniques such as PCA, t-SNE, and UMAP. This project also includes developing an understanding of high-dimensional data visualization.

Main results of the project: The performance of 2D and 3D PCA embedding scatterplots is lower than the t-SNE and UMAP due to the nonlinearity and complexity of datasets for digits and breast cancer datasets. For the metabolomics dataset, points are spread all over meaning that samples exhibit a broad range of metabolic profiles. Both t-SNE and UMAP embeddings show better performance on datasets with increased neighbors and perplexity and with informative initialization rather than random.

Personal key learning:

1 Sujan Darai: Writing code for dimensionality reduction techniques and understanding the high dimensional figures
2 Samra Hamidovic: Writing code for different datasets and creating UMAP, t-SNE, and Scatterplots
3 Matanat Mammadli: Writing code for the metabolomics dataset and visualizing it with PCA, UMAP, and t-SNE

Estimated working hours:

1 Sujan Darai: 7 hours
2 Samra Hamidovic: 7 hours
3 Matanat Mammadli: 7 hours

Project evaluation: 1

Number of words: 2990

Keywords: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP)

1 Introduction

High-dimensional data, characterized by a large number of features (often more than 30), presents challenges such as the curse of dimensionality, computational complexity, and the risk of overfitting. Visualizing high-dimensional data requires techniques that reduce the number of features while retaining significant information to facilitate interpretation. These datasets contain complex features, making it difficult to extract meaningful insights directly.

Techniques such as Principal Component Analysis (PCA), Multidimensional Scaling (MDS), and Linear Discriminant Analysis (LDA) reduce dimensions linearly through projection methods. In contrast, t-SNE is a popular non-linear dimensionality reduction technique that preserves the local structure of data, and UMAP offers both global and local structure preservation in non-linear data. Both linear

and non-linear dimensionality reduction techniques can reduce data to 2D or 3D for visualization purposes.
These dimensionality reduction techniques enable effective visualization of high-dimensional data in lower dimensions, uncovering hidden patterns and relationships that are not apparent in the original high-dimensional space. Visualization tools such as matplotlib, seaborn, and Plotly in Python assist in creating these visualizations for exploratory data analysis.

2 Goal of the project

The main goal of this task is to perform visualization of high-dimensional data employing the digit and breast cancer dataset from the scikit-learn library by using dimensionality reduction techniques such as PCA, t-SNE, and UMAP. One of the purposes is to alter complicated, high-dimensional data into low-dimensional, interpretable 2D or 3D graphical expressions, making it easier to identify buried patterns in high dimensions and quantify relations between features within the data. Further, the task also involves an understanding of the principles and challenges associated with high-dimensional data visualization without losing the important insights and features of the original data during the dimensional reduction process, and the practical application of these techniques using Python visualization tools such as matplotlib, seaborn libraries. Additionally, the task also includes the application of at least one of the t-SNE and UMAP techniques for the MetabComparisonBinaryML dataset.

3 Data and preprocessings

3.1 Data

The toy datasets such as digits and breast cancer from the scikit-learn library are used in this report and data visualization for classification problems. The digits dataset comprises 1797 samples of pictures with 10 classes ranging from 0-9. Each data point is represented with the 8x8 pixel matrix giving a total of 64 features. On the other hand, the breast cancer Wisconsin dataset contains 569 breast cancer instances, with each case depicted by 30 diverse estimations of cell characteristics from breast biopsies. This information makes a difference in creating models that can anticipate whether a tumor is malignant or benign for medical diagnostics. Both datasets are equally important for investigating how strategies like PCA, t-SNE, and UMAP can change complex, high-dimensional information into something meaningful.
For this report, one of the metabolomics datasets, the MTBLS136 dataset was chosen, which consists of serum LC-MS datasets. This dataset consists of 949 named metabolites associated with postmenopausal hormone use. Multiple metabolites were compared in the group of women including 332 estrogen users and 337 estrogen plus progestin users versus 667 non-users.

3.2 Prepropcessing

In the beginning, both the digits and breast cancer dataset were loaded in the Jupyter Notebook from the scikit-learn library. Both datasets contain a numpy array of data and targets. For the digits dataset, the target has 10 classes of numbers

from 0 to 9 while for breast cancer, it has only 2 classes 0 and 1 where 0 represents benign and 1 represents malignant conditions. Since both of the datasets have only numerical columns or features, null values were checked and no null value was found. The digits and breast cancer datasets were ready to use for the dimensionality reduction procedures.
For the metabolomics MTBLS136 dataset, data was uploaded to the Jupyter Lab. It could be easily read as an Excel file and its containing elements were displayed in the Jupyter Lab. The dataset contains SampleIDs, Classes (which is a categorical variable that differentiates between different groups in our study), Hormone (represents different groups in our study, Hormone users: non-users, Estrogen only users and Estrogen + Progestin users) and 949 Metabolites. Class 0 represents Estrogen-only users, class 1 Estrogen + Progestin users, class 2 Non-users and class 3 had missing Hormone values (no entry values), representing nothing. The dataset Excel file was already in good condition, we just handled missing values by filling all NaN (Not a Number) entries with 0. Then the data was ready for the dimensionality reduction methods.

4 Methods

4.1 Principal component analysis

Principal Component Analysis (PCA) is a statistical method utilized to simplify and explore complex datasets by transforming them into a new set of uncorrelated variables known as principal components.[1] In this study, PCA was employed to reduce the dimensionality of the dataset and uncover underlying patterns within the data.
In PCA, the datasets are generally standardized by centering the values (subtracting the mean) and scaling them (dividing by the standard deviation) to ensure equal contribution from all variables. Subsequently, the covariance matrix of the standardized data was computed to understand the relationships between variables. By calculating the eigenvectors and eigenvalues of the covariance matrix, PCA determined the directions of the principal components and the amount of variance explained by each component. This process aided in selecting the most informative components that captured the essential variability in the data.
Through the transformation of the original data into a new coordinate system defined by the principal components, PCA facilitated the exploration of data patterns, noise reduction, and visualization of high-dimensional data in a more manageable lower-dimensional space. Overall, PCA served as a valuable tool in extracting meaningful insights and understanding the underlying structure of the dataset.

4.2 t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding): t-SNE is a popular dimensionality reduction technique commonly used in single-cell transcriptomic data analysis. It aims to visualize high-dimensional data by mapping data points to a lower-dimensional space, typically two dimensions. t-SNE works by modeling the similarity between data points in the high-dimensional space and the low-dimensional embedding. It focuses on preserving the local structure of the data, ensuring that

similar data points are represented close to each other in the visualization. t-SNE minimizes a cost function using gradient descent to optimize the embedding. [2]
In t-SNE, the initialization process involves setting up the initial configuration of data points in the low-dimensional space. This can be done either randomly or based on principal component analysis (PCA). Random initialization places data points randomly in the low-dimensional space at the start of the optimization process, while PCA initialization uses PCA to project the data onto principal components before optimization begins [2]. For our study, the initialization was performed randomly and with PCA on digits and cancer datasets.

4.3 UMAP

UMAP (Uniform Manifold Approximation and Projection): UMAP is another dimensionality reduction technique widely used in single-cell analysis. Similar to t-SNE, UMAP aims to project high-dimensional data into a lower-dimensional space for visualization. UMAP focuses on preserving both local and global structures of the data, making it suitable for capturing complex relationships in the data. UMAP achieves this by constructing a low-dimensional representation of the data based on a fuzzy topological structure. It uses a combination of repulsive and attractive forces to optimize the embedding, resulting in a more balanced representation of the data compared to t-SNE. [2] In Kobak's paper, UMAP utilizes Laplacian eigenmaps (LE) for initialization. LE is an algorithm that helps in achieving a globally accurate embedding of the data, providing a more informed starting point for the optimization process. This initialization method in UMAP contributes to preserving both local and global structures of the data during the embedding process [2]. A different approach was conducted in this study: UMAP was performed on digit and cancer datasets with different neighbor numbers (10, 50, 90) and different distances (6).

5 Results and discussion

Figure 1 shows the results of Principal Component Analysis (PCA) applied to two different datasets (digits and breast cancer) to reduce their dimensions to 2, and visualizes the results using scatter plots. The PCA reduces the datasets to 2 principal components. In the digits dataset, there are 9 classes represented by different colors, and in the breast cancer dataset, there are 2 classes (0 meaning healthy and 1 meaning diseased). Classes with the same color are generally clustered together, although there are some outliers. In the breast cancer dataset, healthy components are more spread out, whereas cancer components are tightly and densely clustered together, indicating that the cancer class has more homogeneous feature patterns.

Figure 2 represents a PCA that reduces the dimensionality of the digits dataset to 3 dimensions and visualizes the relationships between these three principal components. Each cell in this grid represents a scatter plot of one variable against another, allowing us to see pairwise relationships between all variables at once. Here we observe histograms (Frequency distributions of the data along the single Principal Component, PC1 vs PC1, PC2 vs PC2, and PC3 vs PC3) on the diagonal and scatter plots (PC1 against PC2, and PC1 against PC3, PC2 against PC1 and PC2 against PC3, PC3 against PC1 and PC3 against PC2) off-diagonal. PCA

components 1, 2, and 3 refer to the new variables or axes created by the PCA algorithm. These components are linear combinations of the original variables and are designed to capture the maximum variance (1 being the largest amount of variance, representing the direction in which the data varies the most, 2 representing the direction of the second most variance, orthogonal to the first component and 3 adding another layer of orthogonal variance capture, which is helpful in 3D visualizations) in the data in a sequential manner.
On the diagonal line, we see the frequency distribution of the data of the PC1, it has a single peak, indicating that most data points are centered around a particular value. However, the histogram of the PC2 against PC2 shows a multi-modal distribution, indicating multiple clusters. The Histogram of PC3 vs PC3, once again, showcases a single peak.
Off-diagonal we see the digit classes building clusters within their class (colors), indicating that the data naturally separates into distinct groups. However, there are a few outliers, especially on PC2 vs PC3 and PC3 vs PC2 plots, which might suggest that PC2 and PC3 are not sufficient to separate the classes.
Overall, in most scatter plots the classes are well separated, suggesting that the principal components are effective in distinguishing between classes.

Figure 3 shows the t-SNE plot of the breast cancer and digits dataset. t-SNE is applied to both of these datasets to reduce their dimensions to 2, making them suitable for 2D visualization. We used t-SNE with random initialization, and by setting the random state to 42, we made random initialization deterministic and reproducible. Different colors represent different classes in each dataset (2 classes in the breast cancer data and 9 classes in the digits dataset).
Two classes (healthy and diseased) are well-separated and distributed, meaning t-SNE has successfully captured the intrinsic differences between the classes in the breast cancer data. However, there are some class components overlapping, aka outliers, but not a significant amount.
In the digits dataset, each digit class forms its own cluster, which is the ideal case. There are very few overlapping cases or outliers, which suggests that some digits have similar features that t-SNE struggles to distinguish, but these are very few in our case. Clusters are well-separated, indicating that the digits have distinct feature representations. The only barely noticeable overlapping is between classes 1 and 8, which might suggest that those digits have similar features. All clusters are densely packed, indicating consistent feature patterns within a digit class.

Figure 4 illustrates the t-SNE scatterplot matrix (SPLOM) for the digits dataset (4a) and for the breast cancer dataset (4b).
When we look at the t-SNE scatterplot matrix for the digits dataset (4a), on the diagonal line we see histograms or the distribution of data along each component. It is noticeable that all 3 scatter plots (PC1 vs PC1, PC2 vs PC2, PC3 vs PC3) display a single peak, indicating that most data points are centered around a particular value. Off-diagonal scatter plots of the digits dataset show very well separated and grouped by class (9 colors) clusters, which indicates that the data points with similar features are grouped together and that t-SNE effectively captures the underlying structure of the data and shows a good performance. The distance between

similar data points within clusters is very close, while data points from different clusters are further apart, further confirming the effectiveness of t-SNE.
The t-SNE SPLOM for the breast cancer dataset (4b) on the diagonal line shows histograms with multi-modal distribution, suggesting that most data points are distributed around different values. However, the t-SNE Component 3 vs t-SNE Component 3 shows a single peak, indicating that most data points are centered around a particular value. Off-diagonal t-SNE scatter plots for the breast dataset present well-separated and grouped clusters of healthy and diseased classes, indicating effective dimensionality reduction by t-SNE. Only on TC2 vs TC3 and TC3 vs TC2 plots, it is to be observed that healthy and cancer data points overlap each other more than usual, indicating some outliers, which might suggest that TC2 and TC3 are not as sufficient to separate the classes as other t-SNE Components.

Figure 5 shows the t-SNE embeddings for the digits dataset (5a) and breast cancer dataset (5b).
On the left side, we see t-SNE embeddings with random initialization and on the right side with PCA initialization. The perplexity (the effective number of neighbors) of the plots is increased each time. In both datasets, with increased perplexity, clusters of the same classes become denser and more compact, indicating a more effective performance of t-SNE.
On the digits dataset (5a), there is no noticeable difference between the t-SNE plots initialized with random and PCA methods. Both methods yield similar visualizations of the dataset. However, on the breast cancer dataset (5b), the last plots demonstrate a notable difference: the t-SNE plot with perplexity 50 and random initialization shows more densely grouped and well-separated clusters for healthy and cancer classes. In contrast, the t-SNE plot with perplexity 50 and PCA initialization showcases still well-separated but much less dense, more loosely grouped classes of healthy and cancer data points, suggesting that while PCA initialization (wit perplexity = 50) still separates the classes, it does so with less compactness compared to random initialization.
According to the Kobak paper [2], t-SNE embeddings with informative initialization (such as PCA) perform more effectively and preserve global structure better compared to t-SNE with default random initialization.

Figure 6 shows UMAP embedding on the digits and breast cancer dataset.
It is remarkable that with increasing distance (minimum distance between points in the low-dimensional embedding space) clusters of both the digits and breast cancer classes got bigger and more dense.
With increasing neighbors, in the digits dataset, there is not much change in visualization but in the breast cancer dataset components got closer together and built bigger and more compact clusters. When n_neighbors is high, UMAP tends to focus more on preserving global structure, which could lead to more pronounced clustering as the algorithm tries to keep the clusters coherent and well-separated.
However, in both datasets, increasing minimum distance had more effect on grouping bigger clusters than neighbors, which might sound counterintuitive at first, but it can be due to complex interaction between neighbors and minimum spacing between points in the embedding space.

Figure 11 displays PCA embedding of the metabolomics dataset, with four mini-figures. They have perplexity 5 and 50 and are either random or PCA initialized.
It is noticeable, that with PCA initialization, colors or classes are clustered and grouped better together, which is also expected, as mentioned in the Kobak paper [2] t-SNE with PCA initialization performs better or more effectively.
With increased perplexity, classes or data points are also less spread out and more dense, which is also beneficial and indicates that t-SNE performing well.

Figure 16 with all its four mini-figures shows UMAP plots of the metabolomics dataset. They have neighbors 5 and 50, and minimum distance (minimum distance between embedded points in the low-dimensional space) of 0.1 and 0.9.
With increasing minimum distance, classes are clustered better together and they are more dense. It does make sense, because if min_dist is small, it allows for more local structure to be preserved, meaning that nearby points in the high-dimensional space will remain close to each other in the low-dimensional embedding. On the other hand, when min_dist is large, it encourages more global structures to be preserved, leading to a more uniform spread of points in the low-dimensional embedding.
With increasing neighbors, clusters that classes build are larger and more dense.

6 Contributions

Sujan Darai: Implementing and running code for digits and cancer dataset, writing Introduction, Goal of the project and Data & preprocessings
Matanat Mammadli: Implementing and running code for metabolomics dataset, writing Data & preprocessings and Results & Discussion
Samra Hamidovic: Implementing and running code for digits and cancer dataset, writing Abstract, Methods and helped with introduction

7 Appendix

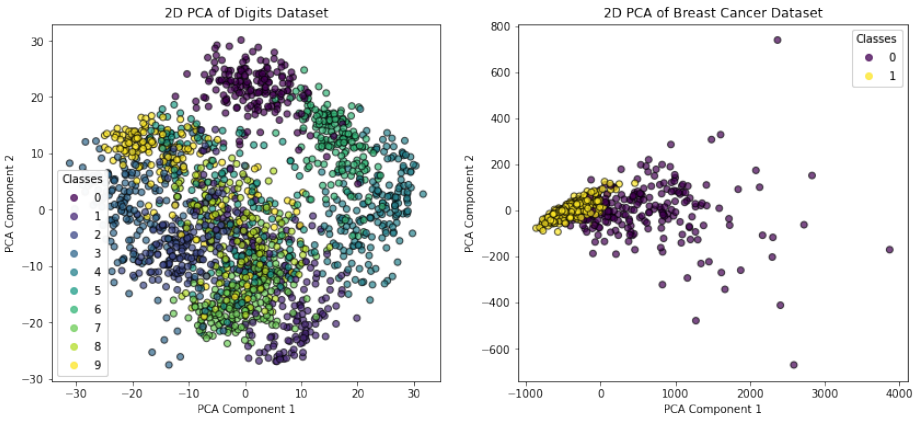
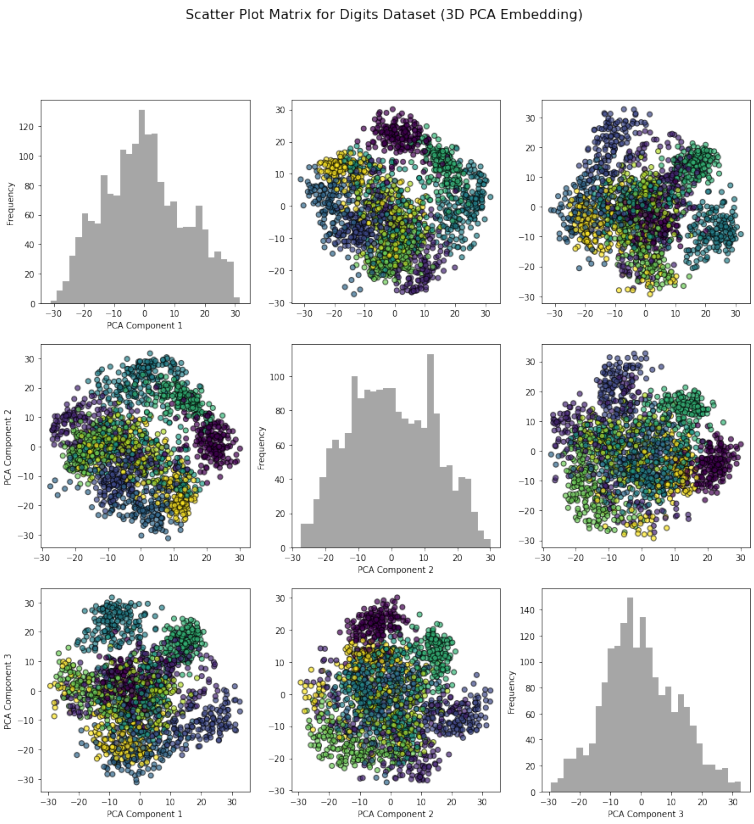
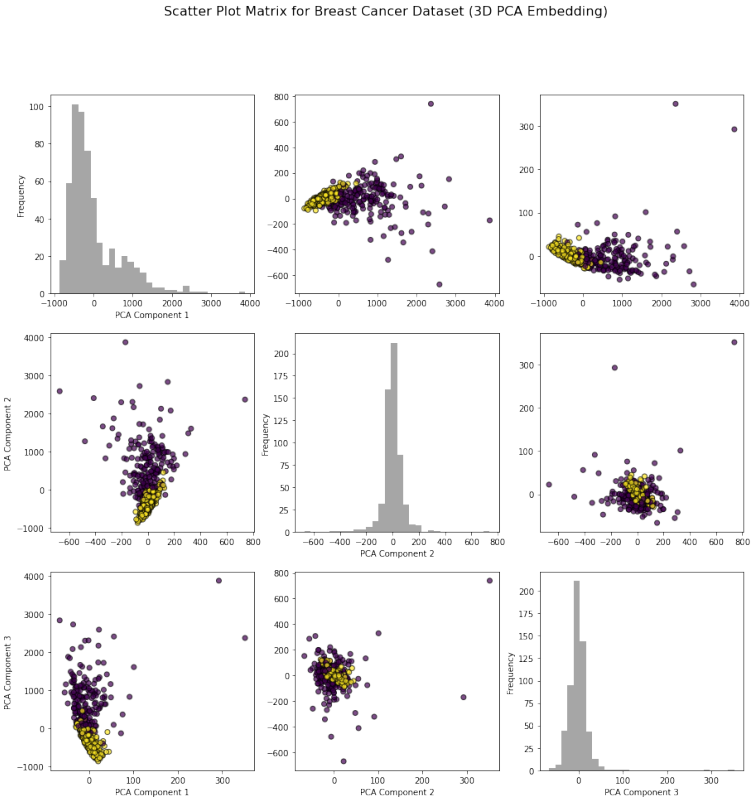


Figure 1: PCA plot of the digits and the breast cancer dataset



(a) on digits dataset



(b) on breast cancer dataset

Figure 2: Scatterplot of the digits and breast cancer dataset with PCA embedding

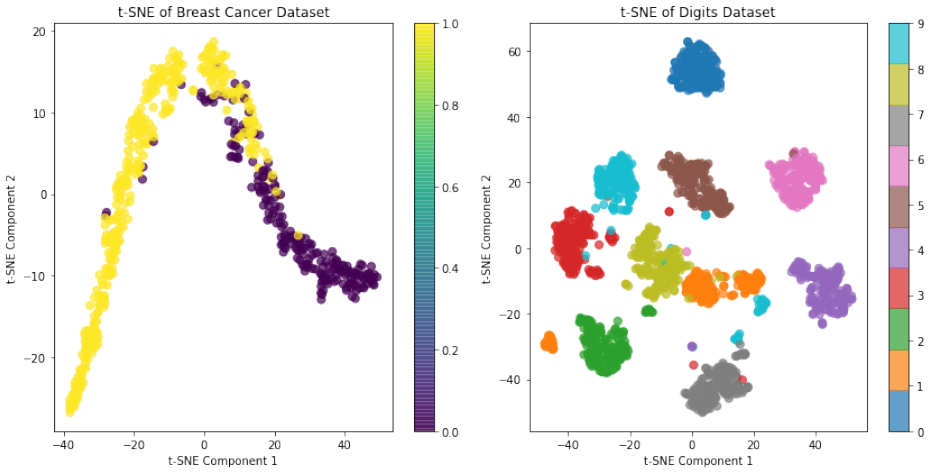
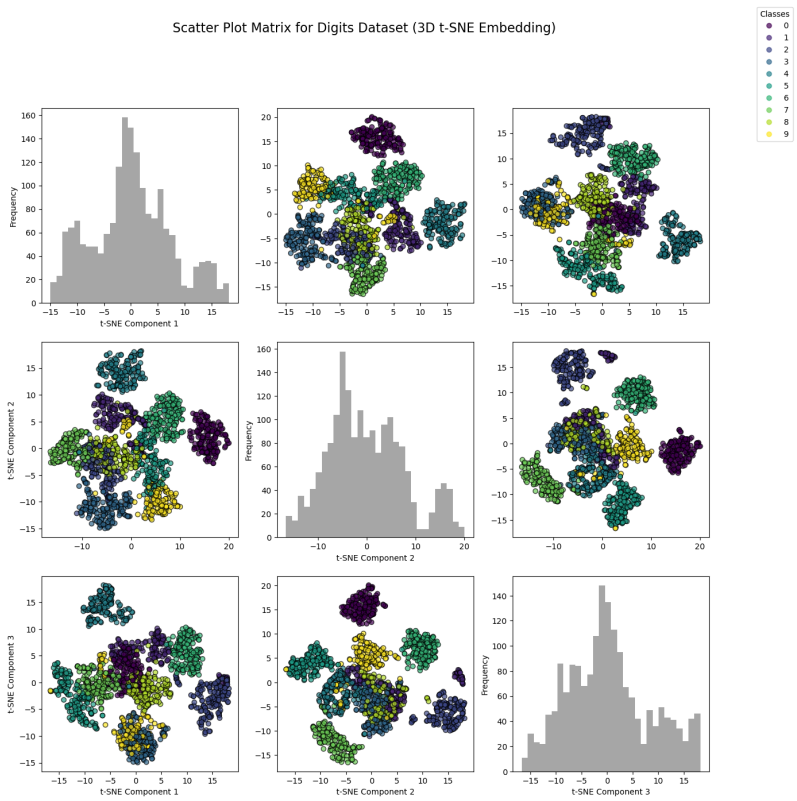
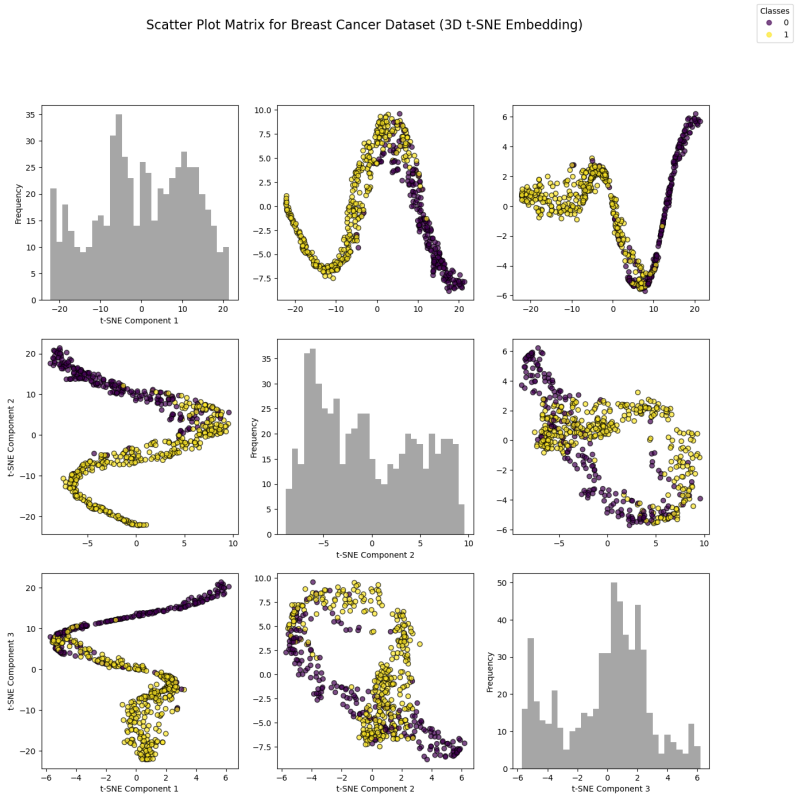


Figure 3: t-SNE plot of the breast cancer and digits dataset

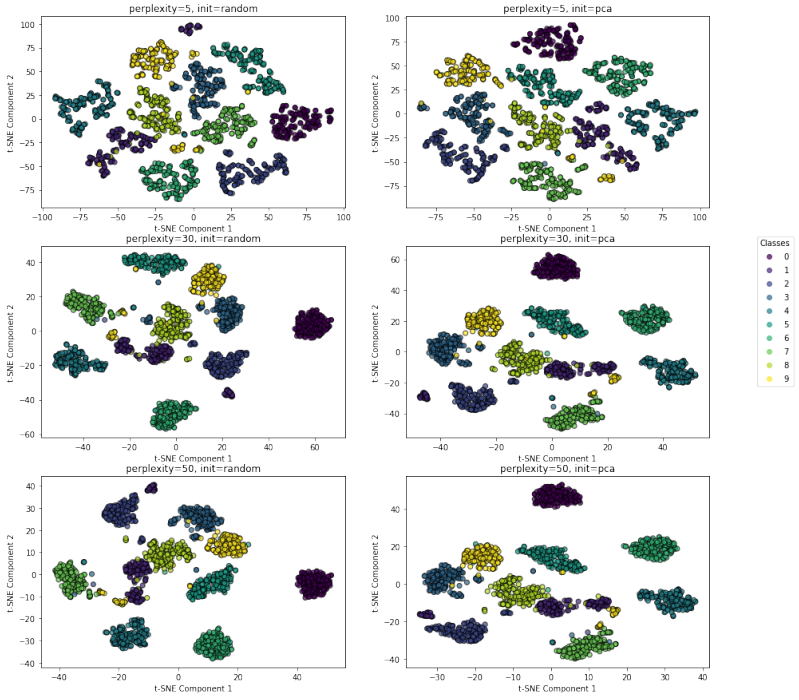


(a) on digits dataset

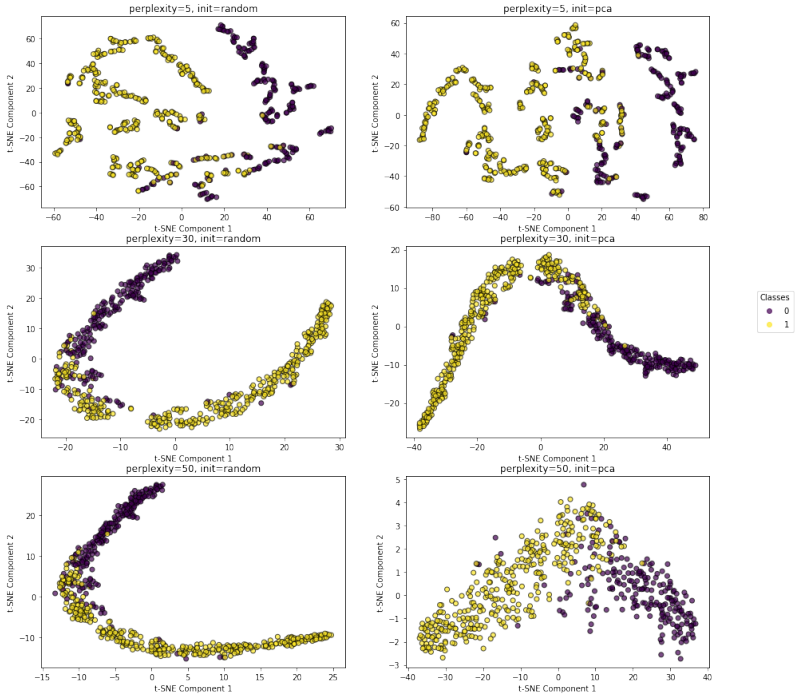


(b) on breast cancer dataset

Figure 4: t-SNE scatterplot on both datasets

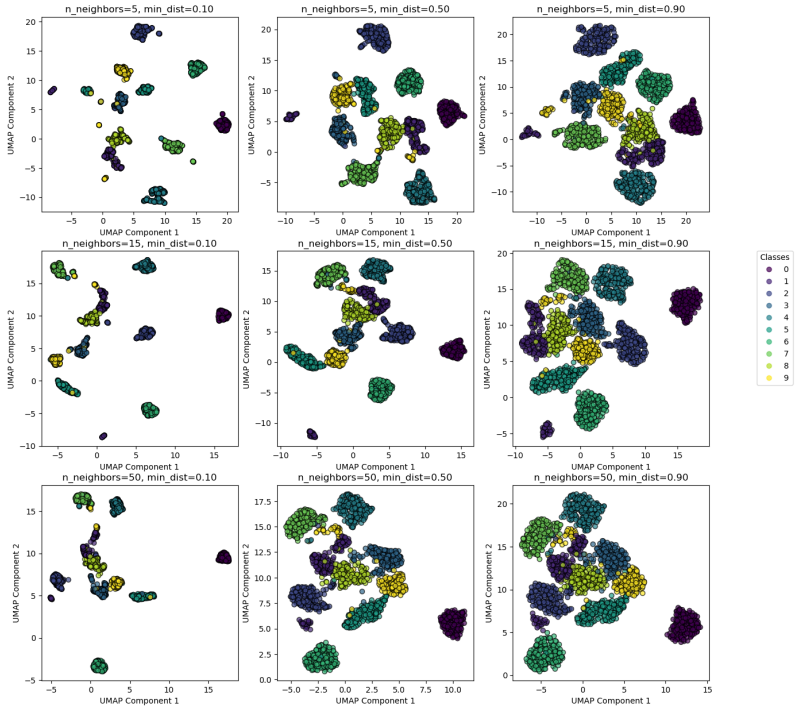


(a) on digits dataset

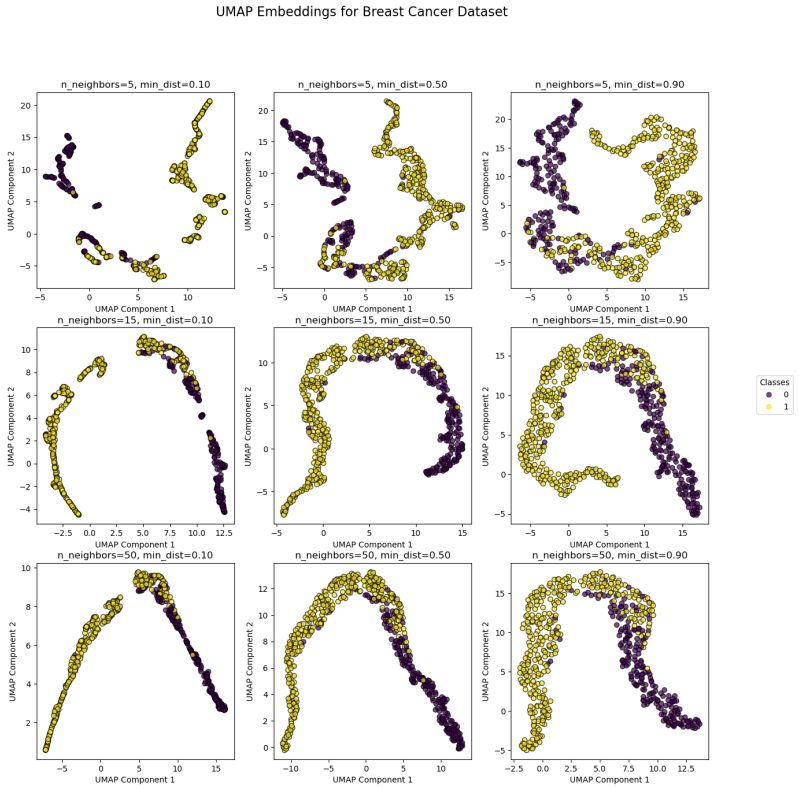


(b) on breast cancer dataset

Figure 5: t-SNE embedding on both datasets with random initialization (left) and PCA initialization (right)



(a) on digits dataset



(b) on breast cancer dataset

Figure 6: UMAP embedding on both datasets with 5 neighbours (top row), 15 neighbours (middle row) and 50 neighbours (bottom row)

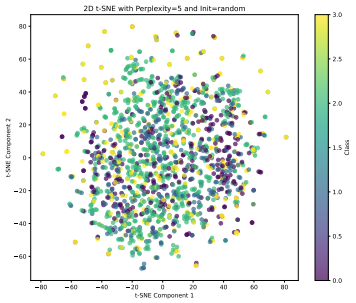


Figure 7: perplexity=5 and random initialization

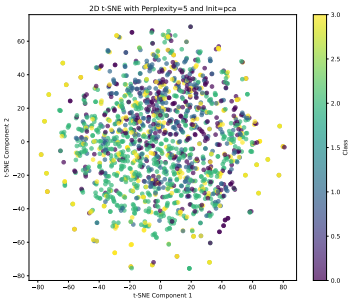


Figure 8: perplexity=5 and PCA initialization

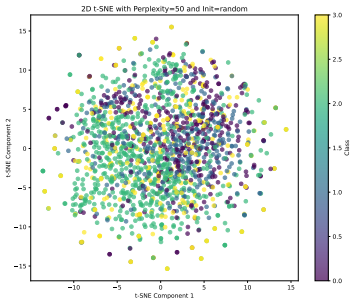


Figure 9: perplexity=50 and random initialization

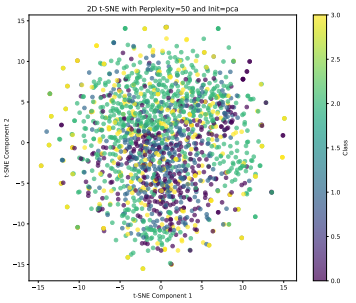


Figure 10: perplexity=50 and PCA initialization

Figure 11: t-SNE plots of metabolomics dataset

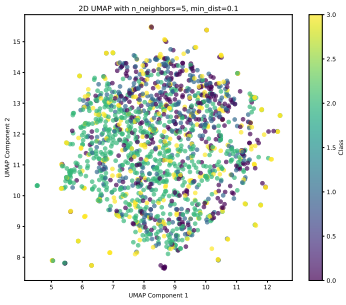


Figure 12: neighbors=5 and min_dist=0.1

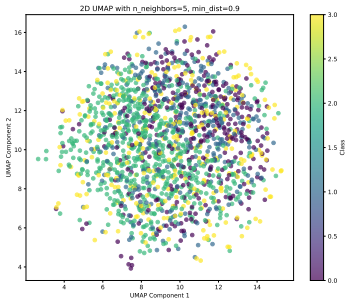


Figure 13: neighbors=5 and min_dist=0.9

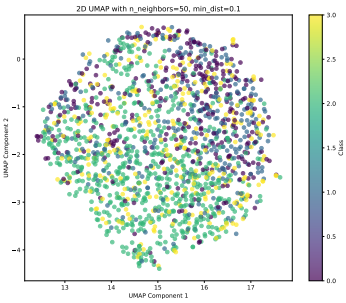


Figure 14: neighbors=50 and min_dist=0.1

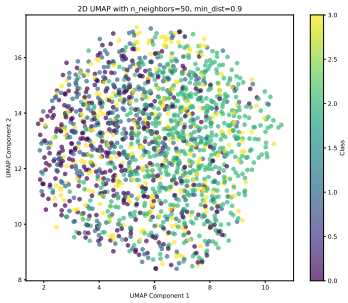


Figure 15: neighbors=50 and min_dist=0.9

Figure 16: UMAP plots of metabolomics dataset

References

1. Rasmus Bro and Age K. Smilde. Principal component analysis. *Analytical Methods*, 6(9):2812–2831, 2014. Publisher: Royal Society of Chemistry.
2. Dmitry Kobak and George C. Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, 39(2):156–157, February 2021.