Dimensionality Reduction Techniques: A Comparative Study

Panagiotis Georgitseas - CSD 4609
Vorgias Dimitris - CSD 4604

February 11, 2025

Abstract

Dimensionality reduction is a crucial preprocessing step in data science and machine learning. It helps in reducing computational cost, mitigating the curse of dimensionality, and improving model interpretability. In this study, we compare several dimensionality reduction techniques applied to two datasets: the Breast Cancer dataset and the Fashion MNIST dataset.

1 Introduction

Dimensionality reduction techniques allow for efficient data processing while preserving important patterns. We investigate various linear and non-linear methods to analyze their effectiveness on structured and unstructured data.

2 Datasets

2.1 Breast Cancer Dataset

The Breast Cancer dataset contains 30 numerical features extracted from digitized images of fine needle aspirates of breast masses. These features describe characteristics such as radius, texture, smoothness, and compactness. Since the dataset is relatively low-dimensional and exhibits linear separability, PCA is expected to perform well.

2.2 Fashion MNIST Dataset

The Fashion MNIST dataset consists of grayscale images of clothing items with a resolution of 28x28 pixels. This dataset is high-dimensional and contains complex, non-linear relationships. Hence, non-linear techniques such as t-SNE and Isomap are expected to provide better separation and clustering.

3 Dimensionality Reduction Techniques

3.1 Principal Component Analysis (PCA)

PCA is a linear technique that transforms the data into a lower-dimensional space by maximizing variance. The transformation is defined as:

$$Z = XW \tag{1}$$

where X is the original data matrix, W is the matrix of eigenvectors of the covariance matrix of X, and Z is the transformed data.

3.2 Kernel PCA

Kernel PCA extends PCA by applying a kernel function to capture non-linear structures in the data. It projects the data into a high-dimensional space before performing PCA:

$$K = \phi(X)^T \phi(X) \tag{2}$$

where $\phi(X)$ is the transformation function and K is the kernel matrix.

3.3 Independent Component Analysis (ICA)

ICA seeks to find statistically independent components in the data by maximizing non-Gaussianity. It is formulated as:

$$X = AS \tag{3}$$

where X is the observed data, A is the mixing matrix, and S is the source signals.

3.4 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE minimizes the divergence between two probability distributions, one in the high-dimensional space and one in the low-dimensional space. The optimization objective is given by:

$$C = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{4}$$

where p_{ij} and q_{ij} represent pairwise similarities in the high-dimensional and low-dimensional spaces.

3.5 Isomap

Isomap extends classical MDS by incorporating geodesic distances computed from the nearest neighbors graph. It preserves the global structure of the data by using:

$$D_{ij} = \min \sum_{(a,b) \in \text{path}} d_{ab} \tag{5}$$

where D_{ij} is the geodesic distance between points i and j.

3.6 Locally Linear Embedding (LLE)

LLE seeks to preserve local relationships between points by reconstructing each point as a linear combination of its neighbors:

$$X_i = \sum_{j \in \mathcal{N}(i)} W_{ij} X_j \tag{6}$$

where W_{ij} are the reconstruction weights.

3.7 Multidimensional Scaling (MDS)

MDS reduces dimensionality by preserving pairwise distances between points:

$$\min \sum_{i,j} (D_{ij} - d_{ij})^2 \tag{7}$$

where D_{ij} is the original distance and d_{ij} is the low-dimensional representation.

3.8 Truncated Singular Value Decomposition (Truncated SVD)

Truncated SVD is a linear method that decomposes the data matrix X into:

$$X = U \Sigma V^T \tag{8}$$

and keeps only the top k singular values.

4 Results and Discussion

Figures below show the 2D and 3D representations obtained using different techniques. PCA performs well for the Breast Cancer dataset due to its linear separability, whereas t-SNE and Isomap provide better clustering in Fashion MNIST.

Classification Results

Method	Random Forest	SVM	Logistic Regression
PCA	0.98	0.97	0.97
Kernel PCA	0.97	0.97	0.92
ICA	0.96	0.98	0.97
Truncated SVD	0.96	0.97	0.98
MDS	0.96	0.96	0.97
LLE	0.96	0.95	0.63
t-SNE	0.96	0.96	0.97
Isomap	0.96	0.96	0.96

Table 1: Accuracy results for different dimensionality reduction methods on the Breast Cancer dataset.

Method	Random Forest	SVM	Logistic Regression
PCA	0.61	0.63	0.63
Kernel PCA	0.62	0.63	0.56
ICA	0.63	0.63	0.62
MDS	0.65	0.66	0.58
LLE	0.72	0.64	0.33
t-SNE	0.82	0.76	0.74
Isomap	0.68	0.65	0.64

Table 2: Accuracy results for different dimensionality reduction methods on the Fashion-MNIST dataset.

4.1 Principal Component Analysis (PCA)

PCA is a **linear** method that maximizes variance in the lower-dimensional space. It is particularly effective for datasets that exhibit linear separability, such as the **Breast Cancer dataset**. As seen in the visualizations, PCA maintains well-defined clusters in the Breast Cancer dataset, allowing models like **SVM and Random Forest** to achieve **high classification accuracy (98%)**.

However, PCA struggles with complex, high-dimensional datasets like **Fashion MNIST**, where the data contains **non-linear relationships**. The resulting 2D representations lack clear class separation, which negatively impacts classification performance.

Despite these limitations, PCA is computationally efficient, requiring only **0.5 seconds for Breast Cancer** and **2.1 seconds for Fashion MNIST**, making it a fast and effective choice when working with large datasets where interpretability is important.

4.2 Kernel PCA

Kernel PCA extends standard PCA by using kernel functions to **capture non-linear structures** in the data. Unlike standard PCA, which assumes linearity, Kernel PCA allows more flexible transformations.

- For **Breast Cancer**, Kernel PCA performs similarly to PCA, achieving high classification performance.
- For **Fashion MNIST**, Kernel PCA **improves over PCA** by better capturing the non-linear variations, but it is still not as effective as **t-SNE or Isomap**.

A key limitation of Kernel PCA is its **higher computational cost**, which can be significant for large datasets.

4.3 Independent Component Analysis (ICA)

ICA is designed to separate **independent latent sources** in data by maximizing statistical independence. Unlike PCA, which focuses on variance, ICA attempts to extract signals that are non-Gaussian and independent.

- In **Breast Cancer**, ICA performed well, though slightly worse than PCA.
- In **Fashion MNIST**, ICA struggled to form well-defined clusters. The extracted features were not as effective in separating the different classes, leading to a **drop in classification accuracy** compared to t-SNE and Isomap.

ICA is particularly useful when working with **signal processing** and **blind source separation**, but it may not always be the best choice for visualizing complex datasets like Fashion MNIST.

4.4 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a **non-linear** technique that is highly effective at capturing **complex structures in high-dimensional data**. Unlike PCA, which seeks to preserve global variance, t-SNE focuses on **local similarities**, making it well-suited for datasets with **non-linear class boundaries**, such as **Fashion MNIST**.

- t-SNE **significantly outperforms PCA and ICA** in Fashion MNIST by forming distinct, well-separated clusters.
- However, it is **computationally expensive** (**7.9s for Breast Cancer, 8.4s for Fashion MNIST**) and does not preserve **global structure**, making it unsuitable for datasets where maintaining relative distances is important.

Despite its high computational cost, **t-SNE remains one of the best choices for high-dimensional, complex datasets** where clustering is critical.

4.5 Isomap

Isomap is an **extension of Multi-Dimensional Scaling (MDS)** that preserves **geodesic distances** rather than Euclidean distances. This makes it well-suited for **highly curved manifolds** where linear techniques fail.

- For **Breast Cancer**, Isomap performed comparably to PCA, maintaining clear class separations.
- For **Fashion MNIST**, Isomap produced **better-separated clusters than PCA but not as distinct as t-SNE**.

A key trade-off with Isomap is its sensitivity to **hyperparameter tuning (number of neighbors)**, and its computational cost, which is **higher than PCA but lower than t-SNE**.

4.6 Locally Linear Embedding (LLE)

LLE is another **non-linear manifold learning technique** that preserves **local relationships** between points. Unlike t-SNE, which models probability distributions, LLE reconstructs each point as a linear combination of its neighbors.

- LLE **worked well on the Breast Cancer dataset** but underperformed in Fashion MNIST due to its sensitivity to **noise and high-dimensional complexity**.
- One major drawback of LLE is that it tends to produce **distorted embeddings** when the dataset contains large variations in density.

While LLE is computationally more efficient than t-SNE, it often requires extensive tuning to obtain meaningful results.

4.7 Multi-Dimensional Scaling (MDS)

MDS seeks to **preserve pairwise distances** in lower dimensions. Unlike PCA, which is variance-based, MDS tries to maintain the structure of the original high-dimensional space.

- For **Breast Cancer**, MDS produced embeddings similar to PCA, leading to **strong classification performance**.
- For **Fashion MNIST**, MDS was **not as effective** as t-SNE or Isomap, as it struggles with complex class structures.

While MDS is useful for preserving global distances, it **does not emphasize local structures**, making it less effective for datasets with **intricate relationships**, such as Fashion MNIST.

4.8 Truncated Singular Value Decomposition (Truncated SVD)

Truncated SVD is closely related to PCA but is optimized for **sparse data matrices**. Instead of computing the full covariance matrix, it operates on the data directly.

- Truncated SVD performed **similarly to PCA** on the **Breast Cancer dataset**, achieving high classification accuracy.
- On **Fashion MNIST**, Truncated SVD **struggled to capture complex structures**, leading to reduced classification performance.

The main advantage of Truncated SVD is its **speed and scalability**, making it a good choice for **large-scale sparse datasets**.

4.9 Key Takeaways

The choice of dimensionality reduction method depends on **the dataset structure** and **intended use case**:

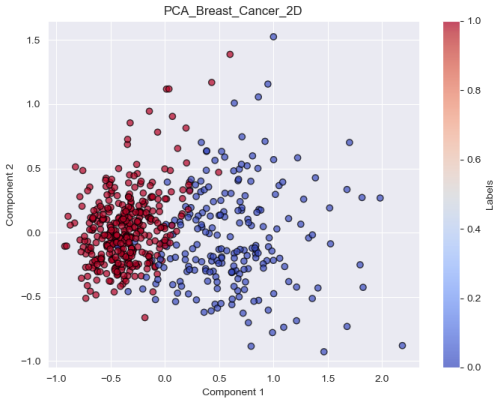
- **Linear Datasets (Breast Cancer)**:
  - **Best Methods:** PCA, Kernel PCA, MDS
  - **Why?** These methods preserve global variance and structure effectively.
- **Non-Linear Datasets (Fashion MNIST)**:
  - **Best Methods:** t-SNE, Isomap
  - **Why?** These methods better capture high-dimensional relationships.

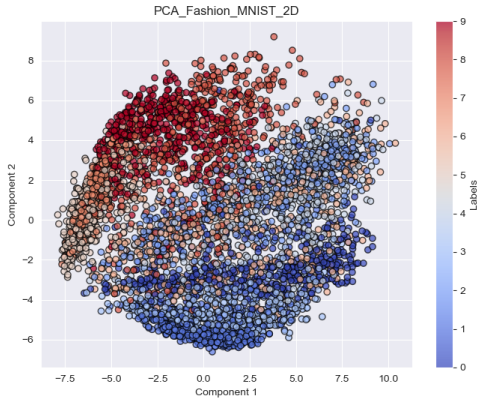
The **ultimate takeaway** is that **PCA is the best choice for structured, linearly separable datasets**, whereas **t-SNE and Isomap excel in handling complex, high-dimensional data with non-linear relationships**.

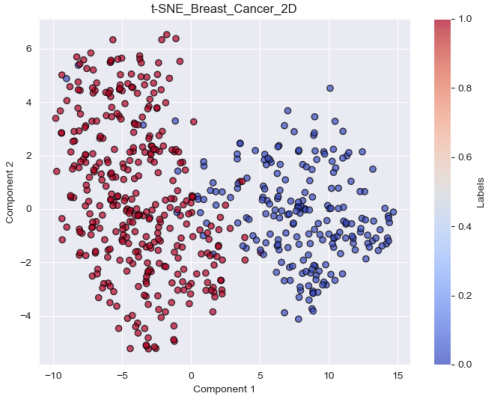
Selecting the right method involves balancing **accuracy, interpretability, and computational efficiency** based on dataset characteristics.

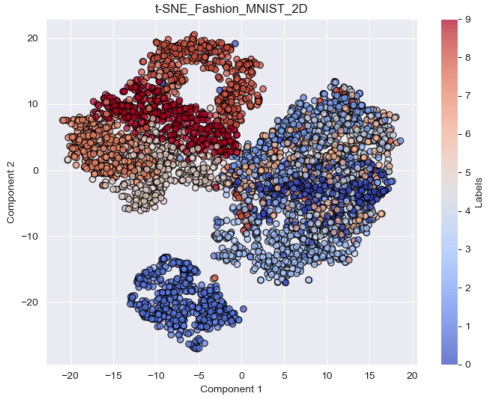
5 Conclusion

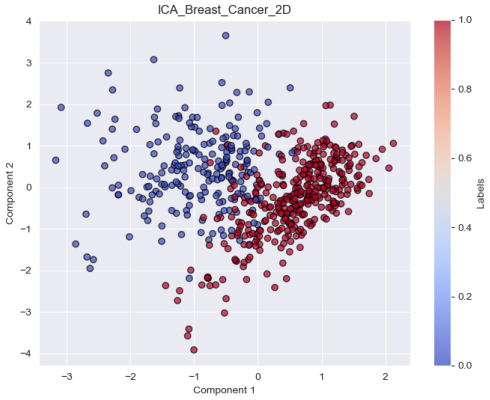
In this study, we analyzed multiple dimensionality reduction techniques applied to two datasets. PCA was highly effective for the Breast Cancer dataset due to its linear separability. However, t-SNE and Isomap provided better clustering for the Fashion MNIST dataset, as they capture non-linear relationships. The choice of technique depends on the structure of the dataset and the objectives of the analysis.

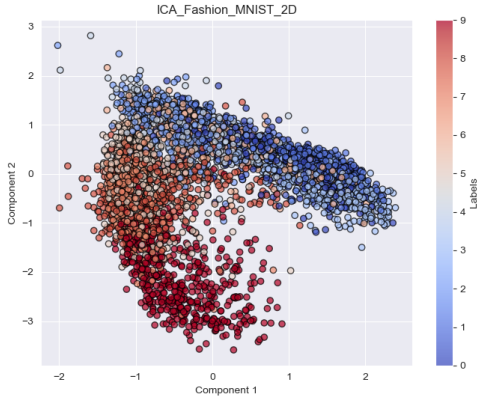

PCA_Breast_Cancer_2D
Component 1
Component 2
Labels

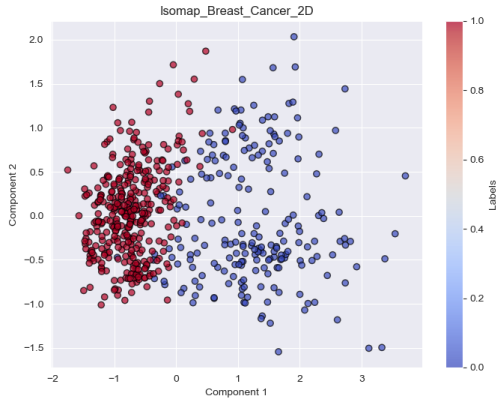

PCA_Fashion_MNIST_2D
Component 1
Component 2
Labels

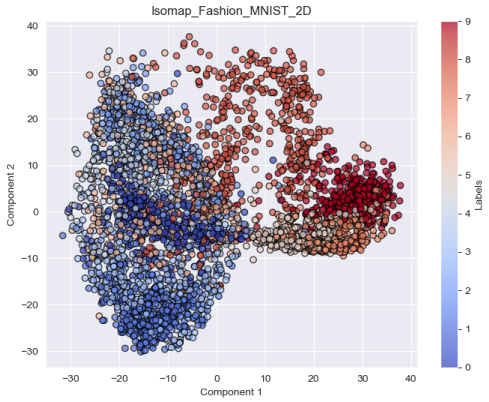

t-SNE_Breast_Cancer_2D
Component 1
Component 2
Labels

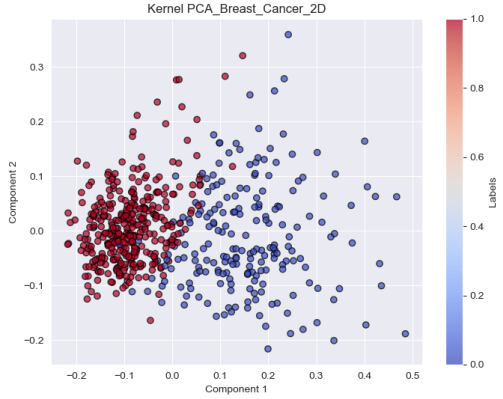

t-SNE_Fashion_MNIST_2D
Component 1
Component 2
Labels

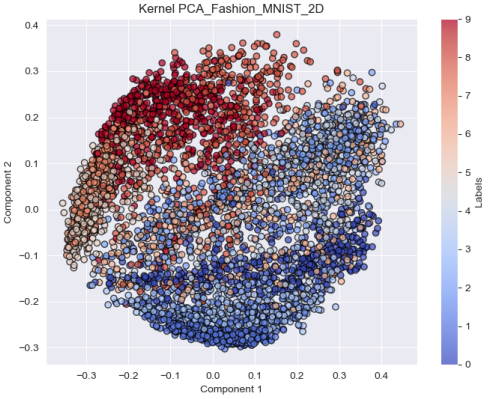

ICA_Breast_Cancer_2D
Component 1
Component 2
Labels

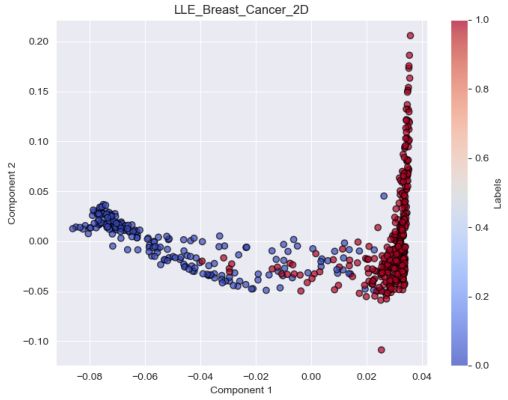

ICA_Fashion_MNIST_2D
Component 1
Component 2
Labels

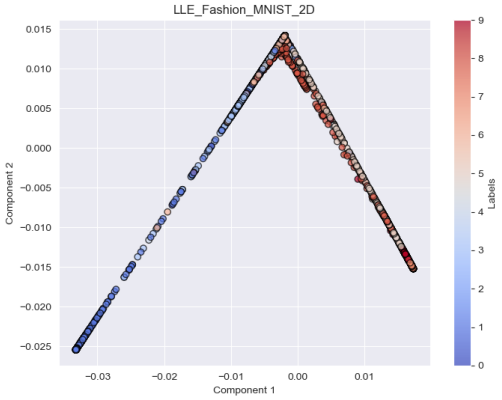

Isomap_Breast_Cancer_2D
Component 1
Component 2
Labels


Isomap_Fashion_MNIST_2D
Component 1
Component 2
Labels


Kernel PCA_Breast_Cancer_2D
Component 1
Component 2
Labels


Kernel PCA_Fashion_MNIST_2D
Component 1
Component 2
Labels


LLE_Breast_Cancer_2D
Component 1
Component 2
Labels


LLE_Fashion_MNIST_2D
Component 1
Component 2
Labels

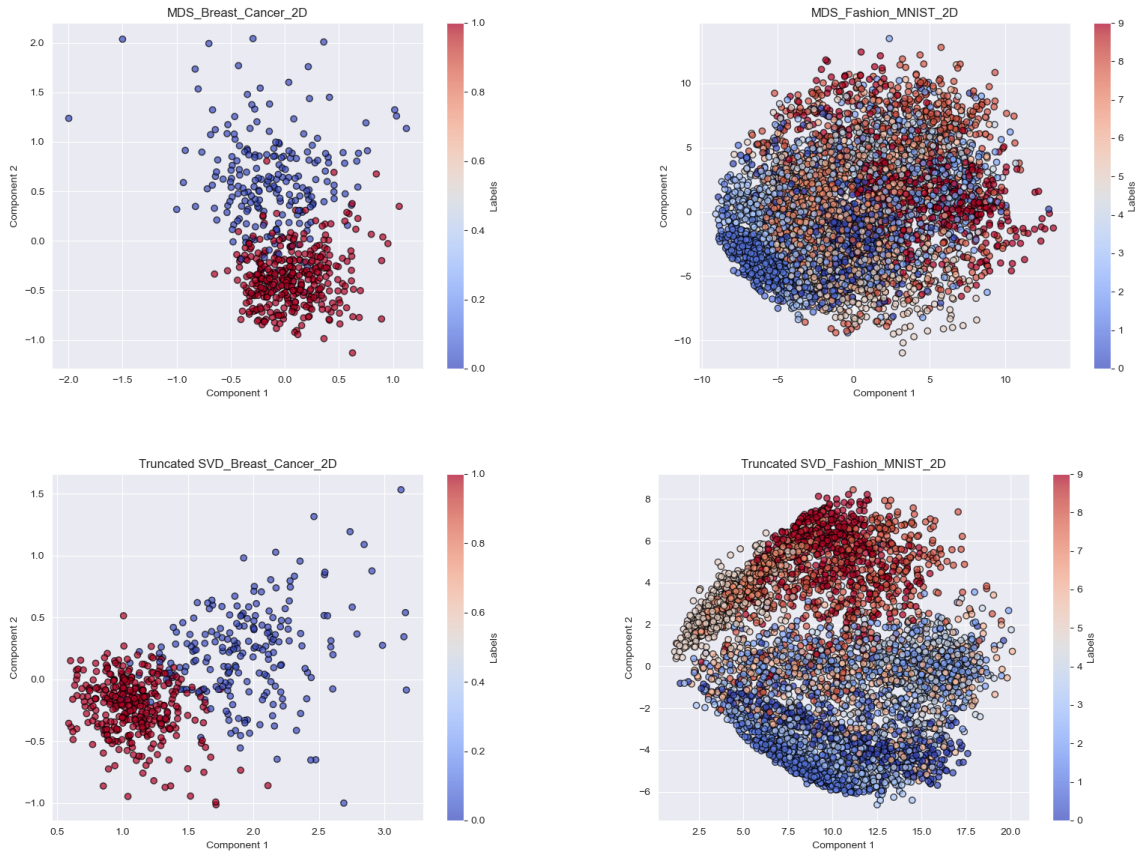
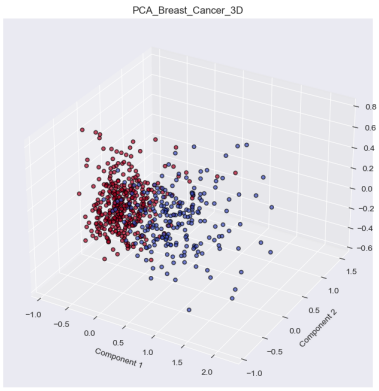
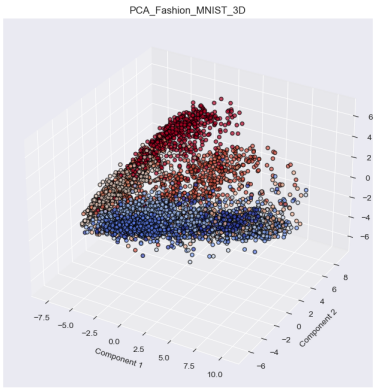
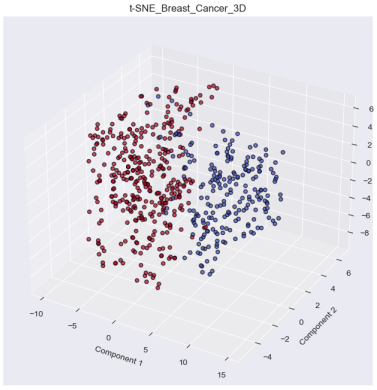
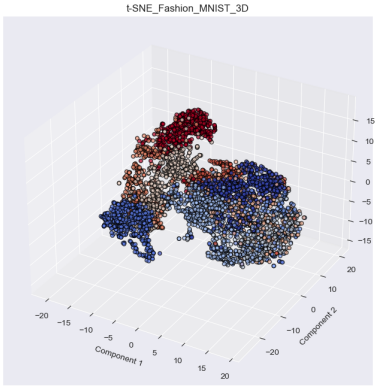


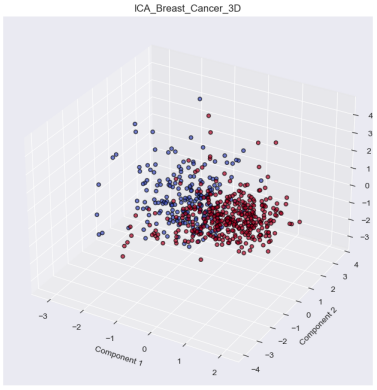
Figure 1: Comparison of some different dimensionality reduction techniques applied to Breast Cancer and Fashion MNIST datasets (2D representations).

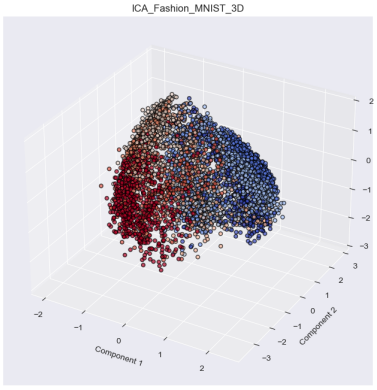

PCA_Breast_Cancer_3D
Component 1
Component 2
Component 3

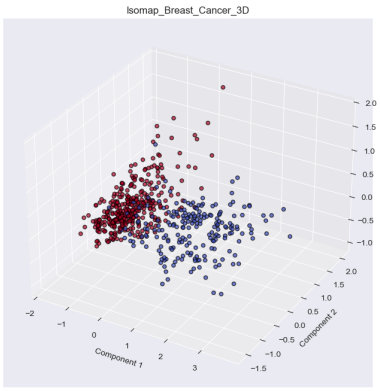

PCA_Fashion_MNIST_3D
Component 1
Component 2
Component 3

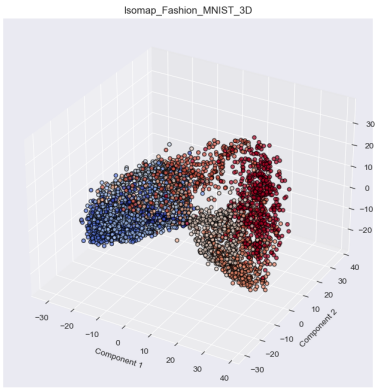

t-SNE_Breast_Cancer_3D
Component 1
Component 2
Component 3

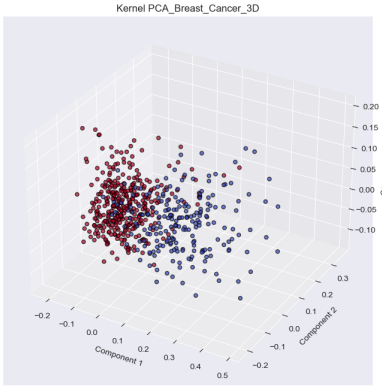

t-SNE_Fashion_MNIST_3D
Component 1
Component 2
Component 3

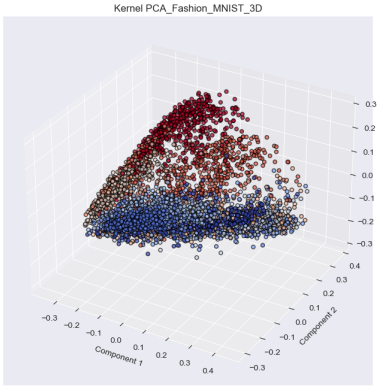

ICA_Breast_Cancer_3D
Component 1
Component 2
Component 3

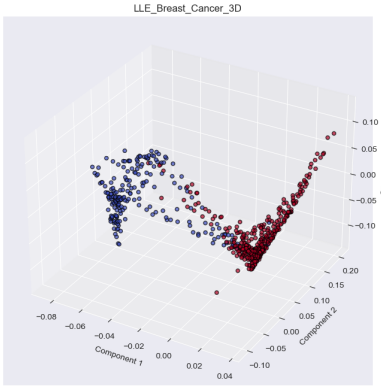

ICA_Fashion_MNIST_3D
Component 1
Component 2
Component 3

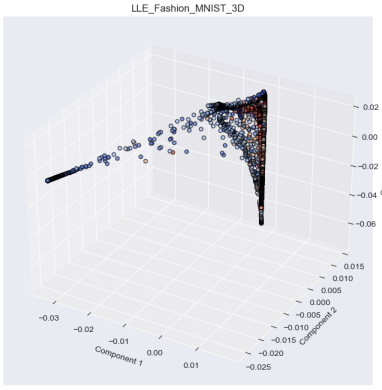

Isomap_Breast_Cancer_3D
Component 1
Component 2
Component 3


Isomap_Fashion_MNIST_3D
Component 1
Component 2
Component 3


Kernel PCA_Breast_Cancer_3D
Component 1
Component 2
Component 3


Kernel PCA_Fashion_MNIST_3D
Component 1
Component 2
Component 3


LLE_Breast_Cancer_3D
Component 1
Component 2
Component 3


LLE_Fashion_MNIST_3D
Component 1
Component 2
Component 3

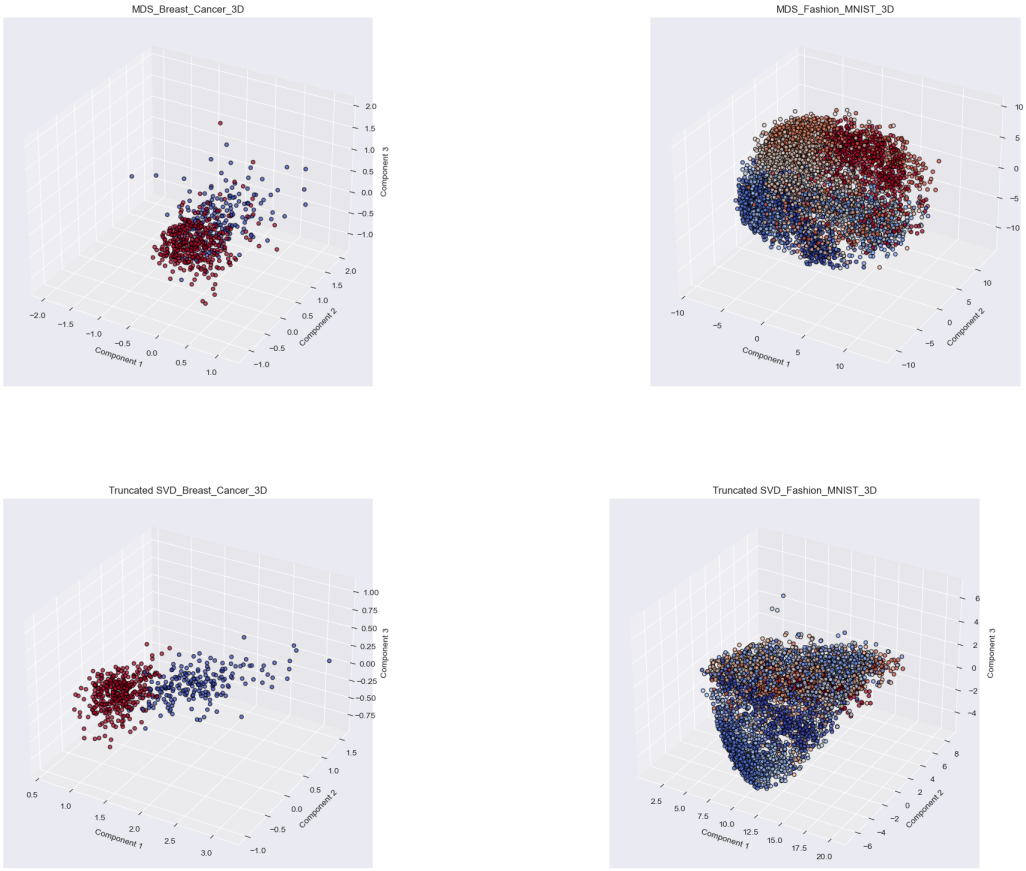


Figure 2: Comparison of some different dimensionality reduction techniques applied to Breast Cancer and Fashion MNIST datasets (3D representations).