WDZD Project

Miłosz Włoch, Maciej Ługowski, Kamil Szkoła

June 2022

Spis treści

1 Introduction

The topic of our project is visualization and analysis of the datasets for selected methods. The main goal of the project is to visualize datasets such as **FMNIST**, **RCV Reuters**, and **Smallnorb** using four selected methods that are:

- T-SNE
- UMAP
- PaCMAP
- IVHD

The work will also focus on describing results produced, understanding them, and analyzing measurements such as the execution time of each method. Ultimately, the goal would be to see which methods are the best at reducing the dimensionality of each set by using four chosen metrics. The project seems to be difficult and some problems may occur. Trying to anticipate some of them it was agreed that the biggest challenge would be visualizing such large datasets. It is possible that this reason will force the authors to shrink the data or make use of Prometheus cluster. Another problem may turn out to be the preparation of data for visualization due to the fact that most of the discussed datasets are not so widely analyzed. In addition, the implementation of the **IVHD** method may also prove to be problematic due to the fact that it is also unfamiliar.

2 Theory

2.1 Sets

1. **Fashion-MNIST** is a collection of 70,000 images of clothes in shades of grey. Each image is 28×28 pixels and belongs to one of ten classes. The collection was intended to serve as a direct replacement for the popular MNIST collection, which is widely used to analyze and test machine learning algorithms.

2. **Smallnorb** is a collection containing images of 50 toys belonging to 5 general categories: *four-legged animals, human figures, airplanes, trucks*, and *cars*. The objects were depicted by two cameras in 6 lighting conditions, 9 elevations (30 to 70 degrees every 5 degrees) and 18 azimuths (0 to 340 every 20 degrees). The collection is designed for experiments in shape-based 3D object recognition and originally contains 48,600 observations.

3. **Reuters** is a collection consisting of 804,409 news articles belonging to 90 classes. The set is used for document classification using machine learning algorithms.

2.2 Methods

1. **T-SNE** (T-Distributed Stochastic Neighbor Embedding) is a dimension reduction method based on t-distribution (t-Distributed Stochastic Neighbor Embedding). It is a nonlinear and unsupervised technique primarily used for multidimensional data mining and visualization. The **T-SNE** algorithm computes a similarity measure between pairs of points in both high and low dimensional space. It attempts to optimize these two similarity measures afterwards. This method handles local structure very well at the expense of visualizing global structure.

2. **UMAP** (Uniform Manifold Approximation and Projection) is a nonlinear dimension reduction method based on topological data analysis techniques that can be used to visualize high-dimensional data in a manner similar to that known, for example, from the **T-SNE** algorithm. We assume that the data is uniformly distributed on a locally consistent Riemannian manifold and that the Riemannian metric is (approximately) locally constant. UMAP then uses approximations of the local manifolds to combine their representations of fuzzy symplectic sets to form a topological representation of the high-dimensional data. Given low-dimensional representations of the data, it can also analogously try to construct an equivalent topological representation. UMAP tries to minimize the cross entropy between the topological representations obtained in this way, and thus seeks a low-dimensional representation that reflects the topological structure of the original (high-dimensional) data as well as possible. This method is characterized by higher speed and better preservation of global structure than T-SNE. T-SNE and UMAP are both highly stochastic and sensitive to parameter change methods, so significantly different results can be obtained after each change.

3. **IVHD** (Interactive Visualization of High-Dimensional Data Tool) is a tool used for interactive visualization of multidimensional data. It was invented to accelerate the visualization of large and multidimensional data sets. This algorithm clearly outperforms modern algorithms in terms of computational and memory complexity while maintaining high quality of the embedded data.

4. **PaCMAP** (Pairwise Controlled Manifold Approximation) is a dimensionality reduction method that can be used for visualization, preserving both local and global structure of the data in original space. Algorithms like T-SNE or UMAP focus on the local structure, TRIMAP on the global structure, but none of them focuses on both. They control their structures by matching parameters,

mainly by the number of neighbors considered. PaCMAP, on the other hand, is characterized by its ability to preserve both global and local structure by dynamically using three types of point pairs: neighbor pairs (pair_neighbors), center pairs (pair_MN), and far-distant pairs (pair_FP). PaCMAP executes faster than T-SNE and UMAP and performs well with visualization even for default parameters.

2.3 Metrics

1. **DR Quality** (Data Reduction Quality) is a quality measure that tries to maintain both global and local structure of original data. The produced output is a graph illustrating the dependence of $R_{NX}(K)$ measure on the K numbers of neighbours. Higher output value indicates on better quality.

2. **KNN Gain** (K-nearest neighbors Gain) - quality measure which calculations are based on the k-nearest neighbors classifier. The produced output is a graph illustrating the dependence of $G_{NN}(K)$ measure on the neighborhood size - K. The higher $G_{NN}(K)$ value means that the original data structure is better preserved.

3. **Trustworthiness** - the trustworthiness was proposed by Venna and Kaski, as a local quality measure of a low-dimensional representation. The metric focuses on the preservation of local neighborhoods, and compares the neighborhoods of points in the low-dimensional representation to those in the reference data. Hence, the trustworthiness measure indicates to which degree we can trust that the points placed closest to a given sample in the low-dimensional representation are really close to the sample also in the reference data set. N parameter defines the size of the neighborhoods to consider.

4. **Shepard Diagram** is a diagram that compares how far apart your data points are before and after you transform them (ie: goodness-of-fit) as a scatter plot. Shepard diagrams can be used for data reduction techniques like principal components analysis (PCA), multidimensional scaling (MDS), or t-SNE.

3 Visualizations

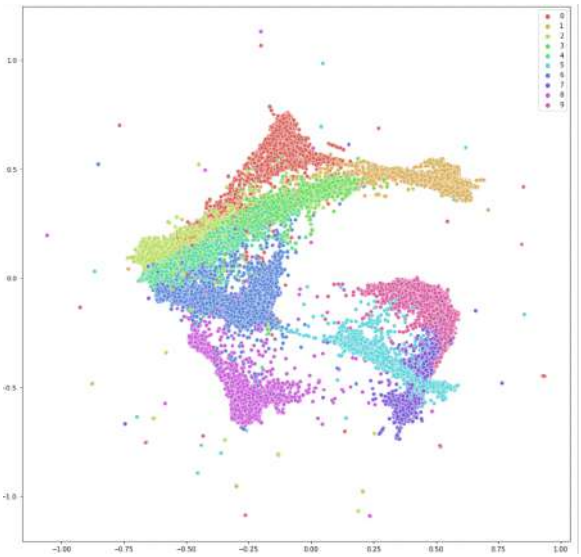
3.1 Time of embedding for chosen datasets and methods

Reduction Time		
Dimensionality reduction technique	Datasets	Time
T-SNE	FMNIST	17:45
	Reuters	32:56
	Smallnorb	8:03
UMAP	FMNIST	1:55
	Reuters	2:59
	Smallnorb	0:59
PaCMAP	FMNIST	0:52
	Reuters	1:02
	Smallnorb	0:29
IVHD	FMNIST	3:04
	Reuters	24:57
	Smallnorb	11:14

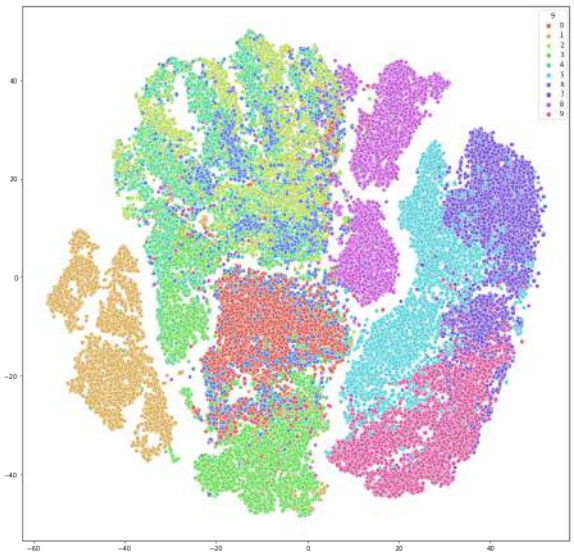
As it can be seen the embedding of Reuters dataset took the most time using each method. The reason behind it is the fact that Reuters is the biggest datased used. In fact it was shrinked to $\frac{1}{8}$ of its original size having 100 000 observations. Furthermore, the method which took the most time to be embeded was T-SNE as it is clearly shown in the table above.
On the other hand, the dataset which embeddings took the least time was smallnorb which was also shrinked having 24 301 observations which is merely $\frac{1}{2}$ of its original size. The fastest method turned out to be PaCMAP with each timing being less or equal to 1 minute.

3.2 FMNIST

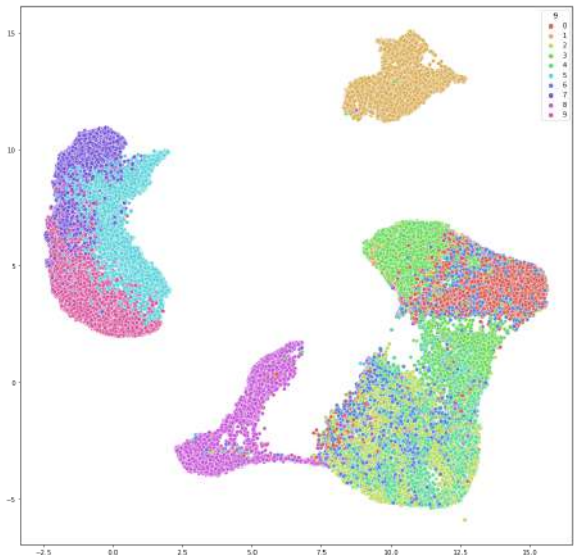
The visualization of the four selected methods for the dataset **FMNIST** is depicted below.



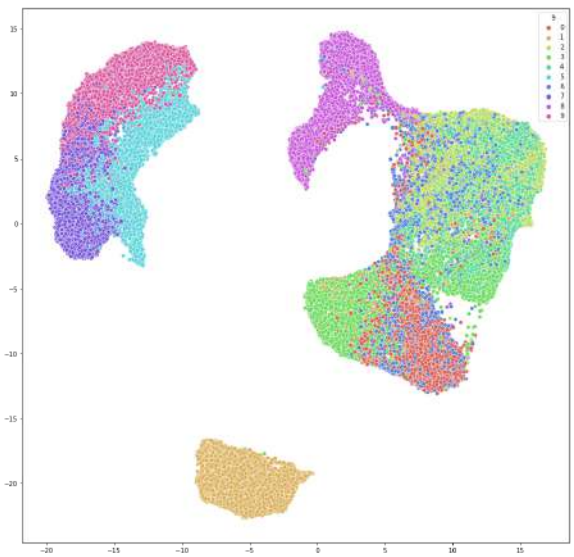
(a) IVHD method



(b) T-SNE method



(c) UMAP method



(d) PaCMAP method

(a) In the visualization of the set using the **IVHD** method, groups of similar points are visible as most classes are clustered together. Although, the distances between clusters are very low, which does not allow to accurately distinguish similar classes from each other. The method used preserved good local structure at the expense of global structure. It is visible that all clusters are concentrated in the middle. The Euclidean metric and the following parameters were chosen:

- iterations - 2500
- nearestNeighborsCount - 2

- randomNeighborsCount - 1
- binaryDistances - 0
- reverseNeighborsSteps - 0
- reverseNeighborsCount - 0
- l1Steps - 0

(b) Visualization of the **FMNIST** set using the **T-SNE** method did not separate the classes as expected. Only a few clusters of similar points are visible, where the rest of the groups blend together. The method also preserved a very good local structure at the expense of the global one, which is just shown by the small distances between points. In this case again similar clusters cannot be distinguished.

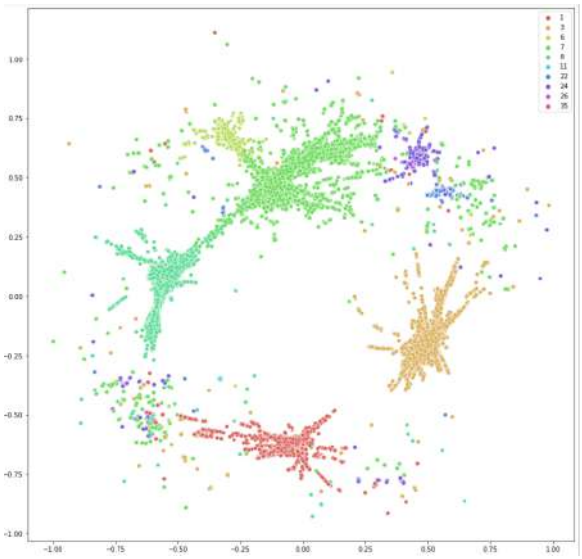
(c) **UMAP** is a method that groups labels globally, resulting in visible distances between clusters. The exceptions are similar groups, which are grouped in a local way, resulting in small distances in the graph. Its significant advantage is the speed of computation.

(d) The visualization using the **PaCMAP** method came out almost identical to the previous method. It's hard to see major differences in the distribution of points between the methods. However, this method is much faster.

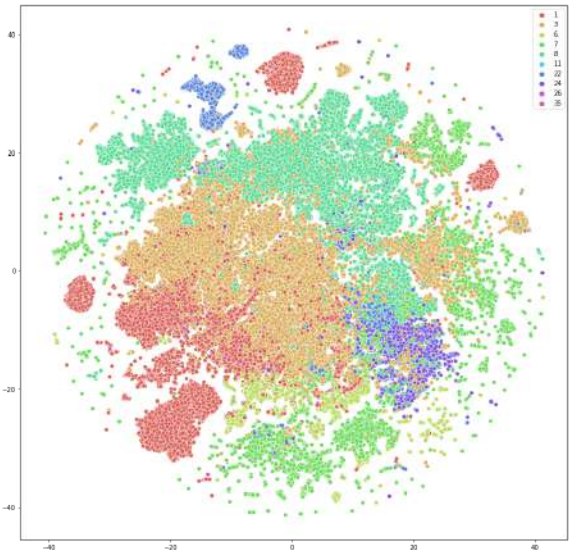
In the **FMNIST** set visualization problem, the **PaCMAP** method proved to be the best method, and it performed very well in separating the clusters with the shortest execution time.

3.3 Reuters

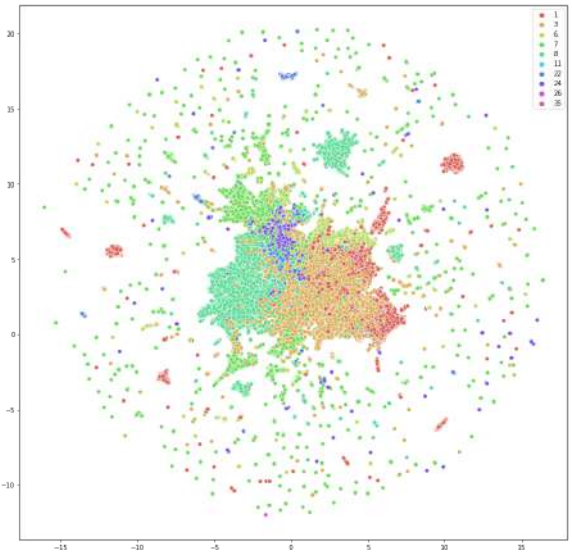
In case of Reuters only 10 most occuring classes were chosen for visualizations as the rest of them existed but in small numbers. The visualization of the four selected methods for the dataset Reuters is depicted below.



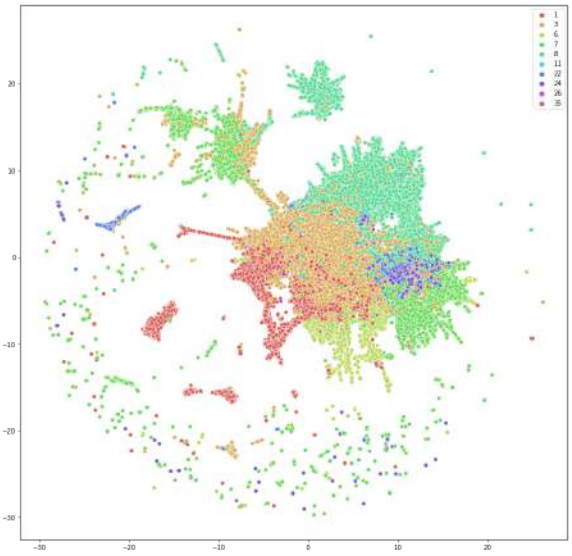
(a) IVHD method



(b) T-SNE method



(c) UMAP method



(d) PaCMAP method

(a) In visualizing the set using the **IVHD** method, groups of similar points are visible, as that most points of a class are clustered together, with a few exceptions. Clusters are scattered and relatively many outlier points are visible in the graph. It is not possible to infer which groups of points are similar relative to each other. The method used preserved both good local and global structure. The Euclidean metric and the following parameters were chosen:

- iterations - 5000
- nearestNeighborsCount - 5
- randomNeighborsCount - 1
- binaryDistances - 1
- reverseNeighborsSteps - 0
- reverseNeighborsCount - 0
- l1Steps - 0

(b) The visualization of the **Reuters** set using the **T-SNE** method failed to separate the classes only for the dark green, blue and purple labels. Most of the points are scattered all over the space, Similar clusters cannot be distinguished as the points mix with each other.

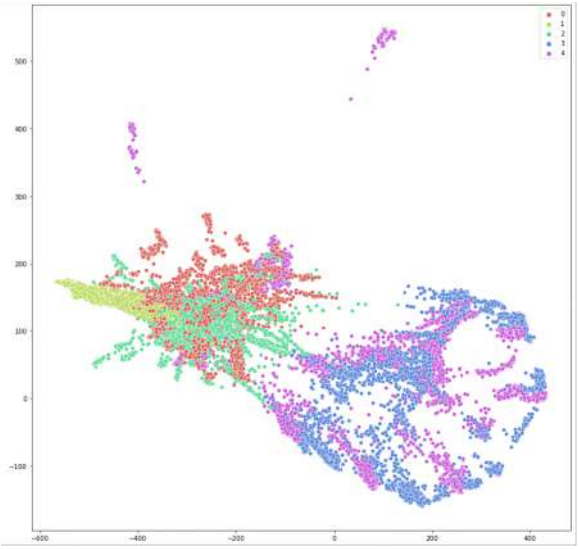
(c) **UMAP** definitely did not perform well when visualizing the **Reuters** set. The center of the graph shows a cluster of all observations with slightly visible groups of similar points at the edges. The points are scattered throughout the space regardless of the label they have.

(d) The visualization using the **PaCMAP** method came out very similar to the previous method. However, a better fit of the points to the clusters can be seen, i.e. fewer outliers are visible. Several clusters containing given labels can be relatively specified.

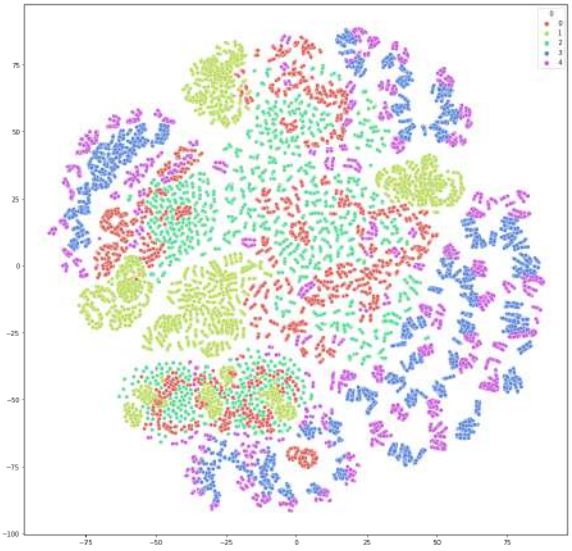
The dimension reduction method with the best class separation performance turned out to be the **IVHD** method, which was the only one that was up to the task of visualizing the processed dataset.

3.4 Smallnorb

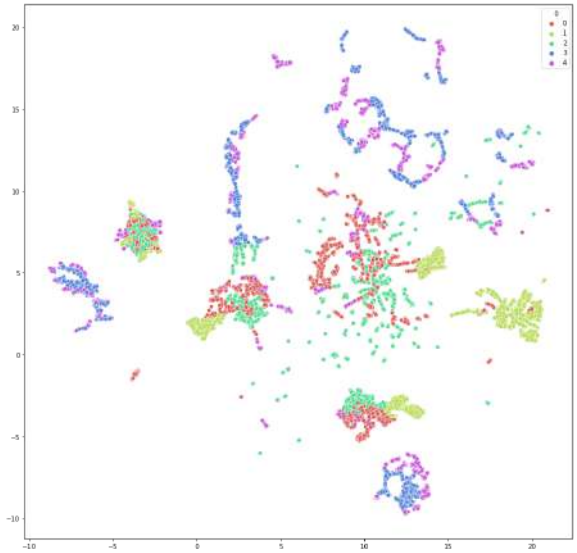
The **Smallnorb** dataset was shrinked to $\frac{1}{2}$ of its original size with 24 301 observations remaining. The visualization of the four selected methods for the dataset is depicted below.



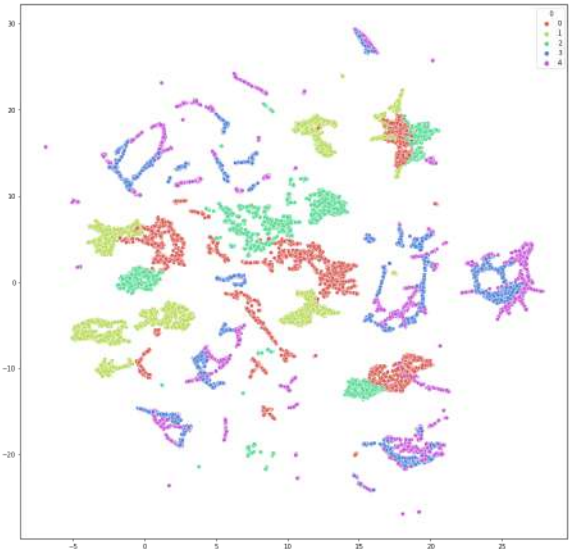
(a) IVHD method



(b) T-SNE method



(c) UMAP method



(d) PaCMAP method

(a) The **IVHD** method did not do well with the **Smallnorb** set. One can distinguish three clusters converging in one area, although the fact that they partially overlap is not good. The blue and pink groups are spread over the space which interferes with the analysis of the set. The method tries to preserve the local structure, not coping with the global separation of classes. The cosine metric and the following parameters were chosen:

- iterations - 6000
- nearestNeighborsCount - 20
- randomNeighborsCount - 1
- binaryDistances - 1
- reverseNeighborsSteps - 0
- reverseNeighborsCount - 0
- l1Steps - 0

(b) Visualizing the **Smallnorb** set using the **T-SNE** method did not separate the clusters as expected. It can be seen that points are scattered all over the graph mixing with each other. Representatives of one group can be seen in every part of the graph. The label with the color purple can be used as an example. It can be inferred that both the cluster with purple and red with dark green are similar groups.

(c) The **UMAP** reduced set using default parameters didn't manage to separate clusters. It tries to preserve local and global structure, unfortunately with mediocre results. To sum up, it is hard to read something sensible from the visualized set, you can only guess.

(d) The **PaCMAP** dimensionality reduction method, despite its speed, was once again not up to the task of finding clusters of similar points. The method tries to preserve both local and global structure, which results in a large dispersion of clusters of observations in the graph. It is a difficult task to make conclusions based on the studied graph. Another hypothesis is that the purple label and the blue label are similar classes, as they mostly mix with each other.

From the presented graphs it can be concluded that the best visualization of **Smallnorb** set was achieved by **T-SNE** method characterized by very good preservation of local structure. The problem of reducing the processed set is its high dimensionality which caused difficulties for the used methods.

4 Metrics

This section addresses the examination of measures of harvest quality for each of the selected dimension reduction methods.

4.1 DR Quality

DR Quality graphs for all three sets are shown below

4.1.1 FMNIST

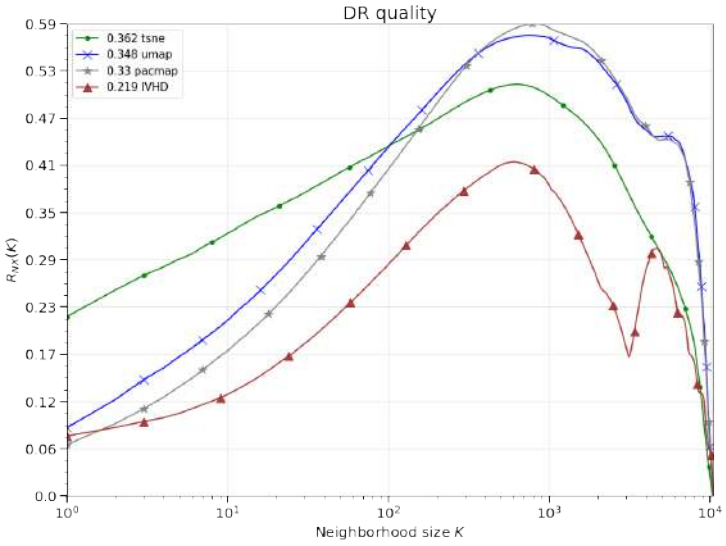


Fig. 4: Dr quality for FMNIST

From the graph it can be read that for a small number of points the **T-SNE** method performs best, where the values of the other methods are very close. As the number of analyzed neighbors increases, an almost logarithmic increase in the coefficient $R_{NX}(K)$ of the **PaCMAP** and **UMAP** methods can be observed. The peak is reached for most methods in the neighborhood of nearly a thousand. Then, the method with the highest quality turns out to be **PaCMAP**.

4.1.2 Reuters

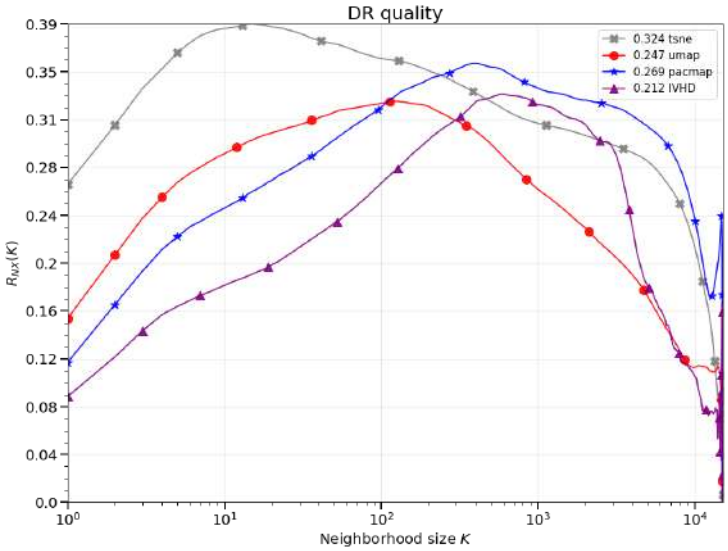


Fig. 5: Dr quality for Reuters

It can be seen that in the presented graph for a low number of points the **T-SNE** method performs best, where the values of the other methods are very close. As the number of analyzed neighbors increases, the value of $R_{NX}(K)$ also increases. The peak is reached for most of the methods in neighborhoods equal to nearly a thousand, except for the T-SNE method where the highest value of the coefficient is reached for 10 neighbors.

4.1.3 Smallnorb

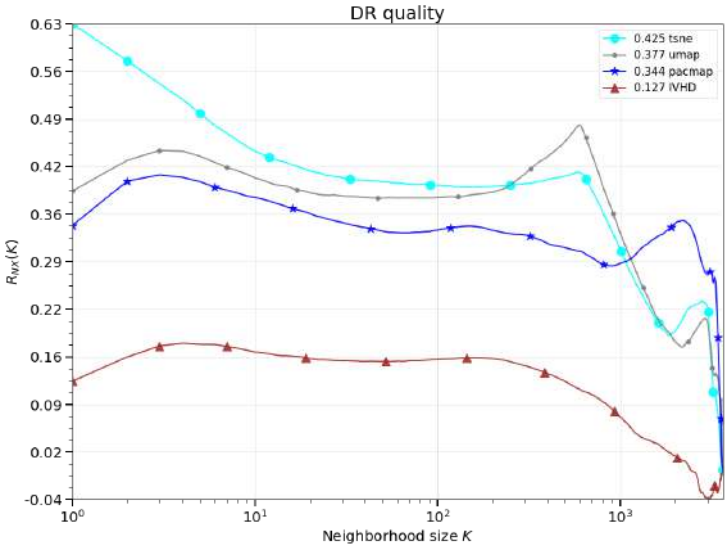


Fig. 6: Dr quality for Smallnorb

The graph presenting the quality measure for the **Smallnorb** set shows the superiority of the **T-SNE** method over the others when considering a small number of neighbors. For the examined neighborhood close to a thousand, the highest value is achieved for **UMAP** while for the number of neighbors much

larger than a thousand, **PaCMAP** turned out to be the best method. The **IVHD** method stands apart from the others, being characterized by relatively small values of the coefficient $R_{NX}(K)$ for any size of the analyzed neighborhood.

4.2 KNN Gain

KNN Gain graphs for all three sets are shown below

4.2.1 FMNIST

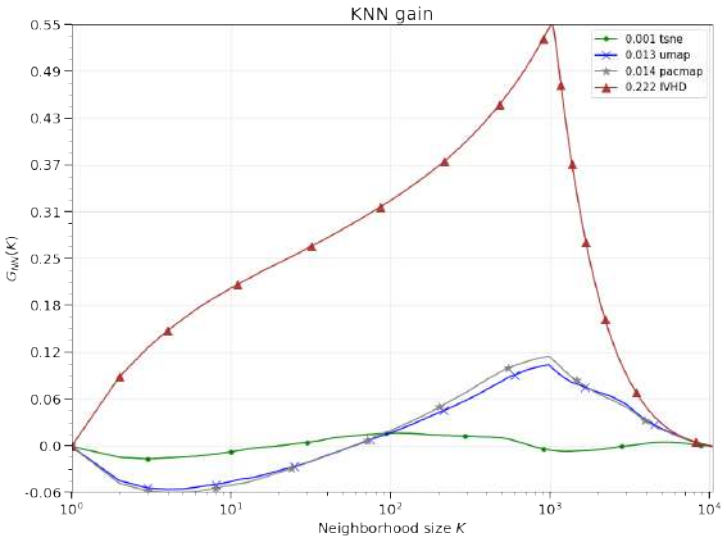


Fig. 7: KNN Gain for FMNIST

In the graph it can be see that the **IVHD** method outclasses the others for any number of neighbors. This is because **IVHD** takes high values of the coefficient $G_{NN}(K)$ regardless of the number of neighbors compared to the other methods which means that this method best represents the original data. The coefficient values of $G_{NN}(K)$ increase as the number of neighbors analyzed increases for all methods except **T-SNE** and reach a peak for the number of neighbors equal to one thousand. The **T-SNE** method for any large neighborhood reaches a fairly constant value. However, for a small number of neighbors it performs better than **UMAP** or **PaCMAP** for a neighborhood of 100.

4.2.2 Reuters

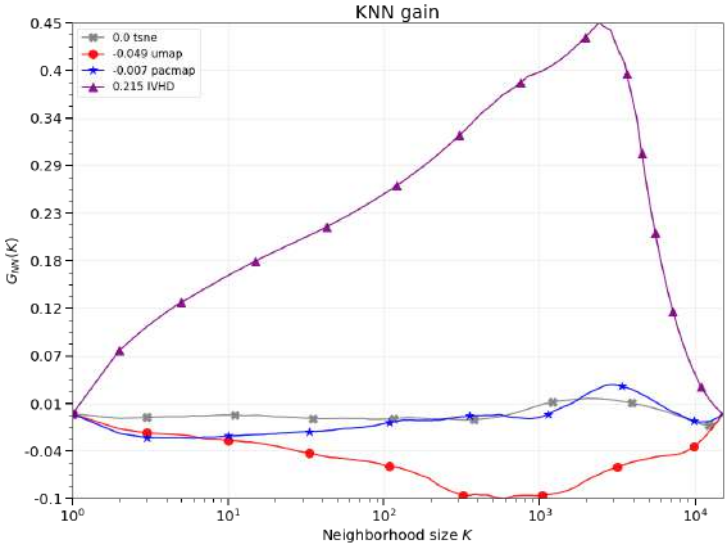


Fig. 8: KNN Gain for Reuters

For KNN gain for the set **Reuters**, the method **IVHD** similarly to the previous set outclassed the others for any number of neighbors. This is because **IVHD** takes high values of the coefficient $G_{NN}(K)$ regardless of the number of neighbors compared to the other methods. The values increase as the number of neighbors analyzed increases for all methods except **UMAP** and reach a peak for the number of neighbors close to one thousand. The **T-SNE** and **PaCMAP** methods for arbitrarily large neighborhoods reach a fairly constant value. **UMAP** reaches the lowest average coefficient value which determines its poor ability to represent the original data.

4.2.3 Smallnorb

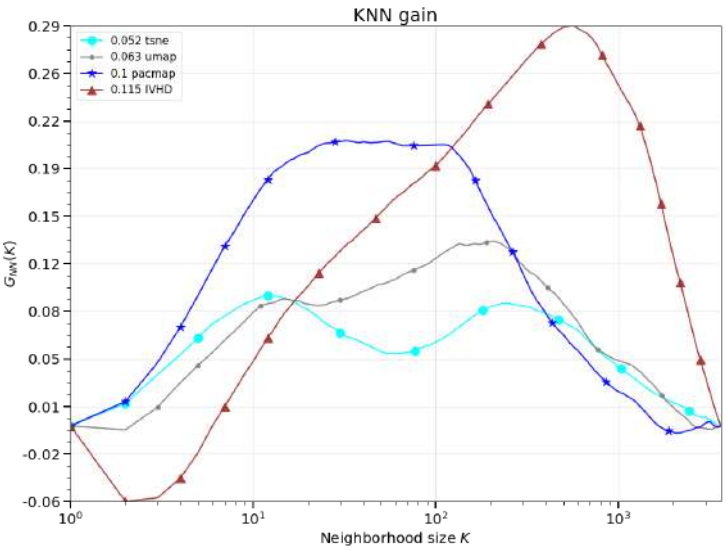


Fig. 9: KNN Gain for Smallnorb

It can be read from the above graph that the method achieving the highest values of the coefficient $G_{NN}(K)$ in a small neighborhood turned out to be **PaCMAP**. With the increase in the number of analyzed neighbors, these values significantly decreased indicating the deterioration of the quality of the tested method. For large neighborhoods the best method is **IVHD**. The other methods achieve similar values, where the better one is **T-SNE** for a small number of neighbors, and for a large number of neighbors the better method is **UMAP**.

4.3 Trustworthiness

The trustworthiness quality measure allows to choose the method which best preserves the original data structure. That means the values closest to one indicate on the higher chance of retaining local structure. The trustworthiness analysis focuses on calculating trust given two metrics - cosine and Euclidean - and the number of neighbors analyzed equal to 5, 100 and 500

4.3.1 FMNIST

FMNIST trustworthiness				
Method	Metric	n=5	n=100	n=500
T-SNE	euclidean	0.992	0.973	0.943
	cosine	0.972	0.933	0.892
UMAP	euclidean	0.98	0.971	0.941
	cosine	0.948	0.934	0.908
PaCMAP	euclidean	0.975	0.969	0.931
	cosine	0.941	0.932	0.909
IVHD	euclidean	0.499	0.504	0.525
	cosine	0.498	0.504	0.524

4.3.2 Reuters

Reuters trustworthiness				
Method	Metric	n=5	n=100	n=500
T-SNE	euclidean	0.988	0.917	0.829
	cosine	0.992	0.94	0.867
UMAP	euclidean	0.96	0.912	0.822
	cosine	0.972	0.937	0.864
PaCMAP	euclidean	0.939	0.905	0.826
	cosine	0.953	0.929	0.87
IVHD	euclidean	0.5	0.504	0.523
	cosine	0.499	0.504	0.523

4.3.3 Smallnorb

Smallnorb trustworthiness				
Method	Metric	n=5	n=100	n=500
T-SNE	euclidean	0.993	0.956	0.921
	cosine	0.97	0.832	0.701
UMAP	euclidean	0.98	0.956	0.92
	cosine	0.922	0.827	0.688
PaCMAP	euclidean	0.974	0.951	0.89
	cosine	0.904	0.801	0.64
IVHD	euclidean	0.498	0.505	0.522
	cosine	0.502	0.505	0.524

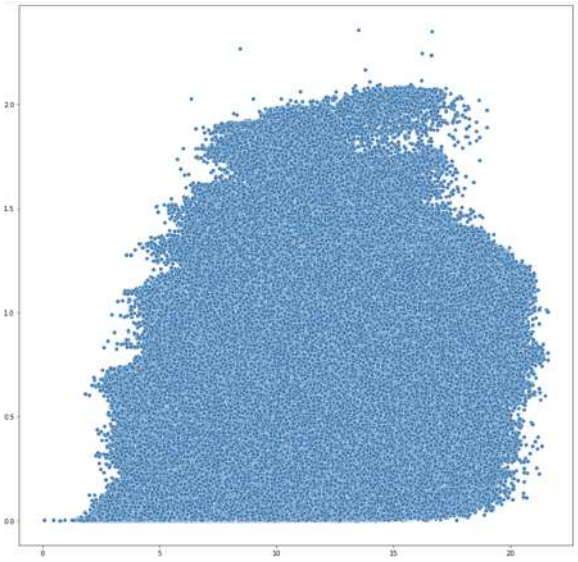
The above tables show the trustworthiness results of the sets **FMNIST**, **Reuters**, **Smallnorb** for 5, 100 and 500 neighbors. By analyzing the values, it can be deduced that for small (n=5), medium (n=100)

and large (n=500) number of neighbors only the **IVHD** method fails. For the Euclidean and cosine metrics the trustworthiness values decrease with larger numbers of neighbors, again the only exception is **IVHD** method as they slightly increase. The best performing method was **T-SNE**.

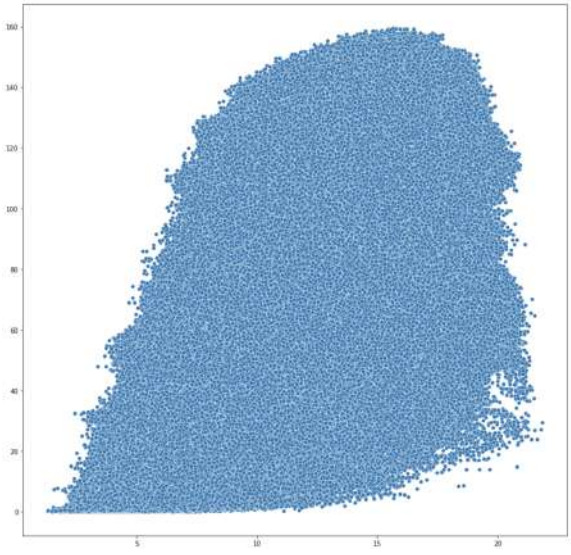
4.4 Shepard Diagrams

Below are shown Shepard diagrams for each set considering dimensionality reduction using the selected methods. This approach will relatively determine for which method the original data structure was best preserved. The X-axis shows the distances between points from the original dataset, and the Y-axis shows the distances between points in the embedded sets. Ideally, dimensionality reduction would result in points forming a diagonal graph, which would mean that as the distance between the original data increases, the distance between the reduced data also increases proportionally. However, this situation rarely exists because the data always loses some portion of its information after the reduction.

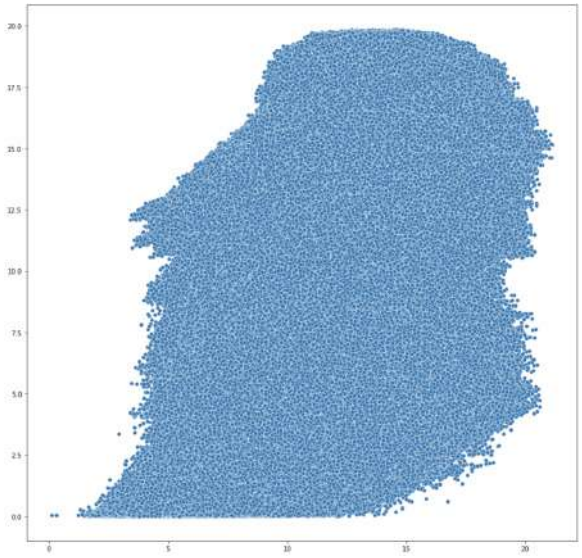
4.4.1 FMNIST



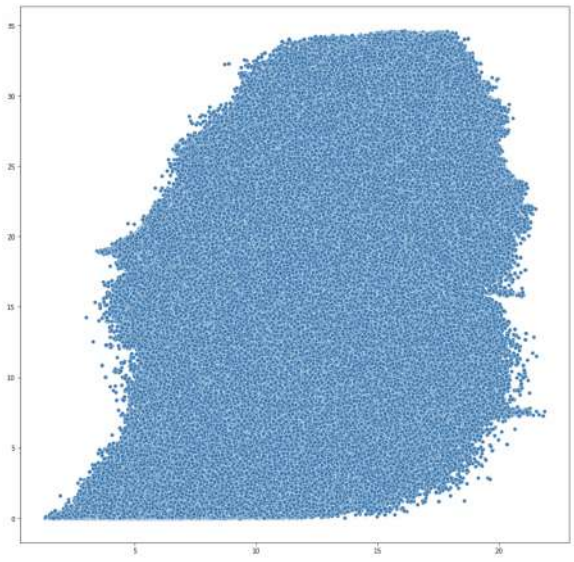
(a) IVHD method



(b) T-SNE method



(c) UMAP method

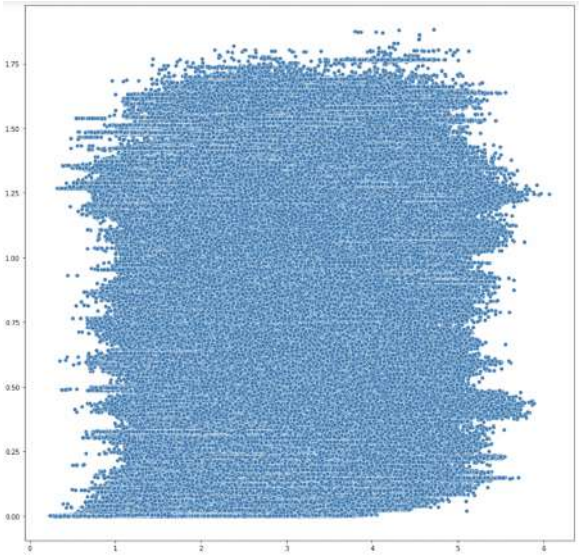


(d) PaCMAP method

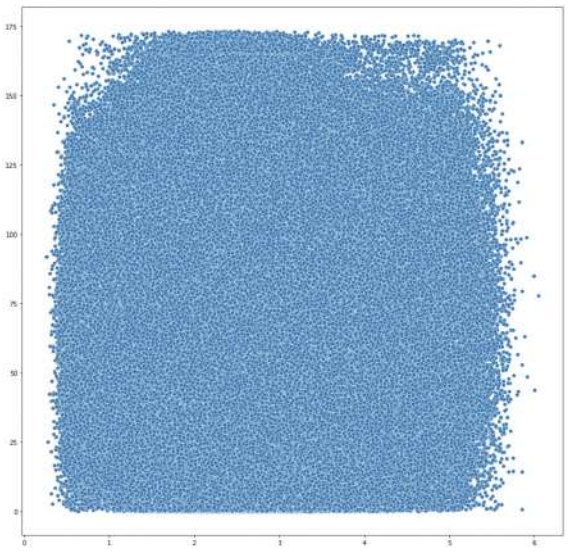
It can be read from Shepard diagram, that all methods produce similar results for the **FMNIST** set. In each of the diagrams, a figure close to the expected one in which the points form a diagonal line can be seen. There is additionally clear noise, which shows up as points far from the diagonal, signifying that

the reduced sets do not retain some portion of the original information. It can be deduced from the above graphs that the worst preserved structure is observed for the **IVHD** method, for which the points form a more flattened graph, where the distances in the embedded data do not exceed a certain threshold.

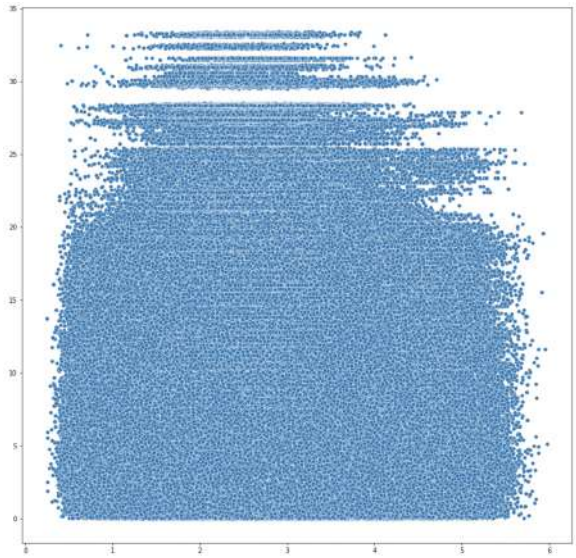
4.4.2 Reuters



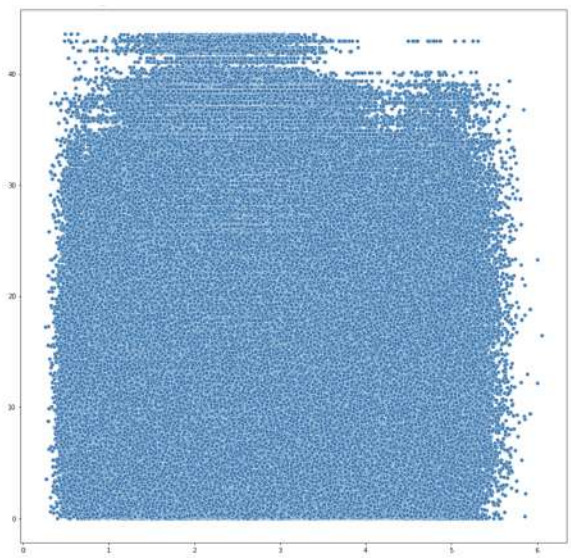
(a) IVHD method



(b) T-SNE method



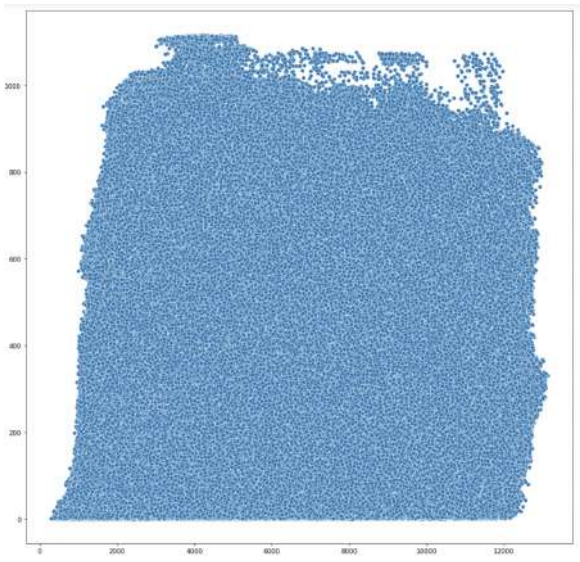
(c) UMAP method



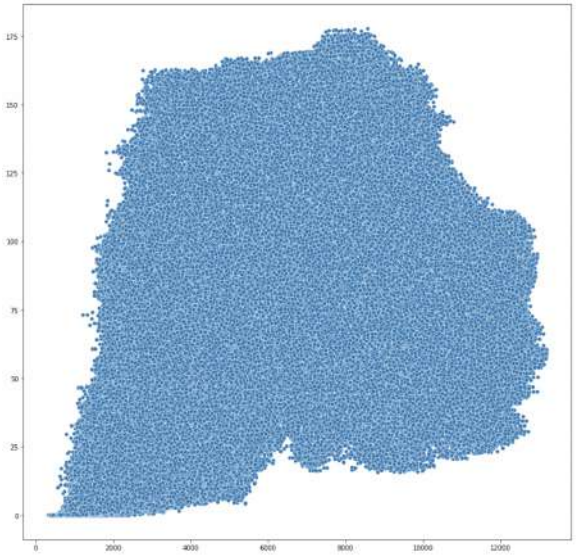
(d) PaCMAP method

For the set **Reuters** one can notice much worse results than those obtained for the previous dataset. In a comparison of all the methods used only **IVHD**, to some extent, attempts to replicate the original data structure. There are noticeable less information losses than the rest. The points in all dimensionionality reduction methods are arranged in a shape similar to a rectangle instead of forming a diagonal.

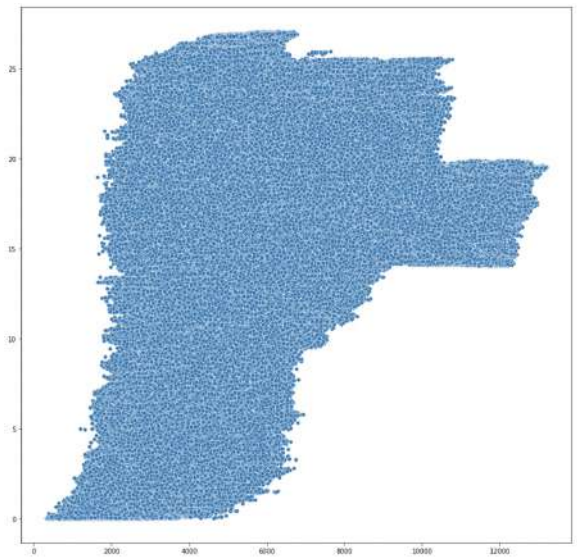
4.4.3 Smallnorb



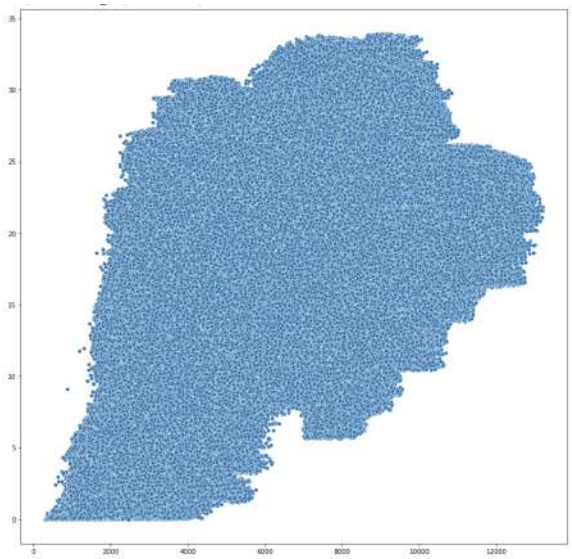
(a) IVHD method



(b) T-SNE method



(c) UMAP method



(d) PaCMAP method

Comparing the diagrams of all the methods performed for the set **Smallnorb**, it can be seen that by far the best at preserving the original data structure was **UMAP**. The only other method that aspires to similar information retention is **PaCMAP**. The other two methods do not cope with the processed set, although **T-SNE** shows clearly better results than the last method used, that is, **IVHD**.