# Performance Evaluation for Different Dimension Reduction Techniques for High-Dimensional Data

***Authors**: Kautilya Kondragunta, Sohan Birajdar, Shreyas Lokesha*

## Abstract:

To ensure that the high-dimensional data visualizations produced are accurate and informative, it is a difficult task that calls for careful consideration of various quality metrics. In this study, we look at how quality metrics are used in the context of gene expression analysis, a popular use for high-dimensional data visualization. We suggest a data transformation algorithm that employs quality metrics to direct feature selection and enhance the readability and interpretability of the resulting visualizations. The Gene Expression Omnibus (GEO) dataset, which contains gene expression profiles for various diseases and conditions, is used to assess the algorithm. Our findings show that the suggested algorithm can successfully decrease the dimensionality of the data while maintaining its valuable characteristics, resulting in better visualization quality and interpretability. We also go over the shortcomings of our strategy and make recommendations for future lines of inquiry in the area of high-dimensional data visualization.

## Introduction:

A fundamental method for visualizing and analyzing high-dimensional data is dimensionality reduction. Gaining insights into the underlying patterns and relationships between the data points becomes more challenging as the dimensionality of the data rises. By reducing the dimensionality of the data while maintaining its useful features, dimensionality reduction techniques like principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) can help solve this problem.

However, choosing the best dimension reduction technique can be difficult because each technique has different advantages and disadvantages. To overcome this problem, we suggest an evaluation framework that rates the effectiveness of various dimension reduction methods using a variety of quality indicators, such as readability, interpretability, scalability, and robustness.

The three commonly used dimension reduction methods—PCA, t-SNE, and UMAP—are the main topics of this paper. Each of these techniques is subjected to our evaluation framework, and we show how the performance of these techniques can be better understood through the visualization of smaller datasets. To evaluate the efficacy of each technique, we specifically calculate the quality metrics before and after reduction and compare the outcomes based on a set of objective quality metrics. Our evaluation framework can assist researchers and practitioners in choosing the best dimension-reduction technique for their unique needs.

We can better understand the advantages and disadvantages of each technique by comparing how well they perform in terms of readability, interpretability, scalability, and robustness. This will also help us to determine which technique is best for a given application. The remainder of the paper includes a discussion of the implications of our findings, the details of our evaluation framework, and the outcomes of our experiments. We also highlight the limitations of our study and suggest areas for future research by contrasting our methodology with related work in the field of dimensionality reduction.

The k-Nearest Neighbors (k-NN) algorithm is a simple, yet effective, method for classification and regression. It operates by considering the 'k' closest instances to a given data point and predicting the class based on the majority class label among those neighbors. By comparing the F1 score, precision, and recall of the k-NN model before and after dimensionality reduction, you can establish the benefits of using such techniques in terms of prediction quality. If the metrics improve after dimensionality reduction, it indicates that the reduced dataset retains the essential information while reducing noise, ultimately leading to better classification performance.

The algorithm will be implemented for evaluating classification metrics using k-Nearest Neighbors (k-NN) and to calculate the F1 score of the k-NN model to establish the benefits of using the dimensionality reduction techniques, which will, in turn, determine the prediction accuracy and quality before and after dimensionality reduction.

One way to evaluate the effectiveness of PCA, t-SNE, and U-Map is by using a measure called Rotating Variance Measure (RVM). RVM is a measure that quantifies the benefits of a PCA, t-SNE, and U-Map transformation by considering the ratio of the variances of the transformed variables. The RVM value ranges from 0 to 1, and a higher RVM indicates that the PCA, t-SNE, and U-Map transformation is more effective at capturing the variance in the original dataset. If RVM is close to 1, it means that the PCA, t-SNE, and U-Map transformations have effectively summarized the variance in the original dimensions using the first two principal components. On the other hand, if RVM is close to 0, it indicates that the PCA, t-SNE, and U-Map transformations haven't captured much variance, and the transformed dimensions may not provide useful information for further analysis or modeling.

In today's data-driven world, effective visualization and interaction techniques are crucial for understanding complex, high-dimensional datasets. Abstraction and Elaboration of visual elements, toggling options to switch between dimensionality reduction approaches, and Instance Selection for viewing detailed information about individual data points. Abstraction and Elaboration enable users to focus on the most relevant visual elements while reducing the cognitive load of interpreting complex visualizations. Toggling options to switch between dimensionality reduction approaches empowers users to compare and evaluate different dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) within the same visualization environment. Instance Selection interaction allows users to examine individual data points within clusters more closely.

By incorporating these interaction techniques, a more intuitive and comprehensive exploration of high-dimensional datasets, these techniques help in enhanced understanding, and contextual insights, help to increase user engagement, encouraging collaboration amongst the team, ultimately enabling users to derive meaningful insights and make informed decisions based on their analysis.

## Related work:

We go over various methods for data classification, dimensionality reduction, and visualization techniques that have been researched in the literature in the section on related work. In order to classify data, Mustakim et al. [1] compared the K-Nearest Neighbor (KNN) algorithm and a modified version, showcasing the modified KNN's superior performance. On the other hand, our project makes use of a variety of dimensionality reduction strategies to produce a classification model that is more reliable and accurate.

Cui et al. [3] proposed data abstraction quality metrics to evaluate how well the abstraction represents the original dataset, while Tatu et al. [2] concentrated on visual analytics tasks and suggested various scatterplot quality measures. Our method, on the other hand, combines various dimensionality reduction strategies to produce a more thorough visualization of high-dimensional data, enabling better comprehension and analysis. Radial visualization (Radviz) was described by Long [4] as a popular technique for multivariate data visualization, and Liu et al. [5] contrasted the most widely used conventional dimensionality reduction techniques. In order to better perform than individual approaches in visualizing high-dimensional data, our project aims to offer a more effective and precise ensemble of dimensionality reduction techniques.

While Johansson and Johansson [7] presented a system for dimensionality reduction that preserves as many important structural properties as possible, Feng and Wang [6] compared PCA and LDA dimensionality reduction algorithms based on the wine dataset. In our strategy, we combine the strengths of various dimensionality reduction techniques in addition to comparing them to produce a potent visualization tool. The idea of Polynomial Principal Component Analysis (Poly-PCA) as a nonlinear dimensionality reduction technique was first introduced by Kazemipour and Druckmann [8]. In our project, dimensionality reduction techniques that are both linear and nonlinear are combined to produce a more flexible visualization system that can handle large datasets.

The quality metrics employed in data visualization analysis and their applications have been covered in a number of papers [9], along with empirical recommendations on the selection of scatterplot and dimension reduction techniques [10]. In order to improve the classification and visualization of high-dimensional data, our project makes use of this knowledge to construct an ideal ensemble of dimensionality reduction techniques. The performance of well-liked dimensionality reduction techniques like PCA, t-SNE, and UMAP was examined by Lewis et al. [11] while Strehl [12] proposed a strategy centered on identifying significant patterns in high-dimensional data. In order to create a better method for handling and visualizing complex datasets, these insights are combined in our project.

The use of pairwise relationships between variables in high-dimensional visual analytics and the significance of interaction in information visualization were both highlighted by Yi et al. [13] and Wilkinson et al. [14]. By allowing users to interact with and explore the generated visualizations, our project incorporates these concepts and promotes a deeper comprehension of the underlying data structures.

## Algorithmic Techniques:

As we know that the high-dimensional data visualizations produced are accurate and informative, but it is a difficult task that calls for careful consideration of various quality metrics. So keeping that in mind, we suggest a data transformation algorithm that employs quality metrics to direct feature selection and enhance the readability and interpretability of the resulting visualizations. To test these algorithms, we have come up with PCA, U-map, and t-SNE and are testing these algorithms on the MNIST dataset for visualization purposes.

So the first algorithm that we used is Principal Component Analysis (PCA), it is a statistical technique that is used worldwide to reduce the dimensionality of a large dataset by transforming it into a new coordinate system. This new coordinate system is capable of capturing the variance in the data. This algorithm works by finding the linear combination of the original variables which captures most of the variance in the data. These linear combinations are called principal components, here the first principal component is in the direction of data that has the most variation followed by the second component which is in the direction orthogonal to the first principal component. PCA is useful in finding the most important features in the dataset as well as PCA can be used to improve the accuracy of other machine learning algorithms.

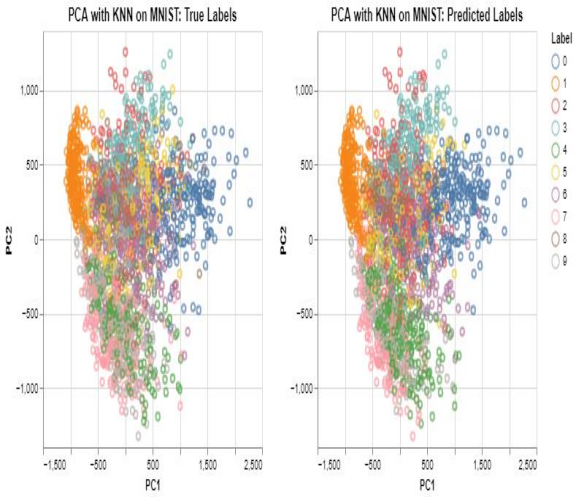
The second algorithm we used is U-Map, U-Map is also a dimensionality reduction technique used for visualizing high-dimensional data. Unlike the PCA technique, the U-map technique uses nonlinear manifold learning to map high-dimensional data to low-dimensional data. This algorithm starts by building a graph that represents the similarities between the high-dimensionality data; this graph is constructed using the nearest-neighbor approach, where each data point is connected to its k-nearest neighbor. U-Map then constructs a low-dimensional embedding of the data by optimizing the graph layout in a way that preserves the local structure of the data. Here, the optimization process is done using stochastic gradient descent, where the objective function is based on both groups of graph structure and the distance between points in the lower dimensional space.

The last dimensionality reduction technique is t-SNE. t-Distributed Stochastic Neighbor Embedding also known as t-SNE is a very powerful nonlinear dimensionality reduction technique used to visualize high-dimensional datasets in a lower-dimensional space, it deals with the two or three dimensions. This algorithm works by calculating the pairwise similarities between two data points in the high-dimensional space and then mapping them to the similarities in a lower-dimensional space using a probability distribution. This algorithm first constructs a probability distribution over pairs of high dimensional data points chosen in a way that has similar points. After that, it constructs a similar probability distribution over the points in the lower dimensional space and minimizes the Kullback - Leibler divergence between these

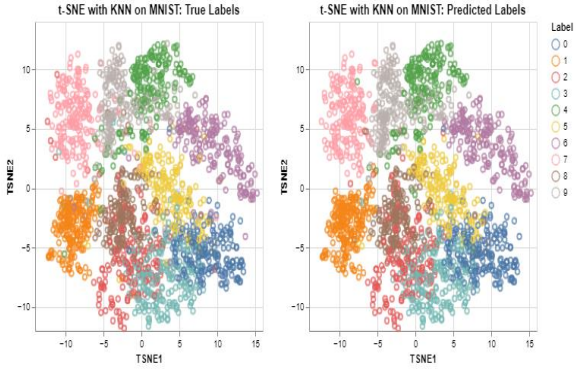
two probability distributions using gradient descent. The main advantage of t-SNE is that it efficiently preserves the local features of the data points that are closer together in the high-dimensional space and are also likely to be close together in the lower-dimensional space. This attribute is particularly useful for complex datasets with multiple clusters and has non-linear relations with features.

All of the above algorithms use scikit-learn. Scikit-learn is a free open-source machine learning library that is used in Python programming. The U-Map algorithm uses 1 U-Map-learn library.

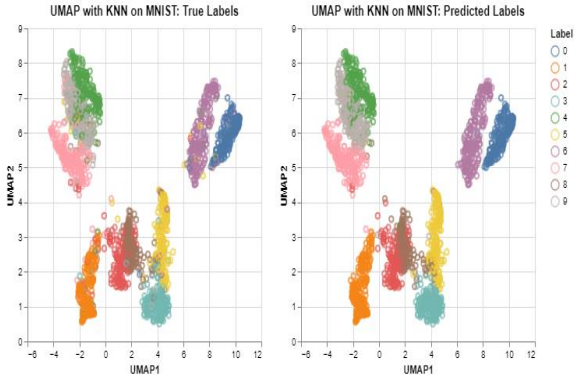
**PCA:**




**t-SNE:**




**U-Map:**




Here, we are using the KNN classifiers (K-Nearest Neighbor classifier) to evaluate our PCA, U-Map, and t-SNE algorithms. We are implementing a KNN classifier with KD tree and parallelism which we have considered using to speed up the implementation process. These two methods make the KNN classifier a lot faster and are generally executed within 3 mins to 4 mins, unlike the vanilla KNN classifier which takes around 25 mins and takes a toll on the system. The KNN classifier uses the entire dataset during the prediction process. The challenges which are faced by vanilla KNN classifiers such as it makes the algorithm computationally expensive when dealing with large datasets, as it requires calculating the distances between each data point are addressed by the KD tree, which is a binary search tree that divides data points into regions to speed up

the search for the nearest neighbors. Furthermore, parallelism is also being used to further improve the efficiency of the KNN classifier algorithm. Parallelism involves dividing the computation across multiple processors or nodes, allowing multiple computations to be performed simultaneously which increases the efficiency of the KNN algorithm.

Using a KD tree and parallelism with the KNN algorithm can significantly reduce the time required for training and prediction, making it more suitable for large datasets.

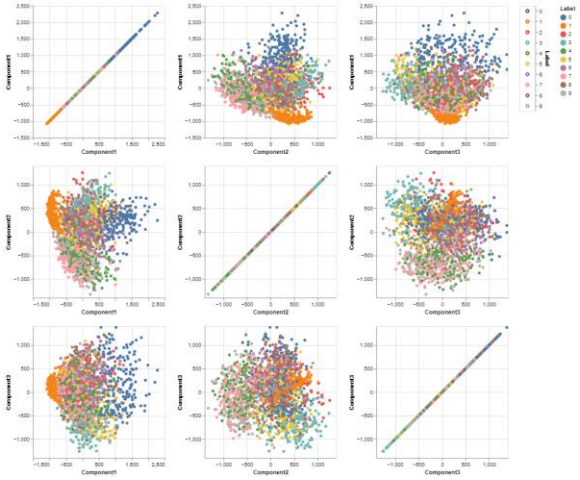
Rotating Variance Measure (RVM) is a measure that quantifies the benefits of a PCA, t-SNE, and U-Map transformation by considering the ratio of the variances of the transformed variables. Here, we apply the RVM to evaluate the PCA, t-SNE, and U-Map algorithms. It can be used in combination with PCA, where it identifies the most important principal component that contributes to the target variable and creates a sparse model that explains the data with fewer features. In t-SNE and U-Map, RVM can be applied to the low dimensional embedding obtained after applying t-SNE or U-Map to perform regression or classification tasks on the reduced dataset.

## Visualization and Interaction:

The benefit of visualization is that it allows a person to understand the dataset and its patterns and trends easily that may not come into notice if not apparent from looking at raw data. The human brain is more receptive to understanding graphical or pictorial information rather than that textual. There are many different ways in which the data can be represented accurately and reflects well on the underlying data. Interactive visualization refers to the visual representation of data that allows users to actively interact with the visualization and manipulate the data. These visualizations, unlike static visualizations, which present data in a fixed format, interactive visualizations give users the ability to explore the data in a more detailed way, which helps to uncover new insights that might not be apparent from a static representation. Interactive visualizations can take many forms, it may include drawdowns to choose components from and simple sliders to see the filtered and sorted data, which allows one to zoom in and out of the dataset, and hover over individual data points to view more information. The main goal of interactive visualization is to create an immersive and engaging experience for users which makes it easier for them to understand the dataset. In our case, we have implemented the interactive scatter plot, which showcases the PCA, t-SNE, and U-Map algorithm's results on the MNIST dataset. We have implemented two separate evaluation techniques, 1. KNN (K-Nearest Neighbors Algorithm) and 2. RVM (Rotating Variance Measure).

Here, we have created interactive visualizations of the dataset which enables users to understand the MNIST dataset and its components. We have compared our results with different metrics and we have plotted separate scatter plots for PCA, t-SNE, and U-Map along with Scatterplot matrices which showcase their results from which we can understand how different components generate different plots according to their properties.

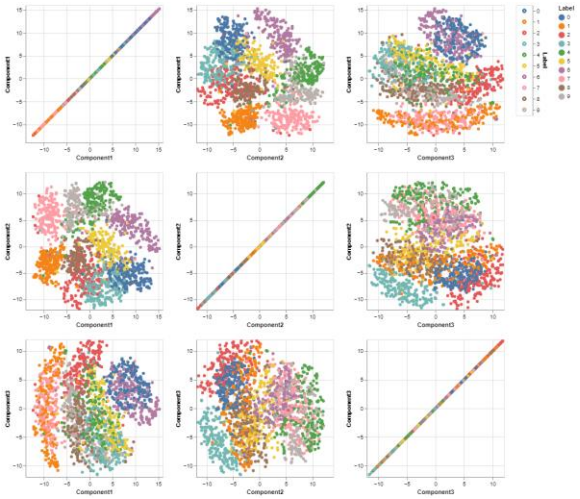
**PCA:**



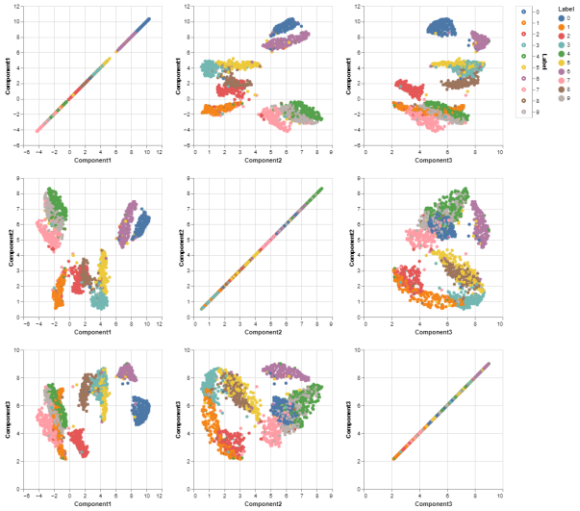
Here, we have implemented an interactive visualization of the PCA algorithm using a scatterplot matrix which helps us to understand how components 1, 2, and 3 interact with each other. This visualization is interactive which lets users zoom in and out and choose which component they want to highlight from the rest of the items in the graph.

**t-SNE:**

Here, we have implemented t-SNE results in an interactive graph, which shows how the different components interact with each other. We can see uniform clusters in the graphs where the components are unique and we can see a straight 45° line which shows that we are comparing two same components on the X-axis and Y-axis.



**U-Map:**



In the above visualization, we can see that the U-Map results are implemented for components 1, 2, and 3. These graphs are also interactive and can let users choose the action they wanna perform on these results to modify them in a way that the user desires. Here, we can see that clusters are a bit separated from each other which shows that no two data items in the graph recede together.

In this way, we have implemented the visualization for our project of dimensionality reduction using PCA, t-SNE, and U-Map using KNN and RVM evaluation techniques. It is important to note that KNN

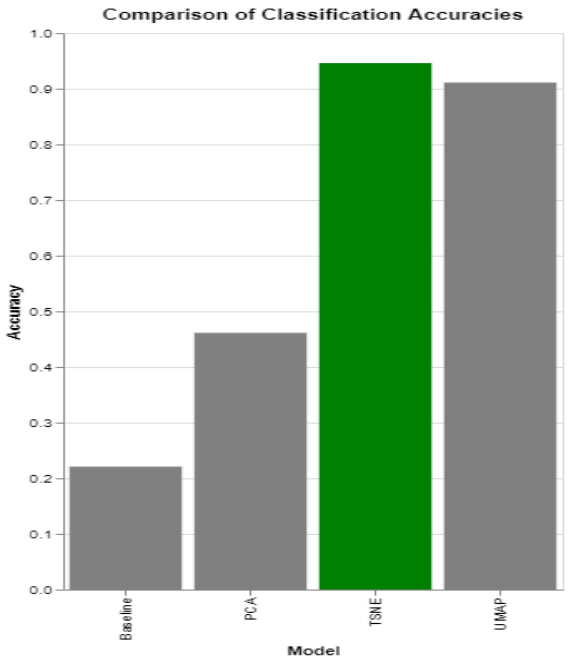
and RVM have been implemented without the use of standard machine learning libraries like scikit-learn or pycaret.

## Evaluation Metrics:

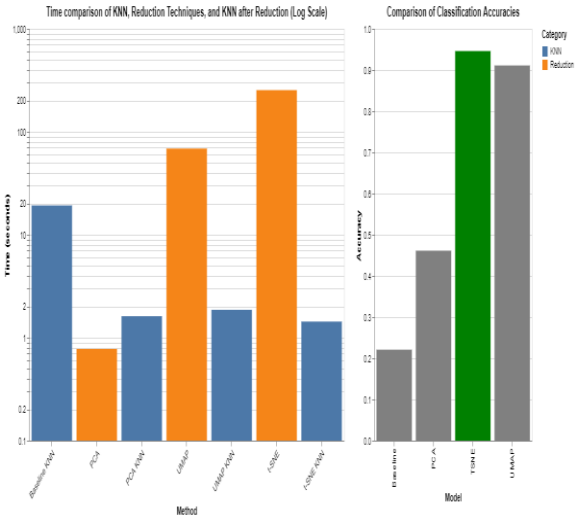
For the evaluation, metrics are referred to as the quantitative measures that are used to evaluate the process and fitness of that particular dataset. These metrics are used to analyze the genetic diversity over a period of time and assess the effects through different iterations. These metrics help to quantify the range in which the data has been improved or worsened over the period of time. These evaluation metrics are used to make complex datasets more readable and understandable. Here we are using scalability, and interpretability. The readability metrics are used where the data set is complex and we have to evaluate the efficiency of presentation in communicating the intended message to the target viewer. Overall, the goal of readability metrics is to create data visualizations that effectively communicate the intended message to the viewer in a clear and understandable manner.

As discussed in the previous stage, we have implemented interpretability and scalability evaluation metrics. When evaluating the performance of dimensionality reduction algorithms such as PCA, t-SNE, and U-Map, it's important to consider the metrics related to readability, interpretability, scalability, and robustness.

In the Interpretability metrics, t-SNE and UMAP are considered highly interpretable since the embedding is obtained through a complex nonlinear transformation, PCA is less interpretable since the principal components have a clear linear relationship with the input variables. In terms of Scalability, PCA is known to be scalable and can handle datasets with millions of samples. t-SNE and UMAP are less scalable and can be computationally expensive for large datasets.




Accuracy is the most straightforward metric and measures the percentage of correctly classified samples out of the total number of samples. While accuracy is a useful metric, it may not always be the best choice, especially when dealing with imbalanced datasets.



## Conclusion:

In this project, we have unveiled the significant potential of using an evaluation framework to assess the effectiveness of different dimensionality reduction techniques, which are crucial in the field of big data. The three techniques that we have used to demonstrate this are PCA, t-SNE and U-Map. We have taken k-NN model to evaluate the effectiveness of dimensionality reduction techniques in terms of prediction quality. The improvement in the F1 score is a strong indication that the essential information is preserved while reducing noise, leading to overall enhanced model performance. Through the introduction of Rotating Variance Measure (RVM), we provided a quantitative measure to evaluate the effectiveness of these demsionality reduction methods. This measure allowed us to quantify the amount of the variance retained in the transformed variables, thereby providing an objective measure of the quality of the dimensionality reduction process.

## Future Scope:

In the future, we want to try and implement Linear Discriminant Analysis, a supervised learning technique that finds the linear combination of features that best separates classes in a dataset. It is commonly used for dimensionality reduction and classification tasks. Another method we want to try out is Random Forest Embedding which is an unsupervised learning technique that uses a variant of Random Forests to embed data into a lower-dimensional space. It is particularly useful for visualizing high-dimensional data and can also be used for clustering and classification tasks.

## References:


1. Mustakim, N Gayatri, Okafalisa, I. Gazalaba. “Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification.”, 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)

2. A. Tatu et al. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In Proc. IEEE Symp. Visual Analytics Science and Technology (VAST), 2009.

3. Q. Cui, M. Ward, E. Rundensteiner, and J. Yang. “Measuring data abstraction quality in multiresolution visualizations.”, IEEE Trans. on Visualization and Computer Graphics, 12:709–716, 2006.

4. T. Long. “iRadviz: an inversion radviz for class visualization of multivariate data visualization.”, SoICT '16: Proceedings of the 7th Symposium on Information and Communication Technology

5. Q. Liu, R.Chen, H. Zhu, H. Fan. “Research and Comparison of Data Dimensionality Reduction Algorithms.”, ICBCI 2017, September 8-11, 2017, Beijing, China

6. S. Feng, H. Wang. “Comparison of PCA and LDA Dimensionality Reduction Algorithms based on Wine Dataset.” , 2021 33rd Chinese Control and Decision Conference (CCDC)

7. S. Johansson and J. Johansson. “Interactive dimensionality reduction through user-defined combinations of quality metrics.”, IEEE Trans. On Visualization and Computer Graphics, 15:993–1000, 2009

8. A. Kazemipour, S. Druckmann. "Nonlinear Dimensionality Reduction Via Polynomial Principal Component Analysis.”,

2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)

9. [BTK11] BERTINI E., TATU A., KEIM D. A.: Quality metrics in high-dimensional data visualization: An overview and systematization. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 17, 12 (2011), 2203–2212.

10. [SMT13] SEDLMAIR M., MUNZNER T., TORY M.: Empirical guidance on a scatterplot and dimension reduction technique choices. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 19, 12 (2013), 2634–2643.

11. [LvdMdS12] LEWIS J. M., VAN DER MAATEN L., DE SA V.: A behavioral investigation of dimensionality reduction. In Proc. 34th Conf. of the Cognitive Science Society (CogSci) (2012), pp. 671–676.

12. [Str02] STREHL A.: Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining. Ph.D. thesis, University of Texas at Austin, 2002.

13. J. S. Yi, Y. a. Kang, J. Stasko, and J. Jacko. Toward a deeper understanding of the role of interaction in information visualization. IEEE Trans. on Visualization and Computer Graphics, 13:1224–1231, 2007.

14. L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. IEEE Trans. on Visualization and Computer Graphics, 12:1363– 1372, 2006.