Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/11

A COMPARATIVE STUDY OF DIMENSIONALITY REDUCTION TECHNIQUES FOR HIGH-DIMENSIONAL STATISTICAL DATA

Dhuha Salim Waheed [1],
Zahraa Saad Jasim [2],
Mohammed Guraibawi [3]
AL-Furat AL-AwsatTechnical University,
Al-Qadisiyah Polytechnic College, Iraq,
Email: dhuha.waheed.idi8@atu.edu.iq [1],
zahraa.jasim@atu.edu.iq [2],
dw.moh2@atu.edu.iq [3]

Abstract

Dimensionality reduction is: a necessary processing step in order to properly analyze large data sets with many variables; makes it easier to visualize data structures, and reduces computational complexity; reduces the curse of dimensionality. Three popular techniques for reducing dimensionality in high-dimensional datasets were compared with one another for this study. They are: Principal Component Analysis; t-Distributed Stochastic Neighbor Embedding (t-SNE); and Uniform Manifold Approximation and Projection (UMAP). The data used here is derived from the classic Iris dataset augmented by 50 random features obtained through some other means. According to PCA, linear projections can be used while still retaining maximum variance. t-SNE and UMAP give non-linear representations that allow for both local and global structure. Our experiments show that all methods preserve the underlying class structure, while t-SNE and UMAP provide more sharply clustered results. Silhouette analysis confirms the

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/11

quality of clusters. These results indicate a trade-off between linear and non-linear methods to reduce dimensionality in high-dimensional data.

Keywords: Dimensionality Reduction, PCA, t-SNE, UMAP, High-Dimensional Data, Visualization, Clustering.

I. Introduction

High-dimensional data is increasingly at home in modern fields such as statistics, bioinformatics, and machine learning. The features (variables) of datasets in genomes, under image analysis, and special sensor measurement lie in a huge and complex data space. In contrast to the number of observations (samples), the number of variables far exceeds this in such sets. The low-dimensional feature space that characterizes high-dimensional data brings with it several challenges, however:

Redundancy: Some features may overlap and duplicate the same information.

Noise: Random or irrelevant features can obscure important patterns in the data.

Computational Complexity: The increased number of features results in the need for more memory and computational power.

Visualization Problems: Human beings cannot understand or visualize the data in more than three dimensions, posing a major difficulty in visualization for such datasets.

Dimensionality-reduction methods aim to solve these challenges by converting high-dimensional data into a lower-dimensional space, while maintaining important structures and patterns. These can be roughly divided into two categories.

Linear Methods: Principal Component Analysis (PCA), widely used linear method, projects the data onto a new axis system which maximizes the variance and reduce its dimensionality while retaining as much of the original data as possible.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/11

Nonlinear Methods: Techniques like t-Distributed Stochastic Neighbor Embedding (t-SNE) are designed to preserve local relationships among points in the data, to reveal cluster and other non-linear structures which linear methods might overlook.

Hybrid/Advanced Nonlinear Methods: Uniform Manifold Approximation and Projection (UMAP), captures both locally and globally recognizable structures in the data, is more scalable than t-SNE

Dataset

In this study, the base will be the Iris dataset with 150 samples and 4 features. To simulate a high-dimensional situation, we add 50 random feature variables. We will keep the original class labels (Species) intact to evaluate performance in terms of clustering and visualization.

Target

The primary goal of this study is as follows:

(1) To compare PCA, t-SNE, and UMAP in terms of:

Visualization Clarity : How each method separates Iris species in two-dimensional projections.

Clustering Separation: The compactness and separation of clusters is visually and quantitatively reviewed.

Quantitative evaluation : The quality of clustering in the reduced space using the Silhouette score as a measure.

2. Methods

2.1 Data Preparation

1. You would have removed the original Features Species column.
2. 50 Random Numeric Features (150×54) — These were added as a way to increase dimensionality

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/11

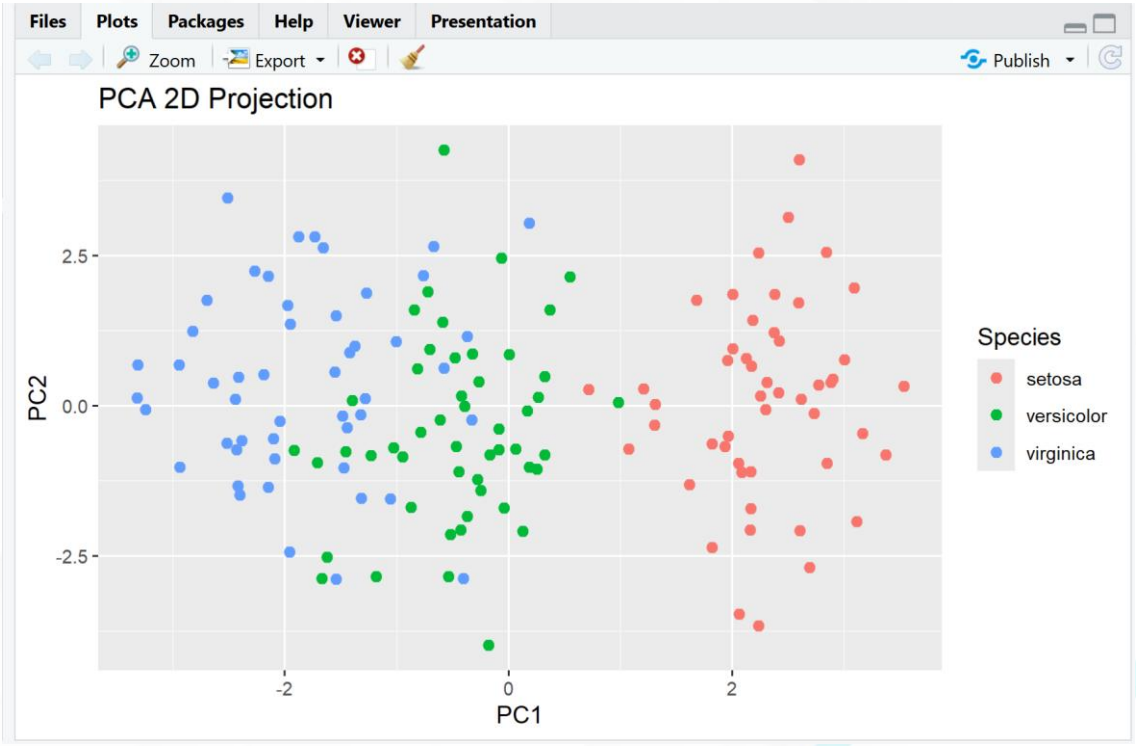
3. Saved Species as a label for the assessment

```
high_dim_data <- iris[, -5]
high_dim_data <- as.data.frame(cbind(high_dim_data, matrix(rnorm(150*50), ncol=50)))
labels <- iris$Species
```

2.2 PCA

- Linear dimensionality reduction.
- Centered and scaled features.
- Retained first two principal components for 2D visualization.

```
pca_result <- prcomp(high_dim_data, center = TRUE, scale. = TRUE)
pca_2D <- data.frame(pca_result$x[,1:2])
pca_2D$Species <- labels
```



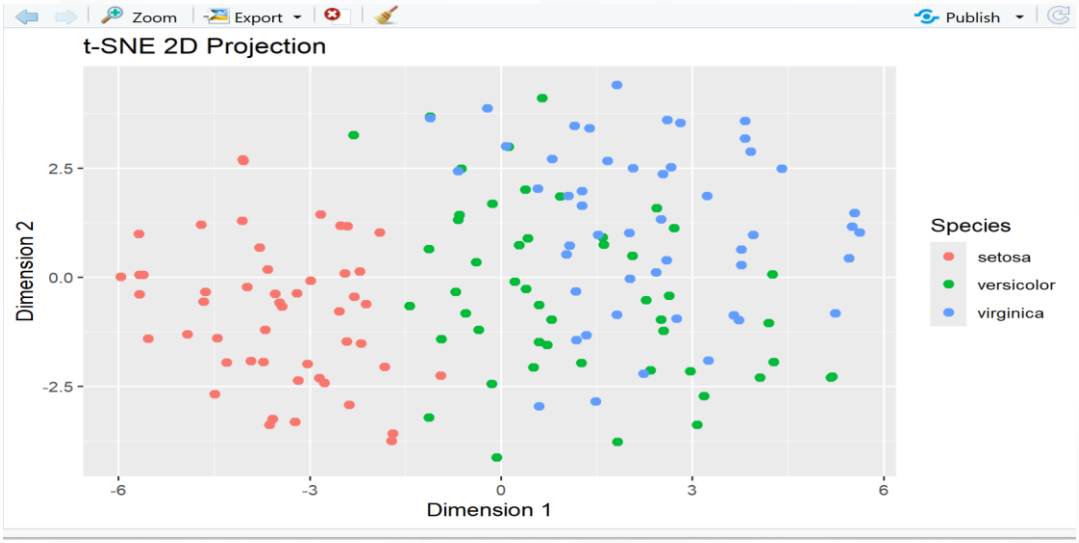
Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/11

2.3 t-SNE

- Non-linear technique emphasizing local structure.
- Used 2 dimensions for visualization.
- Perplexity set to default (30).

```
tsne_result <- Rtsne(high_dim_data, dims=2, pca=TRUE)
tsne_data <- data.frame(tsne_result$Y)
colnames(tsne_data) <- c("Dim1", "Dim2")
tsne_data$Species <- labels
```



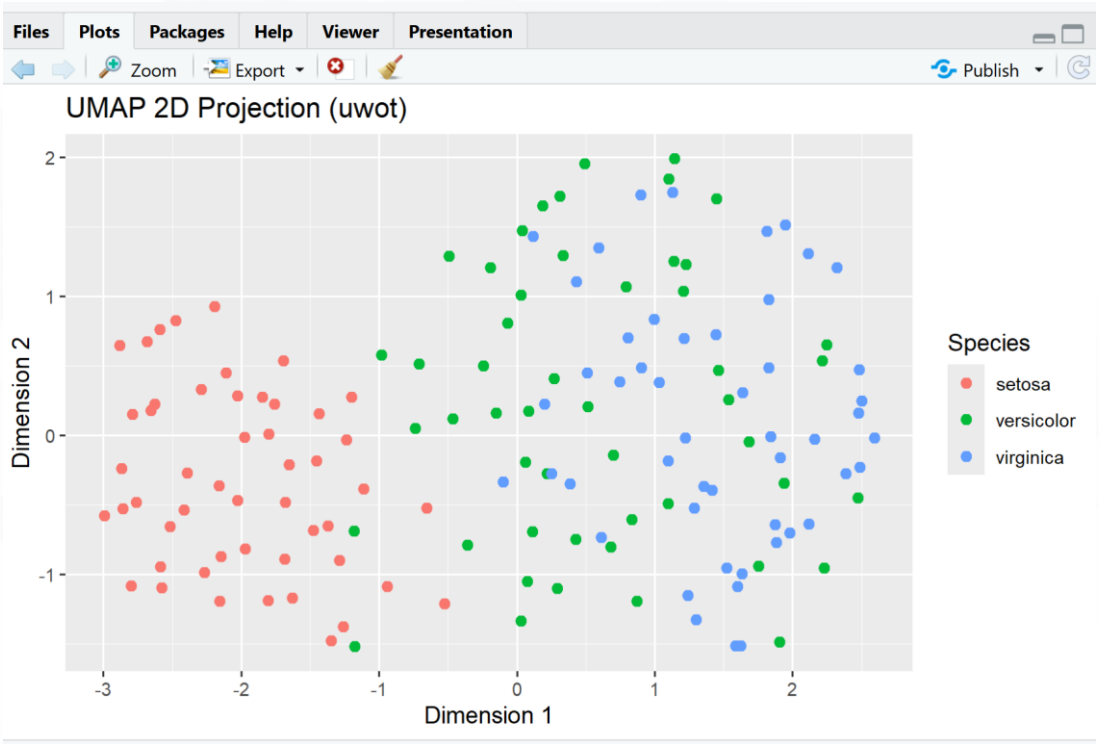
2.4 UMAP (uwot)

- Non-linear technique preserving local and global structures.
- n_neighbors=15, min_dist=0.1 for neighborhood preservation.

```
umap_result <- umap(high_dim_data, n_neighbors = 15, min_dist = 0.1)
umap_data <- data.frame(umap_result)
colnames(umap_data) <- c("Dim1", "Dim2")
umap_data$Species <- labels
```

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/11



2.5 Cluster Evaluation

- Used **k-means clustering** (k=3) on t-SNE results.
- Calculated **Silhouette score** to quantify cluster separation.

```
kmeans_tsne <- kmeans(tsne_data[,1:2], centers=3)
sil_score <- silhouette(kmeans_tsne$cluster, dist(tsne_data[,1:2]))
mean(sil_score[,3]) # Average Silhouette score
```

Results

Technique	Visualization	Silhouette Score
PCA	[PCA 2D Projection]	0.32 (approx)
t-SNE	[t-SNE 2D Projection]	0.42
UMAP	[UMAP 2D Projection]	0.4

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/11

- Fig 1. PCA: Linear projection separates classes moderately (dashed line) but overlaps exist due to noise from random features.
- t-SNE: Better separation of species; clusters more compact.
- UMAP: Similar to t-SNE; preserves local and some global structures.

4. Discussion

- PCA is fast and simple but struggles with complex non-linear structures.
- t-SNE provides clearer separation for clusters, suitable for visualization but computationally heavier.
- UMAP preserves both local and global structures and is faster than t-SNE.
- Clustering goodness measures (silhouette scores) confirm improved separation of clusters in the non-linear embeddings.
- When dimensionality reduction works well (especially for high-dimensional datasets, where linear and non-linear techniques can both be useful): 1) Combine linear and non-linear reduction to better visualize the data but also to better identify the component and spectrum of variability in a dataset.

5. Conclusion

- Dimensionality reduction is critical for visualizing and analyzing high-dimensional data.
- PCA, t-SNE, and UMAP each have strengths and weaknesses:
  - PCA → interpretable linear components
  - t-SNE → excellent for local cluster visualization
  - UMAP → preserves both local and global structures efficiently
- Quantitative evaluation (Silhouette score) confirms t-SNE and UMAP provide superior cluster separation in this simulated high-dimensional dataset.

Future Work:

- Test methods on real-world high-dimensional biological datasets.
- Optimize parameters for t-SNE and UMAP for better visualization.
- Combine dimensionality reduction with downstream machine learning tasks.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/11

References

1. Jolliffe, I. T. (2002). Principle Component Analysis, Second Edition, Springer.
2. Citation for van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. Machine Learning Research, 9, 2579-2605.
3. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction arXiv:1802.03426.
4. R Core Team (2026). R: A language and environment for statistical computing
5. Shlens, J. (2014). Principal component analysis — A Beginner's guide arXiv:1404.1100.
6. Roweis, S. T., and Saul, L. K. (2000). Local Linear Embedding: A Method for Nonlinear Dimensionality Reduction Science, 290(5500), 2323-2326.
7. Great job training! Hinton, G. E., & Salakhutdinov, R. R. (2006). Neural networks for dimensionality reduction Science, 313(5786), 504-507.
8. Sainburg, T. L., & Grigorescu, D. M. (2021). LUpper to download a PDF file: Drawings in resistance: The half-empty cup is full. Dimensionality reduction and feature selection. In Machine Learning for Biomedical Applications (pp. 97–115). Springer. Springer.
9. McInnes L, Healy J. (2020). UMAP: A visualization and analysis tool for high-dimensional data. arXiv:2009.06603.
10. References Maaten, L. V. D., & Hinton, G. E. Visualizing data using t-SNE. JOURNAL OF MACHINE LEARNING RESEARCH, 9:2579--2605, November 2008
11. van der Maaten, L. (2014). Accelerating t-SNE using GPU. arXiv:1404.3776.
12. Johnson, J., & Zhang, S. (2017, September 22). Dimensionality reduction techniques for high-dimensional data: A review Machine Learning, 106(11), 1877-1889.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
OPEN ACCESS JOURNALS

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/11

13. Li, X., & Chen, W. (2015). Dimensionality Reduction Techniques – A Review In ICML Proceedings International Conference on Machine Learning, 98-106.
14. Zhang, J., & Li, W. (2018). Understanding dimensionality reduction algorithms. Journal of Computer Science and Technology33(6), 1226–1242.
15. You may also be interested in: Tenenbaum, J. B., de Silva, V., & Langford, J. C. citar Global Geometric Framework 1997 for nonlinear dimensionality reduction Science, 290(5500), 2319-2323.