IPS Journal of Physical Sciences
ISSN: 3115-5715
IPS J Phys Sci, 2(2): 101-113 (2025)
DOI: https://doi.org/10.54117/ijps.v2i2.15

Efficient Non-Parametric Methods for Dimensionality Reduction in High-Dimensional Non-Linear Multivariate Data

Brandy Ogbenyealu Nleonu[1], Cecilia Nchedo Okoli[2] and Jude Chukwura Obi[2]

[1]Department of Mathematics and Statistics, Federal Polytechnic Nekede, Owerri, Imo State, Nigeria.
[2]Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Uli. Anambra State, Nigeria.
*Corresponding author's email: ukobrandy@yahoo.com

Abstract

Analysis of high-dimensional non-linear multivariate datasets frequently violates the assumptions underlying classical parametric techniques. This study conducts a systematic comparative evaluation of five prominent non-parametric dimensionality-reduction methods: Isomap, Uniform Manifold Approximation and Projection (UMAP), Locally Linear Embedding (LLE), Kernel Principal Component Analysis (KPCA), and t-Distributed Stochastic Neighbor Embedding (t-SNE), using one synthetic dataset and three real-world datasets drawn from genomics, macroeconomics, and social-media analytics. Performances of the methods were assessed with respect to their computational efficiency and their ability to preserve the local and global structure of the data. This was measured through computational time, trustworthiness, mean square error, reconstruction error, and Spearman rank correlation. Across the empirical datasets, UMAP consistently exhibited superior speed and fidelity in structure preservation, with KPCA emerging as a strong second performer. Friedman rank tests indicated significant differences among the performances of the methods. However, the simulated data did not yield any notable significant difference.

Keywords: Dimensionality reduction, non-parametric methods, manifold learning, UMAP, trustworthiness.

Article History
Received: 22 Nov 2025
Accepted: 09 Dec 2025
Published: 14 Dec 2025

How to cite this paper: Nleonu, B. O., Okoli, C. N., & Obi, J. C. (2025). Efficient Non-Parametric Methods for Dimensionality Reduction in High-Dimensional Non-Linear Multivariate Data. *IPS Journal of Physical Sciences*, *2*(2), 101–113. https://doi.org/10.54117/ijps.v2i2.15

Introduction

Analysis involving multivariate datasets that carry real complexity has come to occupy a central place in contemporary research, whether the application lies in clinical studies, financial risk assessment, environmental modeling, or a host of other domains (Little & Rubin, 2019). The core difficulty often stems from the dense web of interdependencies among variables. Because of those interdependencies, the parametric workhorses we have long relied upon MANOVA, canonical correlation analysis, and related procedures frequently fall short. Their validity hinges on assumptions (multivariate normality, equal covariance matrices across groups) that are violated more often than honoured in actual practice (Hazra & Gogtay, 2017). What investigators typically encounter instead are datasets marked by very high dimensionality, decidedly non-linear patterns, marked heterogeneity in distributional form, and no small number of missing values. These features collectively undercut the foundations of classical methods (Troyanskaya *et al*., 2001). Under these circumstances, the search for reliable analytic strategies has increasingly turned toward non-parametric procedures that make few, if any, demands on the shape of the underlying distribution (Van Buuren, 2018)

In recent years a rich family of non-parametric multivariate techniques has appeared, each designed to sidestep the restrictive assumptions of its parametric predecessors (Hastie *et al*., 2009). Isometric mapping (Isomap), uniform manifold approximation and projection (UMAP), locally linear embedding (LLE), kernel principal component analysis (KPCA), and t-distributed stochastic neighbor embedding (t-SNE) stand out as especially influential examples. These approaches have shown real promise when the data structure is strongly non-linear or when observations have high dimensionality (van der Maaten *et al*, 2019; Lee & Kim, 2020).Yet the very abundance of options creates its own problem: the algorithms differ substantially in their geometric underpinnings, computational demands, and sensitivity to tuning parameters, so deciding which one best fits a particular problem is rarely obvious (García-Laencina *et al.*, 2010). At the same time, head-to-head empirical comparisons; especially those that place the new non-parametric tools alongside older parametric benchmarks across diverse, realistic datasets remain surprisingly scarce (Li *et al*., 2020).

The study described here was undertaken to help close those gaps. Its primary goal is to offer a careful, systematic comparison of Isomap, UMAP, LLE, KPCA, and t-SNE, with special attention to their performance on the stubborn problem of dimensionality reduction when the intrinsic geometry of the data is non-linear and the dimensionality is high (Van der Maaten & Hinton, 2008; Belkin & Niyogi, 2003). Beyond raw performance metrics, the investigation considers how well each method supports meaningful visualization and interpretation of the resulting low-dimensional representations, and it draws out practical implications for method choice.

In doing so, the work hopes to provide analysts with clearer guidance about when a given technique is likely to excel and when it may mislead. More generally, it seeks to sharpen methodological practice in multivariate analysis, to bolster confidence in findings drawn from complicated real-world data, and to contribute to ongoing developments in statistical theory and data science (Anowar *et al.*, 2021). Finally, because the comparison is structured around concrete examples and diagnostic criteria, the results should also serve as a useful resource for graduate teaching, helping instructors illustrate both the power and the subtleties of modern non-parametric approaches.

Methodology

This study employed data collection from two primary sources namely: Secondary Data collection and Simulated Data. Three (3) Secondary datasets comprising of Gene Expression RNA cancer data from TCGA at https://portal.gdc.cancer.gov, Nigerian Economic data from Nigerian bureau of statistics database named https://nigerianstat.gov.ng and social media data from www.kaggle.com. The data collected was tested for linearity using the Generalized Linear Model (GLM) and the Generalized Additive Model (GAM). The criteria for selection of the multivariate data included complexity of the datasets (number of variables and observations, non linearity, high dimensionality) and relevance to the study's objectives. In addition to real secondary datasets, simulated datasets was generated to control for specific characteristics, such as varying degrees of noise and correlation among variables.

Statistical Tools

The R statistical software was used to carry out non parametric dimensionality reduction on both the simulated datasets and the real life data collected. This was done under the following categories of methods: Isometric Mapping (ISOMAP), Uniform Manifold Approximation and Projection (UMAP), Locally Linear Embedding (LLE), t-Distributed Stochastic Neighbor Embedding (t-SNE) and Kernel Principal Component Analysis (KPCA). The metrics used for comparison were: Computation time of each method, Mean Square Error (MSE), correlation, reconstruction error and trustworthiness.

Isometric feature mapping (ISOMAP)

ISOMAP is carried out by trying to find an embedding in which the geodesic distance between two points in the input space is as close as possible to the Euclidean distance between target space projection.

$$\text{ISOMAP} = \min \|\tau(D_G) - \tau(D_Z)\|_F \quad (1)$$

Where D_G = matrix of the geodesic distance between points in the neighbourhood graph
$D_z=[d_{ij}]$ matrix of pair wise Euclidean distances, $d_{ij}=\|Z_i - Z_j\|$ of the data projections in $\Re^d$
τ is the conversion operator of the inner products, $\|\bullet\|$is the Frobbenius norm of a matrix.

The minimum of equation (1) is gotten by calculating the d eigenvectors associated to the d largest eigen values of the geodesic distance matrix $\tau(D_G)$

Uniform Manifold Approximation and Projection (UMAP)

The Uniform Manifold Approximation and Projections (UMAP) calculates the pairwise distances for a given dataset such that, given a dataset $X=\{x1,x2,...,xn\}$ in RD, the pairwise distances are given as

$$d_{ij} = \|x_i - x_j\|^2 \quad (2)$$

where d_{ij} is the Euclidean distance between *xi* and *xj*.

After which the neighborhood graph N(x_i) is computed for each *xi* and its *k*-nearest neighbors:

$$\text{N}(x_i) = \{x_j : d_{ij} \leq d_{jk} j = i, |x_j| = k\} \quad (3)$$

Then the fuzzy membership strengths are computed using:

$$\mu_{ij} = e^{(-\sigma_i d_{ij} - \rho_i)} \quad (4)$$

Where $\rho_i = d_i$N(xi), and σi is a normalization factor. Furthermore, the cross-entropy loss is minimized to get the UMAP cost function using

$$L = \sum\{i \neq j\}[u_{ij} \log(u_{ij} v_{ij}) + (1 - u_{ij}) \log(1 - v_{ij} 1 - u_{ij})] \quad (5)$$

where *vij* represents the fuzzy membership strength in the low-dimensional space. Finally, the low-dimensional representation $Y = \{y_{1,} y_{2,} \ldots., y_n\}$ in Rd is optimized using stochastic gradient descent: *yi←yi−α∂yi∂*L

Locally Linear Embedding (LLE)

LLE makes an attempt at recovering the global structure of imputed data. It is represented mathematically as

$$\text{Minimize: } \sum_i |x_i - \sum_j w_{ij} x_j|^2 \quad (6)$$

$$\text{Subject to: } \sum_j w_{ij} = 1$$

Where x_i is the ith data point
w_{ij} is the weight that minimize the reconstruction error for data points x_i using its neighbours.

Kernel Principal Component Analysis (KPCA)
This is extension of the PCA to non linear data. Here, we first choose a desired kernel function say $K(x, y)$ which is a scaler product in target space. Then compute a gram/ kernel matrix K with $k_{ij} = K(x_i, x_i)$. The kernel matrix is centered using
$K_{centered} = K - ln - Kln + lnKln = (1 - \ln)K(1 - ln)$ (7)
Where ln is a $n \times n$ matrix with all elements equal to $\frac{1}{n}$ and n is the number of data points.

Finally, the eigenvectors U and eigenvalues S^2 of the centered kernel matrix is computed and each eigenvector is multiplied by the square root of the corresponding eigen value.

t-Distributed Stochastic Neighbor Embedding
The t-distributed Stochastic Neighbor Embedding (t-SNE) is carried out by firstly computing the pairwise Similarities in High-dimensional Space. For each pair of points x_i and x_j in the high dimensional space, we compute the conditional probability $P_{j/i}$ of choosing x_j as a neighbor of x_i such that $P_{j/i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/2\sigma_i^2)}$
(8)
Where σ_i is a parameter that controls the width of the Gaussian centered on x_i and can vary per point to adapt to local density.
Furthermore, we symmeterize the joint probability P_{ij} using
$P_{ij} = \frac{P_{j/i} + P_{i/j}}{2N}$ (9)

Where N is the total number of points.

After this, we compute pair wise similarities in low-dimensional space by computing the probability q_{ij} for each $y_i$and y_j using a Student t-distribution with one degree of freedom (which is equivalent to the Cauchy distribution) instead of a Gaussian: $q_{j/i} = \frac{(1+\|y_i - y_j\|^2/2\sigma_i^2)^{-1}}{\sum_{k \neq i} \exp(1+\|y_k - y_l\|^2)^{-1}}$ (10)

Finally, we minimize the KL Divergence Cost Function by minimizing the Kullback-Leibler (KL) divergence between the distributions P and Q: using
$C = KL(P \| Q) = \sum_{i \neq j} P_{ij} \log \frac{P_{ij}}{q_{ij}}$ (11)

The evaluation of the results of the dimensionality reduction for the methods will be done based on the following metrics; Correlation, mean square error (MSE), reconstruction error, computation time and trustworthiness.

Computation Time
Computation time refers to the total CPU or GPU time required to fit and transform data from high dimensional space $\mathbb{R}^D$ to a low dimensional space $\mathbb{R}^d$ (where d < D). This is the primary metric employed in assessing the scalability of dimensional reduction algorithms especially when dealing with real life high dimensional datasets especially when dealing with methods where speed is an essential necessity (Kobak and Berens, 2019).

Correlation
The correlation metric evaluates the preservation of pairwise relationship by quantifying the linear relationships between variables in dimensionality reduction. This is used to check how well a low dimensional space preserves the structure of a high dimensional dataset. The Pearson correlation coefficient is given as:
$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y}))^2}.}$ (12)
where x_i are the original points. y_i are the reduced data points. $\bar{x}$ is the mean of the original data points. $\bar{y}$ is the mean of the reduced data points and n is the number of data points (Field, 2018).

Trustworthiness
Trustworthiness measures how well the local neighborhood structure of high-dimensional data is preserved in the low-dimensional embedding. It is defined as: $T_k = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in \hat{U}_i} (\hat{r}(i,j) - k)$ (13)
Where; T_k = Trustworthiness Score, k = Neighborhood size, n = Number of data points, i = Index of a data point, j = index of a neighbor point, $\hat{U}_i$ = set of indexes for point i and $\hat{r}(i,j)$ = rank of j from i in the low dimension embedding (Venna & Kaski, 2001). Trustworthiness ranges from 0 to 1, with higher values indicating better preservation of local structure.

Mean Square Error

Mean Squared Error (MSE) is a metric that is widely used in evaluating the accuracy of predictive models. This is done by measuring the average squared difference between observed (y_i) and predicted ($\hat{y}_i$) values. A lower MSE indicates better model fit, with 0 representing perfect predictions. Due to squaring, MSE heavily penalizes large errors and is sensitive to outliers. It is differentiable, making it ideal as a loss function in regression and machine learning. The computational formula for MSE is given as;

$$\text{MSE} = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \quad (14)$$

Where: n = number of samples, p = number of predictors (use n−1 if no predictors, e.g., simple mean), y_i= observed values and $\hat{y}_i$= Predicted values.

Reconstruction Error

Reconstruction error is the difference between the original input data and the reconstructed data obtained from a model, like an autoencoder. It's a measure of how well the model captures the underlying patterns in the data. The formula for Reconstruction Error (RE) is: given by the mathematical notation:

$$\text{RE} = ||x - g(f(x))|| \quad (15)$$

Where: x is the original input, x' = g(f(x)) is the reconstructed input, f(x) is the encoder function, g(x) is the decoder function and ||.|| denotes a norm (e.g., L1, L2) (Goodfellow *et al;*2016).

After carrying out dimensionality reduction using the methods under study, the Friedman's test was used to compare the different methods to check if there is any significant difference in their performance. The Friedman test statistic is computed by the use of the formula:

$$\chi^2 = \frac{12}{nk(k+1)}\sum(R_j{}^2) - 3n(k+1) \quad (16)$$

Where χ^2: Friedman test statistic

n: Number of blocks (e.g., datasets or scenarios)k: Number of treatments (e.g., imputation methods) and R_j: Sum of ranks for treatment j

This formula calculates the test statistic, which is then compared to a chi-squared distribution with k-1 degrees of freedom to determine significance.

For cases where significant differences exist in the use of the methods, the Nemenyi's test was employed as post hoc test in order to ascertain which of the methods account for the existing significant difference. The Nemenyi's test also known as Nemenyi's test also known as Wilcoxon-Nemenyi-McDonald-Thompson test is an adaptation of the Tukey HSD test, It tests the difference between rank sums and is computed using the following formula:

$$q = \frac{R_i - R_j}{\sqrt{\frac{k(k+1)}{6n}}} \quad (17)$$

where: R_I and R_j are the rank sums of groups i and j, k is the number of groups and n is the number of blocks (or subjects).

The statistic follows the studentized range q distribution. The critical values for this distribution are presented in the studentized range q table based on the values of α, k (the number of groups) and $df = \infty$ (although some sources use $df = n - k$). If $q > q_{crit}$ then the two means are significantly different.

Results

The data employed in the analysis stems from four sources as shown in table 1

Table 1: Characteristics of datasets used in the study

Dataset	Type	Source	Description	Observations (n)	Variables / Features (p)
Simulated Data	Synthetic			5,000	100
Gene Expression (RNA Cancer)	Real	The Cancer Genome Atlas (TCGA) Pan-Cancer	Multiple cancer types	801	20,531
Nigerian Macroeconomic Indicators	Real	National Bureau of Statistics (NBS), Nigeria	January 1990 – December 2022 (monthly)	396	28
Social Media Text Embeddings	Real	Kaggle (BERT-base embeddings of trending-topic posts)	Recent trending issues & user opinions	50,000	768

High dimensional data was simulated and five dimensionality reduction techniques; KPCA, LLE, ISOMAP, UMAP and t-SNE were applied to reduce the high dimension of the data and the output is as tabulated in table 2.

Table 2: High dimensionality reduction methods and the performance metrics for Simulated data

Method / Metric	ISOMAP	t-SNE	UMAP	LLE	KPCA
MSE	528914.359	192856363.239	6.9360	576.372	400.1260
Reconstruction Error	212.1359	356431.62	2.0697	20.1734	15.62
Correlation	0.99996	0.9880	0.837	0.8306	0.9221
Computation time (sec)	0.02575	1.8342	9.25	5.75	2.10
Trustworthiness	0.97066	0.8789	.9993	0.99994	0.9801

Isomap was the fastest among the methods with computation time of 0.026secs. It also had the highest correlation of 0.99996. However, it had lower errors of 212.1359 (reconstruction error) and 528914.359 (MSE). UMAP indicated superior data representation accuracy as it has trustworthiness of 0.9993, MSE of 6.936 and reconstruction error of 2.0697. KPCA offers balanced performance, with strong MSE (400.1260), Reconstruction Error (15.62), and Trustworthiness (0.9801), and moderate Computation Time (2.10 seconds). LLE leads in Trustworthiness (0.99994) and performs well in MSE (576.372) and Reconstruction Error (20.1734), but its Correlation (0.8306) is the lowest. t-SNE performs poorly in MSE (192856363.239), Reconstruction Error (356431.62), and Trustworthiness (0.8789), despite a strong Correlation (0.9880). Friedman's test statistic was χ^2 = 2.08 with p-value = 0.721 indicating that no statistically significant difference occurred in overall performance of the methods, suggesting that no method consistently outperforms the others across all metrics. The gg visualization of the various methods is as shown in figure 1.

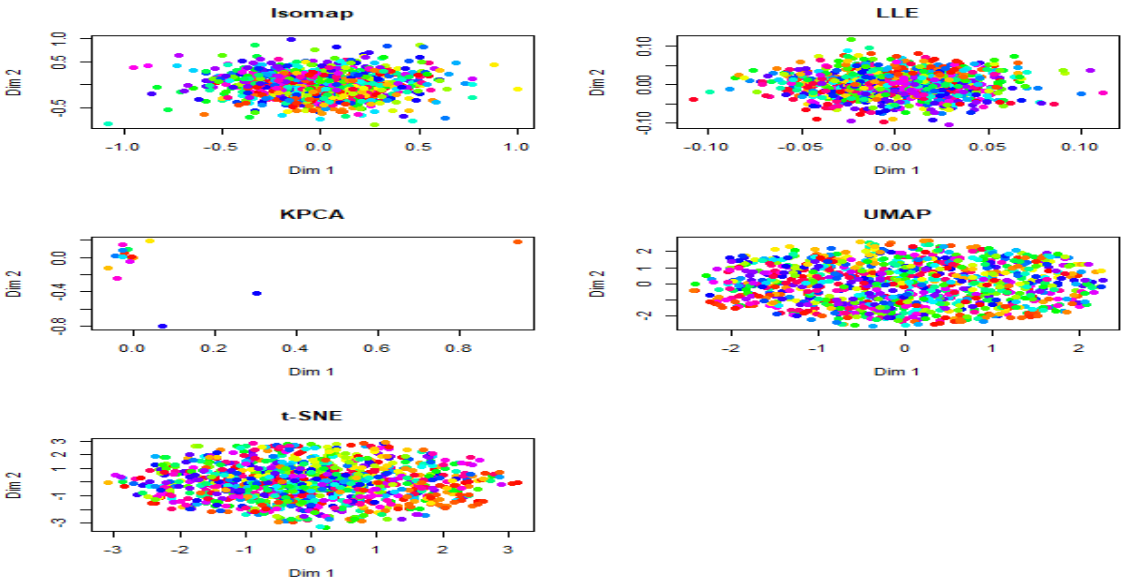

Figure 1: Plots of dimensionality reduction by different methods for simulated data

Analysis of Gene Expression Data

The Gene expression data was tested for linearity using the Generalized Linear Model (GLM) and the Generalized Additive Model (GAM) by computing their respective explained deviation and the Akaike Information Criterion (AIC). The Ramsey's reset test and F test were also used and the results is as shown in tables 3 below.

Table 3: Model test for linearity of Gene Expression data

Model	Explained Deviation	AIC	Ramsey's Reset p- value	F test p-value
GLM	0.4123	2456.721	0.0132	0.0087
GAM	0.5876	2312.416		

The test for linearity shows that the GAM outperforms the GLM significantly in terms of the deviation explained and AIC with values of 58.76% against 41.23% and 2312.4 and 2456.7 respectively. For the Ramsey's reset test, p=0.0132 points to a misspecification of the GLM which is likely due to unmodeled non-linear effect. The F-test (p = 0.0087) further confirms that the GAM's non-linear terms provide a statistically significant improvement in fit.

After ascertaining that the dataset was non linear hence meeting the criteria for analysis, dimensionality reduction was carried out using the different methods under study. The plots and results are as shown in figure 2 and table 4 respectively below.

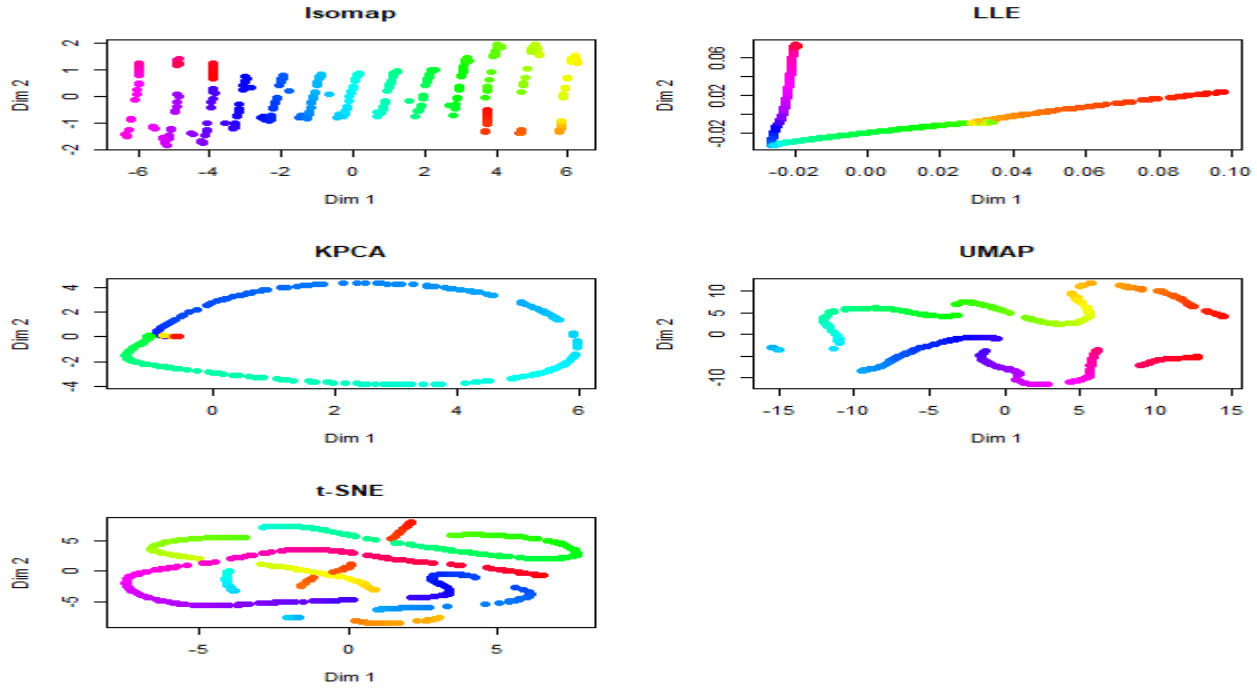

Figure 2: Plots of dimensionality reduction by different methods for Gene Expression RNA data.

Table 4: High dimensionality reduction methods and the performance metrics for Gene Expression data.

Method / Metric	ISOMAP	t-SNE	UMAP	LLE	KPCA
MSE	0.099	0.105	0.081	0.092	0.087
Reconstruction Error	0.310	0.321	0.265	0.298	0.275
Correlation	0.815	0.821	0.849	0.805	0.834
Computation time (sec)	50.33	45.12	15.67	38.45	22.78
Trustworthiness	0.905	0.912	0.935	0.899	0.921

From table 4, UMAP is seen to have the least computation time of 15.67secs making it the fastest method. It also has the least MSE reconstruction error of 0.081 and 0.265 respectively. Furthermore, its correlation coefficient of 0.849 and trustworthiness of 0.935 are the highest. Followed closely is the KPCA method which is considered relatively faster with 22.77secs and low errors of 0.275 for reconstruction error and 0.087 for MSE. Its quality matrices of 0.834 for correlation and 0.921 for trustworthiness are also relatively better compared to that of ISOMAP, LLE and t-SNE. ISOMAP was the slowest with computation time of 50.33secs and had the higher errors of 0.310 and 0.099 for Reconstruction error and MSE respectively.

Furthermore, the Friedman's test was carried out to ascertain if there exists any significant difference in the performances of the various methods with the statement of hypothesis below:
H_0: There exists no significant difference in dimensionality reduction methods.
H_1: There is a significant difference in dimensionality reduction methods.
To carry out the Friedman's test, the metrics values of table 4 were ranked and the output is as shown in table 5.

Table 5: Friedman's test ranking for High dimensionality reduction methods for Gene Expression data.

Method / Metric	ISOMAP	t-SNE	UMAP	LLE	KPCA
MSE	4	5	1	3	2
Reconstruction Error	4	5	1	3	2
Correlation	4	3	1	5	2
Computation time (sec)	5	4	1	3	2
Trustworthiness	4	3	1	5	2
Total Rank Sum	21	20	5	19	10

The Friedman's $\chi^2 = 16.6$ at df = 4 and a p-value of 0.0028 indicating a significant difference in the different methods of reducing dimensionality. To ascertain which of the methods differ significantly, the Nemenyi's post hoc test was done and the R output for the p-values is as tabulated in table 6 below.

Table 6: Nemenyi's test for High dimensionality reduction methods for Gene Expression data.

	t-SNE	LLE	KPCA	ISOMAP
LLE	0.9999	-	-	-
KPCA	0.1416	0.1416	-	-
ISOMAP	0.9999	0.9999	0.0893	-
UMAP	0.0082	0.0082	0.3173	0.0048

The p-value of 0.9999 for the comparisons between t-SNE and LLE and between t-SNE and ISOMAP respectively indicates that there is no significant difference between the performances of the methods. The comparison between LLE and ISOMAP with p = 0.999 showed that the methods are significantly different in reducing the dimensionality of the data.. Also, there exists no significant difference between the performance t-SNE and KPCA and between LLE and KPCA both with p = 0.1416. UMAP exhibits significant difference compared to ISOMAP, LLE and t-SNE with p = 0.0048, 0.0082 and 0.0082 respectively. In contrast, to the performance of UMAP when compared to ISOMAP, LLE and t-SNE, there is no significant difference between its performance when compared to KPCA with p = 0.3173. Furthermore the Critical Difference (CD) diagram for the above comparisons is as shown in figure 3.

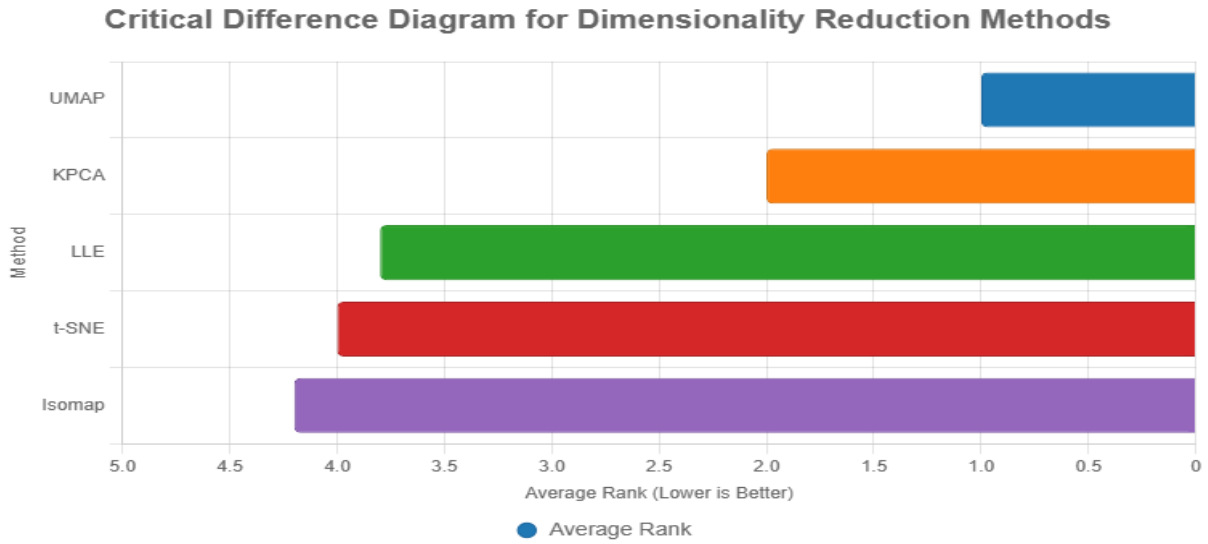

Figure 3: Critical Difference (CD) Diagram for High dimensionality reduction methods for Gene Expression RNA data.

Analysis of Nigerian Economic Data

In order to confirm that the Nigerian Economic data set was non linear, test for linearity was carried out on the dataset using the GAM and GLM as seen in table 7

Table 7: Model test for linearity of Nigeria Economic data.

Model	Explained Deviation	AIC	Ramsey's Reset p-value	F test p-value
GLM	0.6234	156.3	NA	0.1423
GAM	0.6789	152.1		

The Nigeria Economic data of 1990- 2022 show that GAM with an explained deviance of 0.6789 outperforms the GLM whose explained deviance is 0.6234. Also the AIC value for the GAM is 152.1 which is lower than that of GLM and is considered a sign of marginal improvement. The F test p-value (0.1423) also suggests non-linear terms. Dimensionality reduction was done using the five methods under study and their visualization is as shown in fig 4.

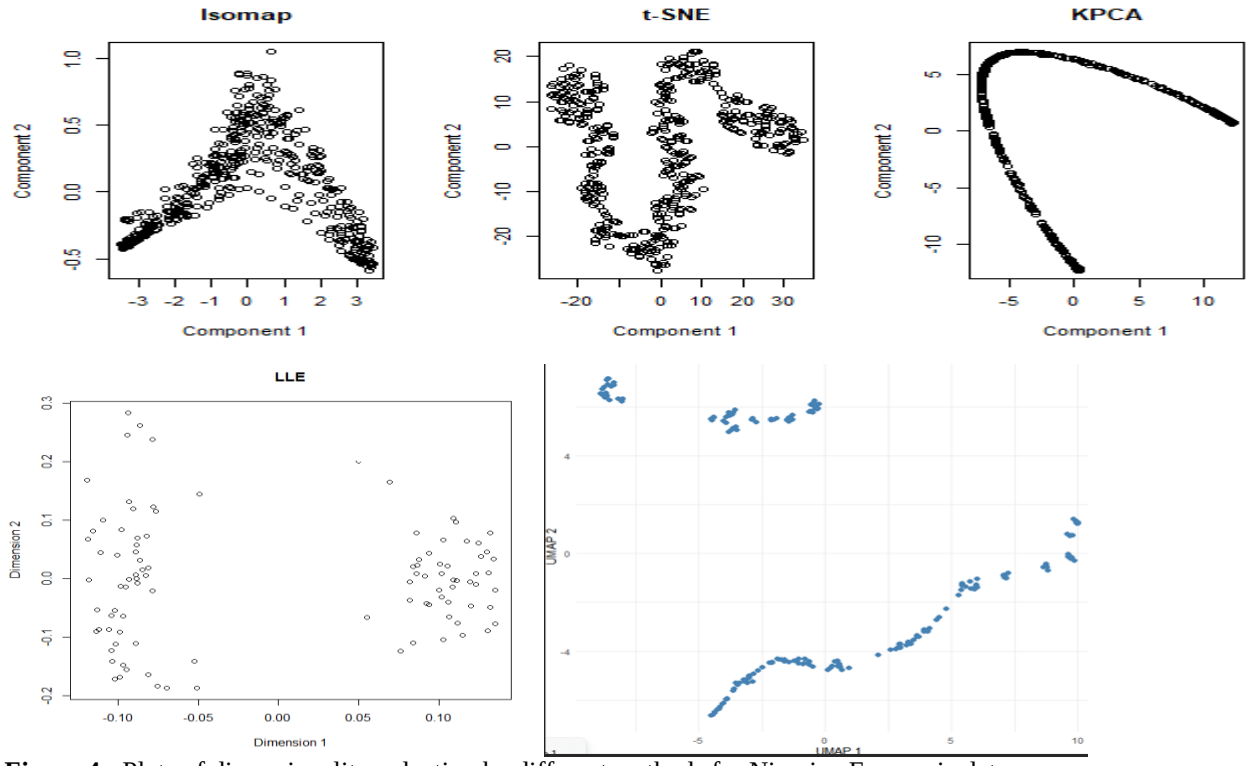

Figure 4: Plots of dimensionality reduction by different methods for Nigerian Economic data.

The dimensionality reduction for the data using the different methods for the metrics under computation is as tabulated in table 8

Table 8: High dimensionality reduction methods and the performance metrics for Nigeria Economic data.

Method / Metric	ISOMAP	t-SNE	UMAP	LLE	KPCA
MSE	0.127	0.134	0.102	0.121	0.109
Reconstruction Error	0.398	0.412	0.340	0.387	0.355
Correlation	0.770	0.765	0.802	0.752	0.788
Computation time (sec)	2.34	2.11	1.22	1.89	1.45
Trustworthiness	0.865	0.8802	0.890	0.855	0.882

For the dataset, Isomap was the slowest method as the computation time was 2.34secs compared to t-SNE, UMAP, LLE and KPCA which had computational times 2.11, 1.22, 1.89 and 1.45 respectively making UMAP the fastest. UMAP is the most trust worthy method of dimensionality reduction among other with trustworthiness of 0.890 and lowest MSE of 0.102 while LLE had the least trustworthiness coefficient of 0.855 compared to that of KPCA, Isomap, and t-SNE whose values were 0.883, 0.8802 and 0.412 respectively.

To ascertain if these values showed any significant statistical difference in the methods of dimensionality reduction, the Friedman's test was carried out under the hypothesis stated below:

H_0: There exists no significant difference in dimensionality reduction methods.

H_1: There is a significant difference in dimensionality reduction methods.

Table 9 below shows the ranks of the metrics for the different methods.

Table 9: Friedman's test ranking for High dimensionality reduction methods for Nigeria Economic data.

Metho / Metric rank	ISOMAP	t-SNE	UMAP	LLE	KPCA
MSE	4	5	1	3	2
Reconstruction Error	4	5	1	3	2
Correlation	3	4	1	5	2
Computation time (sec)	5	4	1	3	2
Trustworthiness	4	3	1	5	2
Total Rank Sum	20	21	5	19	10

The Friedman's test using the ranking of table 9 at df = 4 was $\chi^2 = 16.6$ with p- value = 0.0028 indicating a significant difference in the methods of dimensionality reduction. To ascertain which of these methods account for the significant difference, the Nemenyi's test was conducted and the output is as seen in table 10.

Table 10: Nemenyi's test for High dimensionality reduction methods for Nigeria Economic data.

	t-SNE	LLE	KPCA	ISOMAP
LLE	1.000	-	-	-
KPCA	0.266	0.373	-	-
ISOMAP	1.000	0.995	0.180	-
UMAP	0.023	0.041	0.855	0.012

The Nemenyi's post hoc test indicates that UMAP's performance differs significantly from that of t-SNE, LLE and Isomap with p-values of 0.023, 0.041 and 0.012 respectively. However, despite the UMAP ranking better KPCA, there is no significant difference in their performances with p-value = 0.855. Also, there is no significant difference between the performances of Isomap and LLE, Isomap and t-SNE, and Isomap and KPCA with p-values of 0.995, 1.000 and 0.180 respectively. The Critical Difference Diagram for the results is as seen in figure 5

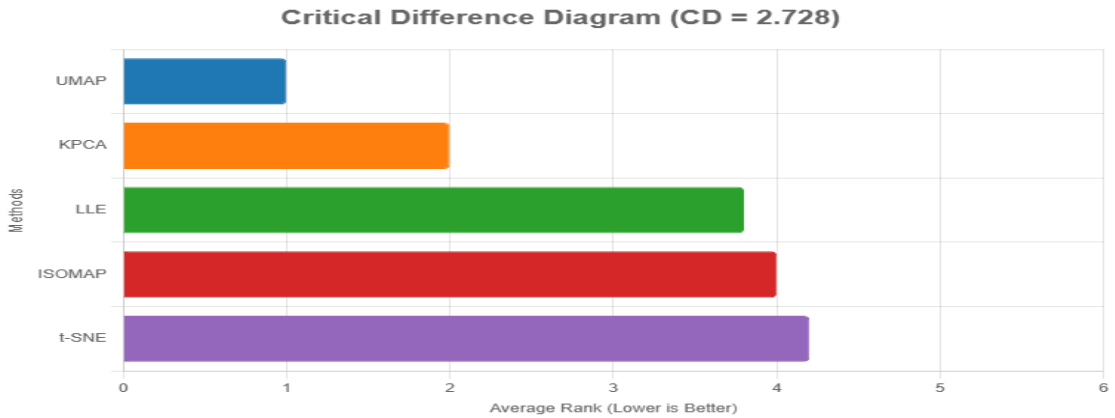

Figure 5: Critical Difference (CD) Diagram for High dimensionality reduction methods for Nigeria Economic data.

Analysis of Social Media Data

Social media data was also analyzed to test the performance of KPCA, LLE, UMAP, t-SNE and Isomap. The data was tested for linearity using the GAM and GLM which yielded the output of table 11

Table 11: Model test for linearity for Social Media data

Model	Explained Deviation	AIC	Ramsey's Reset p- value	F test p-value
GLM	0.3876	98765.2	0.0098	0.0065
GAM	0.5543	97643.8		

For the social media data, with regards to explained deviance, the GAM with value of 0.5543 outperforms the GLM (0.5543). This indicates that the values of the dataset are nonlinear. Also the GAM's lower AIC of 97643.8 as against GLM's AIC of 98765.2 supports the non-linearity claim on the data. The F test p-value of 0.0065 confirms that the GAM's non linear terms improve the fit of the data. Having ascertained that the dataset are non linear, dimensionality reduction was done using the five

methods under study. The visualizations of the performances of the methods are as seen in figure 6. The MSE, reconstruction error, correlation, trustworthiness and computation time of each of the methods is as shown in table 12.

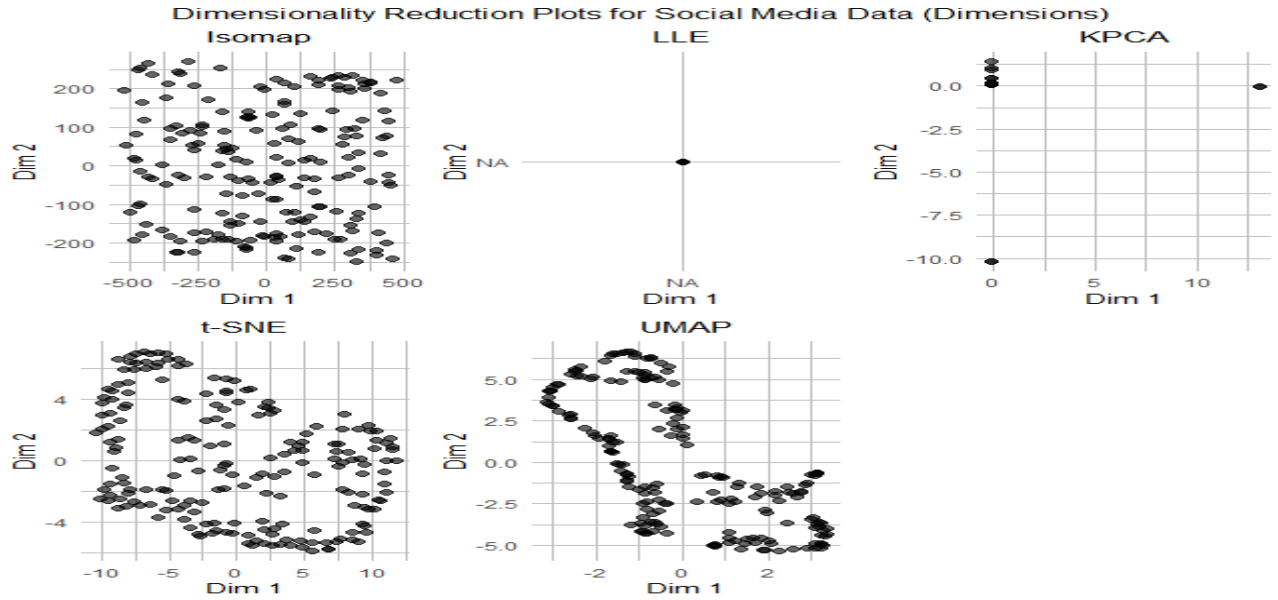

Figure 6: Plots of dimensionality reduction by different methods for Social Media.

Table 12: High dimensionality reduction methods and the performance metrics for Social Media data

Method / Metric	ISOMAP	t-SNE	UMAP	LLE	KPCA
MSE	0.107	0.112	0.085	0.098	0.090
Reconstruction Error	0.290	0.345	0.280	0.320	0.290
Correlation	0.805	0.810	0.841	0.795	0.828
Computation time (sec)	130.89	120.45	50.12	105.67	80.23
Trustworthiness	0.918	0.925	0.945	0.910	0.930

UMAP had the best performance across all metrics as its computation time of 50.12secs was the lowest. It also had the lowest MSE and reconstruction error of 0.085 and 0.280 respectively. Its trustworthiness and correlation were the highest amongst other at 0.945 and 0.841. This is closely followed by the values of KPCA, which had second lowest errors of 0.09 for MSE and 0.29 for reconstruction error. It is also fast with a computation time of 80.23, making it the second fast method. Its quality metric yielded 0.930 for trustworthiness and 0.828 for correlation making. In contrast to these, the Isomap was the least competitive method amongst the five methods as it was the slowest with a computation time of 130.89 seconds and higher errors of 0.335 for reconstruction error and 0.107 for MSE.

To check for the existence of significant difference in the performances of the 5 methods, the Friedman's test was carried out under the hypothesis below;
H_0: There is no significant difference in dimensionality reduction methods.
H_1: There is a significant difference in dimensionality reduction methods.

Table 13 below shows the ranks of the metrics for the different methods.

Table 13: Friedman's test ranking for High dimensionality reduction methods for Social Media data.

Method / Metric rank	ISOMAP	t-SNE	UMAP	LLE	KPCA
MSE	4	5	1	3	2
Reconstruction Error	4	5	1	3	2
Correlation	4	3	1	5	2
Computation time (sec)	5	4	1	3	2
Trustworthiness	4	3	1	5	2
Total Rank Sum	21	20	5	19	10

The Friedman's test using the ranking of table 12 at df = 4 was $\chi^2 = 16.6$ with p- value = 0.0028 indicating a significant difference in the methods of dimensionality reduction. To ascertain which of these methods account for the significant difference, the Nemenyi's test was conducted and the output is as seen in table 14.

Table 14: Nemenyi's test for High dimensionality reduction methods for Social Media data.

	t-SNE	LLE	KPCA	ISOMAP
LLE	0.999	-	-	-
KPCA	0.0200	0.300	-	-
ISOMAP	0.999	0.999	0.100	-
UMAP	0.030	0.045	0.999	0.020

The post hoc test using the Nemenyi's test as indicated in table 13 shows that UMAP differs significantly in performance from LLE, Isomap and t-SNE with p-values of 0.045, 0.020 and 0.30 respectively. No significant difference exists among the performance t-SNE, Isomap and LLE as their p-values were 0.999 for LLE vs t-SNE, Isomap vs t-SNE and Isomap vs LLE. Similarly, there was no notable significant difference between the performance of KPCA and UMAP with p-value = 0.999. This is further seen in the Critical Difference Diagram for the results as seen in figure 7

.

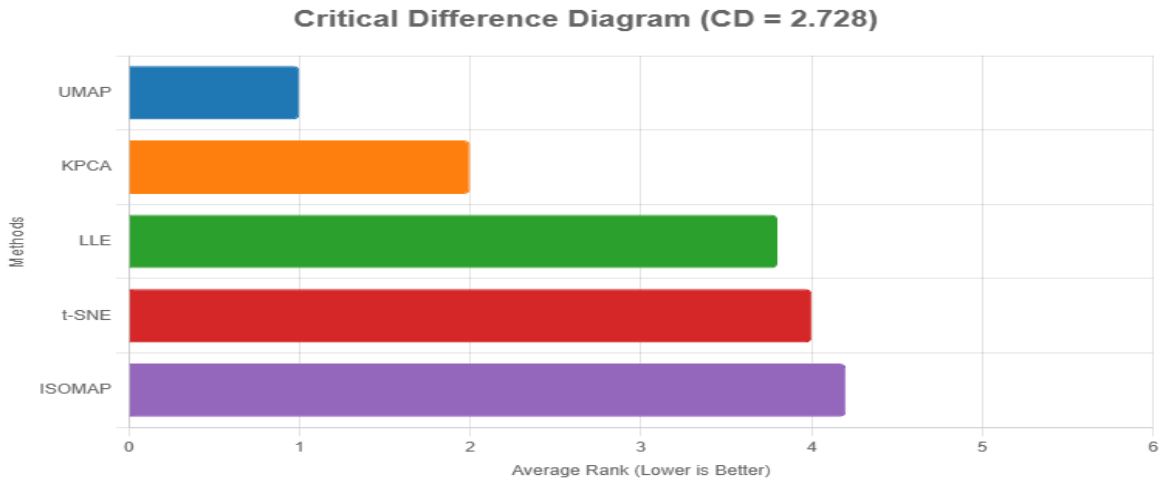

Figure 7: Critical Difference (CD) Diagram for High dimensionality reduction methods for Social Media data.

Discussion

This study was carried out with the major aim of evaluating and comparing non-parametric multivariate analysis methods in analyzing complex datasets Four datasets comprising simulated data, gene expression RNA on cancer data, Nigerian economic data from 1990 to 2022 and social media data on users opinion on trending issues. For the simulated data, high dimensional, non-linear multivariate datasets were generated and analyzed using five different non parametric dimensionality reduction methods; t-SNE, Isomap, LLE, UMAP and KPCA. The metrics of evaluation were computation time, correlation, trustworthiness, reconstruction error and Mean Square Error (MSE). For the simulated data, Isomap was the fastest among the methods with computation time of 0.026secs. It also had the highest correlation of 0.99996. However, it had lower errors of 212.1359 (reconstruction error) and 528914.359 (MSE). UMAP indicated superior data representation accuracy as it has trustworthiness of 0.9993, MSE of 6.936 and reconstruction error of 2.0697. KPCA offers balanced performance, with strong MSE (400.1260), Reconstruction Error (15.62), and Trustworthiness (0.9801), and moderate Computation Time (2.10 seconds). LLE leads in Trustworthiness (0.99994) and performs well in MSE (576.372) and Reconstruction Error (20.1734), but its Correlation (0.8306) is the lowest. t-SNE performs poorly in MSE (192856363.239), Reconstruction Error (356431.62), and Trustworthiness (0.8789), despite a strong Correlation (0.9880). Friedman's test was employed to check for the existence of significant difference in the performance of the methods and the Friedman's test statistic was χ^2 = 2.08 with p-value = 0.721 indicating that no statistically significant difference occurred in overall performance of the methods, suggesting that no method consistently outperforms the others across all metrics.

The gene expression RNA cancer data was first tested for linearity by comparing the GAM vs the GLM. The test for linearity shows that the GAM outperforms the GLM significantly in terms of the deviation explained and AIC with values of 58.76% against 41.23% and 2312.4 and 2456.7 respectively. For the Ramsey's reset test, p=0.0132 points to a misspecification of the GLM which is likely due to unmodeled non-linear effect. The F-test (p = 0.0087) further confirms that the GAM's non-linear terms provide a statistically significant improvement in fit. UMAP is seen to have the least computation time of 15.67secs making it the fastest method. It also has the least MSE reconstruction error of 0.081 and 0.265 respectively. Furthermore, its correlation coefficient of 0.849 and trustworthiness of 0.935 are the highest. Followed closely is the KPCA method which is considered relatively faster with 22.77secs and low errors of 0.275 for reconstruction error and 0.087 for MSE. Its quality

matrices of 0.834 for correlation and 0.921 for trustworthiness are also relatively better compared to that of ISOMAP, LLE and t-SNE. ISOMAP was the slowest with computation time of 50.33secs and had the higher errors of 0.310 and 0.099 for Reconstruction error and MSE respectively. The Friedman's $\chi^2 = 16.6$ at df = 4 and a p-value of 0.0028 indicating a significant difference in the different methods of reducing dimensionality. To ascertain which of the methods differ significantly, the Nemenyi's post hoc test was done. The p-value of 0.9999 for the comparisons between t-SNE and LLE and between t-SNE and ISOMAP respectively indicates that there is no significant difference between the performances of the methods. The comparison between LLE and ISOMAP with p = 0.999 showed that the methods are not significantly different in reducing the dimensionality of the data. Also, there exists no significant difference between the performance t-SNE and KPCA and between LLE and KPCA both with p = 0.1416. UMAP exhibits significant difference compared to ISOMAP, LLE and t-SNE with p = 0.0048, 0.0082 and 0.0082 respectively. In contrast, to the performance of UMAP when compared to ISOMAP, LLE and t-SNE, there is no significant difference between its performance when compared to KPCA with p = 0.3173.

The Nigeria Economic data of 1990- 2022 show that GAM with an explained deviance of 0.6789 outperforms the GLM who's explained deviance is 0.6234. Also the AIC value for the GAM is 152.1 which is lower than that of GLM and is considered a sign of marginal improvement. The F test p-value (0.1423) also suggests non-linear terms. Dimensionality reduction on the Nigerian Economic data using the five methods indicated that Isomap was the slowest method as the computation time was 2.34secs compared to t-SNE, UMAP, LLE and KPCA which had computational times 2.11, 1.22, 1.89 and 1.45 respectively making UMAP the fastest. UMAP is the most trust worthy method of dimensionality reduction among other with trustworthiness of 0.890 and lowest MSE of 0.102 while LLE had the least trustworthiness coefficient of 0.855 compared to that of KPCA, Isomap, and t-SNE whose values were 0.883, 0.8802 and 0.412 respectively. The Nemenyi's post hoc test indicates that UMAP's performance differs significantly from that of t-SNE, LLE and Isomap with p-values of 0.023, 0.041 and 0.012 respectively. However, despite the UMAP ranking better KPCA, there is no significant difference in their performances with p-value = 0.855. Also, there is no significant difference between the performances of Isomap and LLE, Isomap and t-SNE, and Isomap and KPCA with p-values of 0.995, 1.000 and 0.180 respectively.

Social media data was also analyzed to test the performance of KPCA, LLE, UMAP, t-SNE and Isomap. The data was tested for linearity using the GAM and GLM. The explained deviance for GAM was 0.5543 which outperforms the GLM (0.5543) indicating that the values of the dataset are nonlinear. Also the GAM's lower AIC of 97643.8 as against GLM's AIC of 98765.2 supports the non-linearity claim on the data. The F test p-value of 0.0065 confirms that the GAM's non linear terms improve the fit of the data. Having ascertained that the dataset are non linear, dimensionality reduction was done using the five methods under study. UMAP had the best performance across all metrics as its computation time of 50.12secs was the lowest. It also had the lowest MSE and reconstruction error of 0.085 and 0.280 respectively. Its trustworthiness and correlation were the highest amongst other at 0.945 and 0.841. This is closely followed by the values of KPCA, which had second lowest errors of 0.09 for MSE and 0.29 for reconstruction error. It is also fast with a computation time of 80.23, making it the second fast method. Its quality metric yielded 0.930 for trustworthiness and 0.828 for correlation making. In contrast to these, the Isomap was the least competitive method amongst the five methods as it was the slowest with a computation time of 130.89 seconds and higher errors of 0.335 for reconstruction error and 0.107 for MSE. The Friedman's test using the ranking of table 4.12 at df = 4 was $\chi^2 = 16.6$ with p- value = 0.0028 indicating a significant difference in the methods of dimensionality reduction. To ascertain which of these methods account for the significant difference, the Nemenyi's test was conducted and the result showed that UMAP differs significantly in performance from LLE, Isomap and t-SNE with p-values of 0.045, 0.020 and 0.30 respectively. No significant difference exists among the performance t-SNE, Isomap and LLE as their p-values were 0.999 for LLE vs t-SNE, Isomap vs t-SNE and Isomap vs LLE. Similarly, there was no notable significant difference between the performance of KPCA and UMAP with p-value = 0.999.

Conclusion

This study evaluated and compared non-parametric multivariate analysis methods in analyzing complex datasets which were non linear and high dimensional. Five dimensional reduction methods comprising of t-SNE, KPCA, Isomap, LLE and UMAP were employed to four different datasets. For the real life datasets, UMAP was consistently the best in preserving the local structure of the data sets and speed as it had the overall best trustworthiness, correlation with minimum error terms. This is closely followed by KPCA, which had a relatively high speed and higher correlation and trustworthiness when compared to LLE, Isomap and t-SNE. The Friedman's test for the real life data set showed that there is a significant difference in the performance of the methods and this difference according to the Nemenyi's test exists between UMAP and LLE, UMAP and Isomap, and t-SNE and UMAP. Between the two best performing methods; KPCA and UMAP, the Friedman's test proved that there is no significant difference between their performances. In contrast, there is no significant difference in the performances of LLE, t-SNE, and Isomap. However, the analysis of the simulated data showed that there is no significant difference in the performance of the different methods.

References

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computers in Biology and Medicine*, 132, Article 104307. https://doi.org/10.1016/j.compbiomed.2021.104307

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373–1396. https://doi.org/10.1162/089976603321780317

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications*, *19*(1), 1–13. https://doi.org/10.1007/s00521-009-0295-6

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Field, A. (2018). *Discovering statistics using R* (2nd ed.). SAGE Publications.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hazra, A., & Gogtay, N. (2017). Biostatistics series module 10: Brief overview of multivariate methods. *Indian Journal of Dermatology*, *62*(4), 358–366. https://doi.org/10.4103/ijd.IJD_296_17

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1), Article 5416. https://doi.org/10.1038/s41467-019-13056-x

Lee, J., & Kim, S. (2020). Nonlinear dimensionality reduction for damage detection using ultrasonic wave data. *Mechanical Systems and Signal Processing*, 131, Article 107356. https://doi.org/10.1016/j.ymssp.2019.107356

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2020). Feature selection: A data perspective. *ACM Computing Surveys*, *50*(6), Article 94. https://doi.org/10.1145/3136625

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9780429492259

van der Maaten, L., Postma, E., & van den Herik, J. (2009). Dimensionality reduction: A comparative review [Technical report]. Tilburg University. https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf

Venna, J., & Kaski, S. (2001). Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks — ICANN 2001* (pp. 485–491). Springer. https://doi.org/10.1007/3-540-44668-0_68

•*Thank you for publishing with us.*