EXPLORING DIMENSIONALITY REDUCTION TECHNIQUES: A MULTIFACETED ANALYSIS OF PCA, t-SNE, AND UMAP ON DIVERSE DATASETS

Pwint Phyu Khine [1], Thet Thet Hlaing[2] & Soe Mya Mya Aye[3]

Abstract

As the amount of data volumes grows, the increasing dimensionality of datasets poses significant challenges for data analysis and interpretation. Dimensionality reduction techniques play a critical role by selecting or synthesizing essential dimensions to manage heterogeneous data regardless of structures (structured or unstructured) effectively. This research explores various dimensionality reduction methods, applied to multiple datasets, to enhance analytical insights and visualization during the exploratory phase.

Keywords: Dimensionality Reduction, Machine Learning, Heterogeneous Data, Big Data, PCA, tSNE, UMAP

Introduction

Dimensionality reduction is the process of projecting high-dimensional data into a lower-dimensional subspace, capturing its essential features to alleviate the "curse of dimensionality" (Lespinats et al., 2022). This technique is essential for improving machine learning model performance by simplifying the dataset and enabling efficient handling of high-dimensional spaces (Bingham & Mannila, 2001). While dimensionality reduction facilitates data analysis, careful consideration is necessary to preserve the inherent characteristics of both structured and unstructured data. Structured data, with defined attributes, presents unique challenges in feature selection, especially for large-scale datasets. Unstructured data such as images and time series, in contrast, brings complexities due to its high dimensionality and varied formats. The curse of dimensionality arises because high-dimensional data often lacks structure, making traditional clustering and organizational methods less effective. To address this, dimensionality reduction methods condense data into the most meaningful components, which allows for more robust generalization, improved model efficiency, and reduced redundancy and noise. In applications, reducing data dimensions enables faster processing, better classification in image analysis, and more insightful data interpretation in healthcare. Ultimately, dimensionality reduction offers a systematic way to address the challenges of heterogeneous data, facilitating stronger insights and decisions.

Theoretical Background

Curse of Dimensionality

As dimensionality increases, the volume of the data space expands rapidly, resulting in “data sparsity”. Data points are spread thinly across space which is similar to objects such as planets, satellites, rocks, etc. floating in literal space. A minuscule difference in direction will make a moving vehicle such as a rocket change its direction, landing on a foreign planet instead

[1] Department of Computer Studies, University of Yangon
[2] Department of Computer Studies, University of Yangon
[3] Department of Computer Studies, University of Yangon

of earth. This sparsity means that effective data analysis requires exponentially larger datasets to provide reliable results. In many scenarios, each dimension variable can assume several values leading to a large number of possible combinations. This phenomenon is called "combinatorial explosion". As dimensions increase, the sampling volume grows exponentially, creating additional complexity. Each added dimension drastically increases the size of the sampling space, meaning far more points are required to maintain the same correctness, intensifying the effects of both combinatorial explosion and data sparsity.

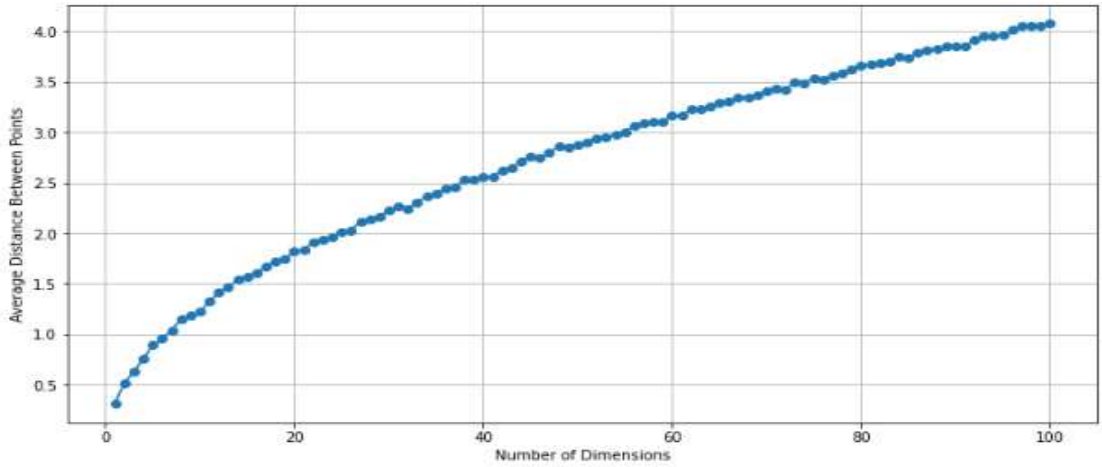


Figure 1: Curve of Dimensionality: Average Distance Between Points as Dimensions Increase leading to data sparsity and combinational explosion

Types of Dimensionality Reduction Methods

Dimensionality reduction techniques are commonly categorized as linear or non-linear and can also be grouped by computational approach, including matrix-based, manifold learning, and deep learning methods. Each approach offers specific methods tailored to different data structures and analysis needs. Matrix-based techniques, like Principal Component Analysis (PCA), capture maximum variance by projecting high-dimensional data onto a lower-dimensional linear subspace (Jolliffe & Cadima, 2016). PCA is primarily suited for structured data but can be adapted for non-linear data through Kernel PCA, which applies kernel functions to map data with complex relationships. Manifold learning techniques, such as t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), address non-linear relationships by assuming data lies on a lower-dimensional manifold within high-dimensional space. t-SNE emphasizes local data clusters (Hinton & Roweis, 2002; Maaten & Hinton, 2008), while UMAP captures both local and global structures (McInnes et al., 2018). Deep learning methods, particularly Autoencoders, use neural networks to create a latent representation, capturing non-linear features. Variations like Stacked Autoencoders and Deep Belief Networks (DBNs), including Restricted Boltzmann Machines (RBMs), further enhance the ability to capture complex patterns, making them effective for high-dimensional, unstructured data, such as images and audio. These methods—PCA, t-SNE, UMAP, and Autoencoders—are feature projection techniques, producing new, informative features in a lower-dimensional space, facilitating Exploratory Data analysis (EDA), visualization, and modeling. To establish a robust basis for implementing dimensionality reduction techniques, this section explores the theoretical foundations of PCA, t-SNE, and

UMAP. A thorough understanding of these methods is essential to justify their application and effectiveness in the context of this research.

Principal Component Analysis (PCA)

For data X with n number of samples and p features, PCA is by first centering the data by subtracting the mean of each feature from the corresponding feature values.

The covariance matrix ∑ of the centered data is calculated to understand the relationships between the features.

$$X_{\text{centered}} = X - \text{mean}(X)$$

$$\Sigma = \frac{1}{n} X_{\text{centered}}^{T} X_{\text{centered}}$$

Eigenvalue decomposition is performed to extract eigenvalues λi and eigenvectors v_i of the covariance matrix ∑ for maximum variance.

$$\Sigma \mathbf{v}_i = \lambda_i \mathbf{v}_i$$

Sorting and selecting the components to give the required top k eigenvectors in a new feature space is performed where V_k is the matrix with the top k eigenvectors as columns, and Z represents the transformed reduced data. The selected PC are then projected for visualization.

$$Z = X_{\text{centered}} \mathbf{V}_k$$

Incremental PCA is used for large memory footage datasets i.e. digit datasets. IPCA is good for big data as it performs the above-mentioned PCA step incrementally based on available memory space.

t-Stochastic Neighbor Embedding (t-SNE)

t-SNE emphasizes preserving the local structure by placing similar data points closer together in lower dimensions, making it highly effective for visualizing clusters. The t-SNE model transformed into a two-dimensional space **c**alculates the Pairwise Affinities in high dimension. The conditional probability of affinity $p_{j|i}$ is calculated for data points x_i and x_j where variance parameter σ controls the perplexity i.e. number of neighbors:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

tSNE got its name from the use of Student's t-distribution for the similarity of points in a low dimensional plane. The data point in the low dimensional plane is represented as q_{ij}.

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

tSNE finds the embeddings y_i that minimizes the divergence between high dimensional distribution P and low dimensional distribution Q by using Kullback-Leibler (KL) divergence.

$$\mathrm{KL}(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Various perplexity values were tested (e.g., 5 to 50) to evaluate their impact on the embedding quality. The denser dataset will require more perplexity however it comes with a longer time in computation.

Uniform Manifold Approximation and Projection (UMAP)

Manifold learning techniques assume that high-dimensional data lies on a lower-dimensional, curved manifold within a higher-dimensional space. While t-SNE emphasizes preserving local structures, UMAP balances both local and global structures, making it effective for visualizing complex data patterns, especially in clustering. UMAP constructs a weighted graph in high dimensions, capturing inherent structure through a fuzzy topological model, which is then projected to a lower-dimensional space. Built on Riemannian geometry and topological data analysis, UMAP uses fuzzy simplicial sets to model data structure, enabling computational efficiency and meaningful, interpretable embeddings. This robustness makes UMAP widely applicable in fields like bioinformatics, NLP, and image processing, where both cluster separation and structure preservation are essential.

Various neighborhood sizes (n_neighbors) and minimum distances (min_dist) were tested to optimize both the granularity of local structure and the extent of global structure preservation in the embedding. For each data point x_i, find its k-nearest neighbors and calculate the edge weights using a kernel function. This is based on the distance between points, transformed into a fuzzy set representation. The probability of connection p_{ij} is defined with p_i as the distance to the nearest neighbor of x_i and σ_i is used to fix the number of neighbors:

$$p_{ij} = \exp\left(-\frac{\|x_i - x_j\| - \rho_i}{\sigma_i}\right)$$

In the low-dimensional space, a similar fuzzy simplicial complex is built, initialized randomly or with spectral embedding. Optimizing the embedding in low-dimensional embedding is done by calculating the cross-entropy between the high-dimensional and low-dimensional fuzzy simplicial sets, where qij is the probability of connection in the low-dimensional space, dependent on the current embedding. By applying UMAP, Visualization for both distinct clusters and gradients within continuous datasets is more efficient.

$$C = \sum_{(i,j)} p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}}$$

These dimensionality reduction techniques, each tailored for different types of data and analysis objectives, allow for more efficient storage, processing, and visualization of high-dimensional datasets, improving computational efficiency and interpretability. These methods are particularly beneficial for **Exploratory Data Analysis** (EDA), especially in the **Visualization** stage for complex, high-dimensional data.

Implementation

Datasets

This study applies dimensionality reduction techniques to different datasets, demonstrating distinct behaviors based on the inherent nature of each dataset. Synthetics data is also generated to illustrate the overall operations using Swiss roll problem. Data points in higher dimension 3D plane is projected in 2D plane by using PCA, tSNE and UMAP. The resulted visualization shows whether they are possible for reconstruction, synthesis generation, and local structure preservation, and global data structure.

PCA is not well-suited for the Swiss Roll due to its inability to handle nonlinear data effectively. It shows moderate reconstruction error and fails to preserve local structure. t-SNE captures local structure better (high trustworthiness and continuity) but has high KL divergence, indicating difficulties with global distribution preservation as it fails to maintain global structure. UMAP also preserves local relationships well but slightly less than t-SNE. Its lower KL divergence and cross-entropy suggest a somewhat better balance between local and global structures. For this dataset, UMAP and t-SNE both provide good representations of local structure, making them preferable for visualizing continuous, nonlinear data like the Swiss Roll. In open box problem, PCA can maintain its almost entire global structure in lower dimension, and UMAP somehow can do it leading to possible potential for data compression and synthetic data generation. t-SNE is impossible in these areas although further researches should be done to investigate these cases.

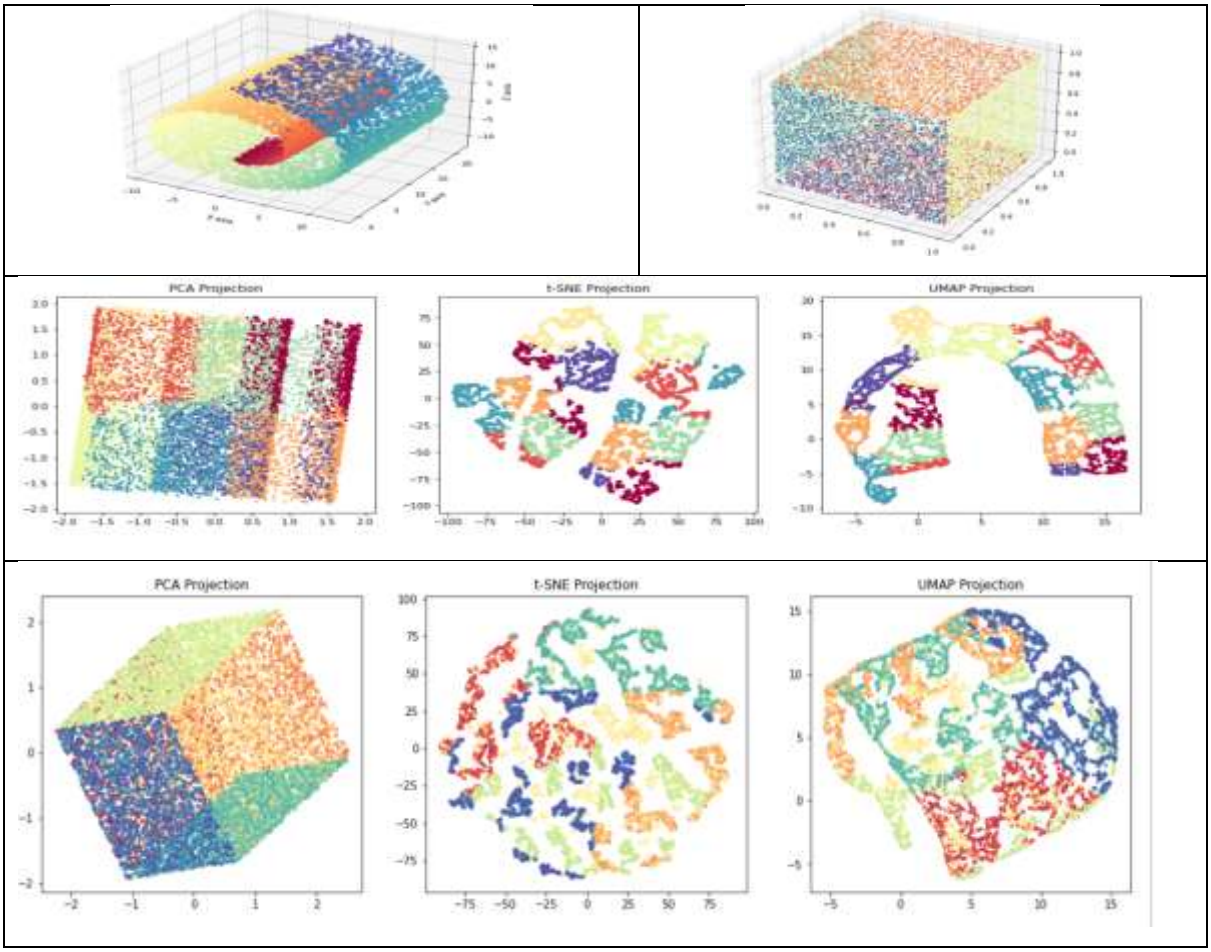
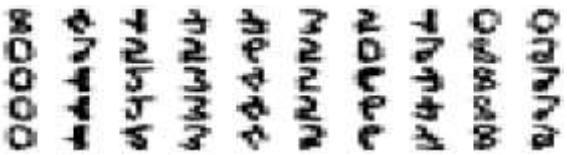


Figure 2. Swiss Roll Problem and Open Box problem projected into lower dimensional by using PCA, tSNE and UMAP

Various datasets from Python's sklearn library are used to showcase how dimensionality reduction can aid data analysis and visualization in structured and unstructured data contexts. Both structured and unstructured datasets from Python's sklearn library, highlight the unique behaviors of each. For structured data, the Breast Cancer Dataset includes 569 samples and 30 features derived from cell nuclei images, such as radius and texture, supporting predictive modeling for tumor malignancy (Malignant = 0, Benign = 1). The Diabetes Dataset features 768 samples with 8 health metrics like BMI and age, aiming to analyze disease progression through continuous target scores; higher values indicate more severe progression. Both datasets provide insight into how dimensionality reduction can improve data analysis and visualization in medical applications. For unstructured data, the Handwritten Digits Dataset includes 1,700 8x8 pixel images of digits (0-9), ideal for exploring image classification and high-dimensional data visualization. For big data, (mnist_784', version=1), contains 70,000 samples in total. This includes 60,000 training images and 10,000 testing images of handwritten digits are used alternatively to ensure some of the visualization results. These datasets demonstrate dimensionality reduction's utility in refining analysis and visualization for diverse data types.

(a) Cancer and Diabetics datasets (structured data)

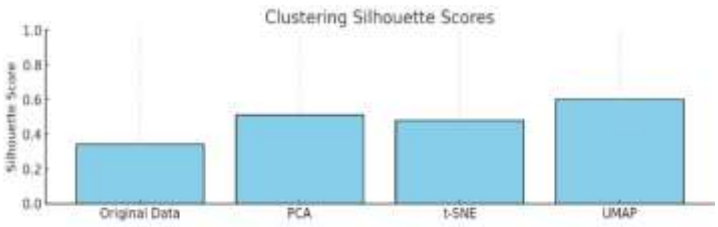


(b) Hand-written Digital Image Dataset (unstructured data)

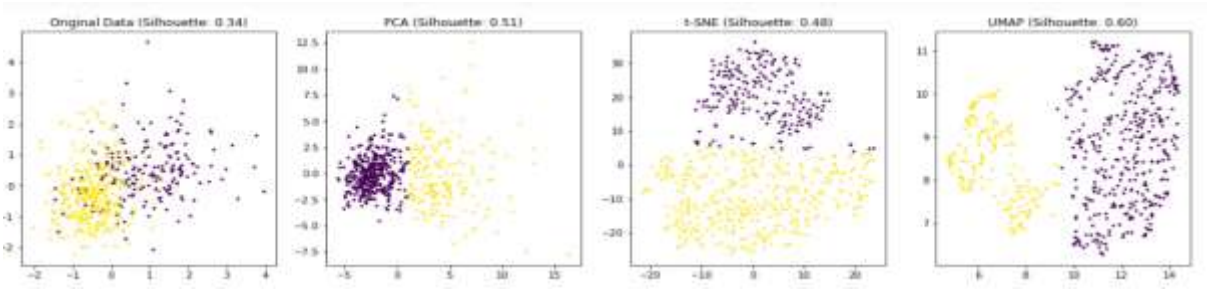
Figure 3. Diverse Datasets for structured and unstructured data

Analysis of datasets for PCA, tSNE, UMAP

The Cancer Dataset has two groups i.e. benign and serious. When visualized in a scatter plot, it is already clear. However, the steadily increased silhouette score on PCA, tSNE, and UMAP shows that the clustering ability increases with each method even without the hyperparameter tuning. Baseline cancer dataset has low silhouette score and the relationship between clusters could not be seen clearly. However, after applying dimensional reduction, the data points in lower dimension show better silhouette score (between somewhat good) in all three cases even without the expensive parameter tuning.



(a) Silhouette Scores for original data projection and dimensionality reduction methods



(b) Scatter plot visualization for original data projection and dimensionality reduction methods

Figure 4. Cancer Dataset with PCA, tSNE, and UMAP

The Diabetes Dataset has continuous values and is used for regression. When visualized in a scatter plot, the data points are unable to reveal any patterns in PCA. However, tSNE and UMAP with default values show two possible clusters.

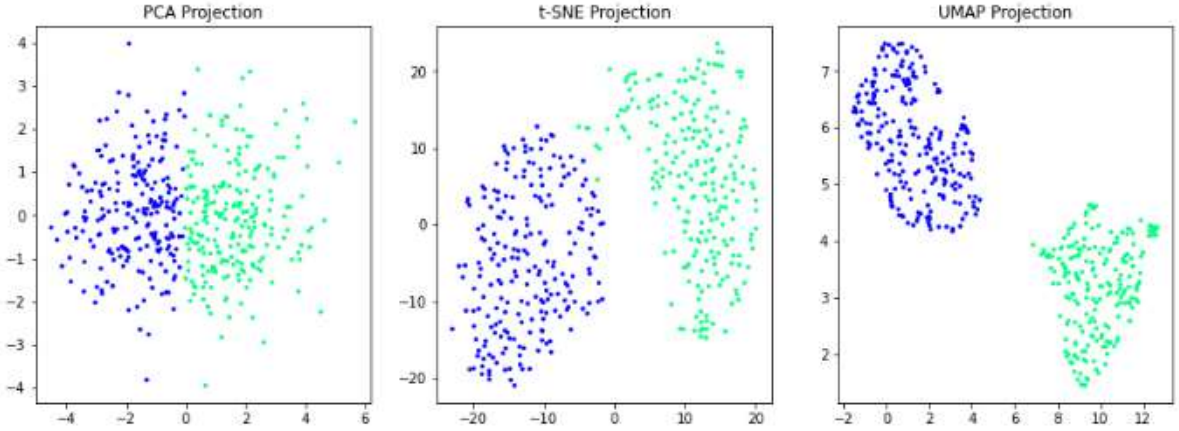


Figure 5. PCA, tSNE, and UMAP for Diabetes Dataset

The most prominent result comes in the unstructured image dataset MINST. PCA is unable to distinguish between clusters. However, both tSNE and UMAP show clear clusters in lower dimensions which greatly helped in further analysis processes.

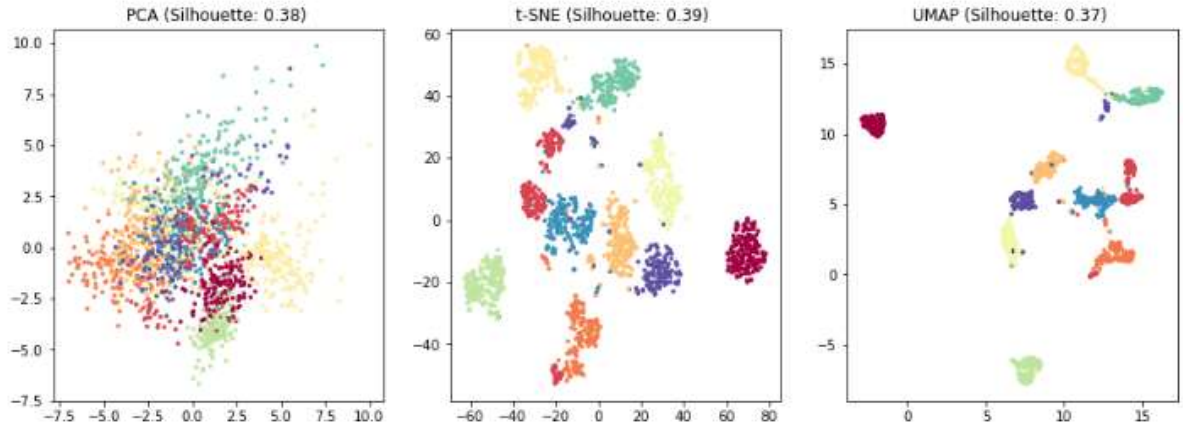


Figure 6. PCA, tSNE and UMAP on digit dataset

The other problem encountered is the volatile nature of tSNE. Based on the perplexity parameter, it generates different data at different parameter values. The silhouette scores vary in other methods such as UMAP but their visualization did not change much in lower dimensional

space showing that they are in stable condition. However, tSNE had an additional problem in unstability in the analysis process making them used only for visualization to understand patterns in only specified parameter values.

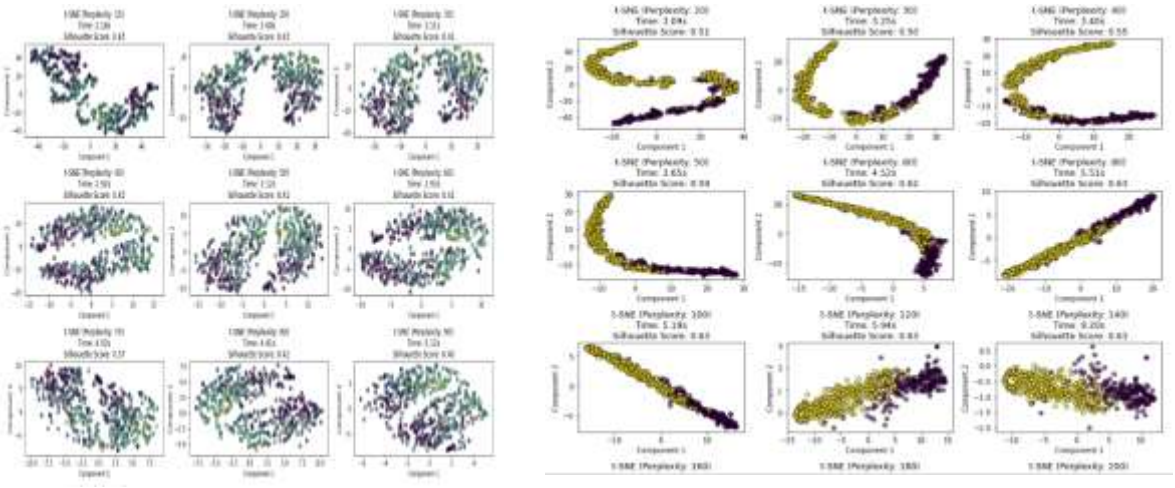


Figure 7. Volatile structure of diabetes dataset (left) and cancer dataset (right) using tSNE, leading to erroneous interpretation

In digit datasets, it creates false clusters when different parameter values are used as shown in the figure. Instead of 10 clusters, they create smaller erroneous clusters in visualization.

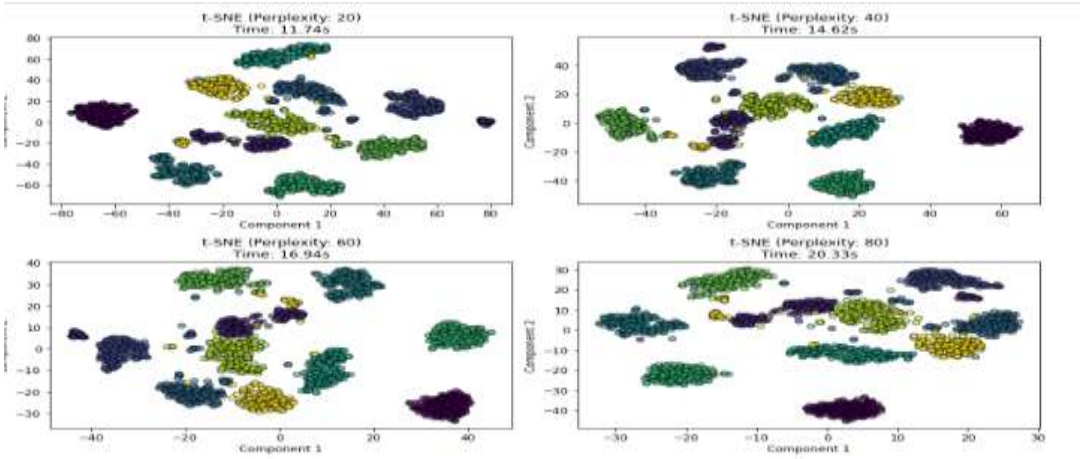


Figure 8. t-SNE creates false clusters in the digit dataset. Clusters are generated correctly or incorrectly based on different parameter values.

Result and Discussion

This paper explores the dimensionality reduction techniques on structured and unstructured data for preprocessing of machine learning datasets. Dimensionality reduction techniques such as PCA, t-SNE, and UMAP are crucial for visualizing high-dimensional data and facilitating subsequent analyses. Each method has unique characteristics, and their effectiveness requires specific evaluation metrics.

The "eye-rolling" approach is commonly adopted, where practitioners visually inspect various visualizations especially 2D or 3D scatter plots of the first two or three principal components. This qualitative evaluation allows researchers to identify patterns, clusters, or outliers without the need for formal quantitative metrics. Although traditional evaluation metrics

like silhouette scores are not typically employed with PCA, insights can still be drawn from the resulting data representations.t-SNE is a non-linear dimensionality reduction method specifically designed for visualizing high-dimensional data. Its effectiveness is often evaluated using quantitative metrics, with the silhouette score being a primary option. This score measures how similar an object is to its cluster compared to other clusters, ranging from -1 to 1, where higher values indicate better clustering. Similarly, UMAP is a non-linear technique that excels in preserving both local and global data structures, making it particularly effective for complex datasets. Evaluation metrics for UMAP include silhouette scores and other clustering metrics, such as the Davies-Bouldin index and adjusted Rand index. Silhouette score is used to have more understanding of data. For the Cancer Dataset, which consists of benign and malignant groups, scatter plot visualizations reveal clear separations. Notably, the silhouette scores steadily increase across PCA, t-SNE, and UMAP, indicating enhanced clustering ability with each method, even without hyperparameter tuning. In contrast, the Diabetes Dataset, characterized by continuous values and primarily used for regression, presents chal lenges in visualizing patterns. PCA visualizations fail to reveal any discernible patterns among data points. However, t-SNE and UMAP, using default parameters, uncover two potential clusters, demonstrating their capability to identify underlying structures within the data. In addition to structured datasets, t-SNE, and UMAP excel in analyzing unstructured image data, such as the MINST dataset. Here, PCA struggles to distinguish between clusters, while both t-SNE and UMAP effectively reveal clear clusters in lower dimensions, significantly aiding subsequent analytical processes.

Table 1. Comparison Analysis for PCA, tSNE, and UMAP

Properties	PCA	t-SNE	UMAP
Incremental Learning	Incremental PCA available	Stochastic Gradient Descent, Randomized	Can handle well in its own
Outlier Handling	Result affected	Handles well	Handles well
Computational Complexity	$O(d^2n+d^3)$	$O(n^2)$	O(n log n)
Data Restoration	Yes	No	Somewhat
Hyperparameter	-	Perplexity (neighbors within a cluster)	Number of Neighbors, Min Distance
Structured Data	Maintains global structure	Often loses global structure	Preserves both local and global structure (to an extent)

PCA is one of the most robust techniques used in dimensionality reduction. PCA reduces the dataset to two principal components while retaining as much variance as possible. As they maintain the data structure, it is easier to reconstruct data, therefore, PCA is good at the restoration of original data. TSNE is very volatile as the visualization changed with each perplexity score for breast cancer, diabetes, and digit dataset. It even creates false clusters in digits. t-SNE optimizes the dataset's structure in lower dimensions using probability distributions, allowing for clear clustering, especially useful in high-dimensional data. However, their volatile nature makes them unsuitable for the reconstruction of original data. UMAP somehow maintains both local and global structures, making it effective for visualizing datasets

with complex structures. Dimensionality Reduction is a necessary tool which could be often in machine learning procedures before data are exported for machine learning algorithms or when machine learning algorithms are required to reduce dimension to increase their accuracy in classification or make them more defined in clustering. To perform effective analytics. DM methods will be required to apply as necessary in any machine learning process pipelines. Different Dimensionality Reduction Techniques are applied for heterogenous data – PCA, tSNE, and UMAP on structured data (breast cancer, diabetics) and (handwritten digit).

By employing these dimensionality reduction techniques, researchers can enhance their understanding of data structures and improve the robustness and reliability of their findings in various analytical scenarios. Overall, the choice of evaluation metrics should align with the specific objectives of the analysis and the characteristics of the dataset.

Conclusion and Future Work

Dimensionality reduction techniques are essential for feature extraction, data analysis, and visualization, simplifying high-dimensional datasets into manageable representations. This transformation reveals hidden patterns, enhances model interpretability, and boosts overall performance. As data scales in volume and complexity, effective dimensionality reduction becomes increasingly critical for efficient analysis. Future research will prioritize deep learning-based dimensionality reduction, particularly through autoencoders, which enable learning complex data representations useful for tasks like anomaly detection, classification, and generative modeling. Applying these techniques to big data will also streamline processing and facilitate the extraction of valuable insights from large-scale datasets. Advances in dimensionality reduction are anticipated to drive progress across machine learning, AI, and data science. This research examines dimensionality reduction challenges across diverse datasets, promoting a deeper understanding of methods for handling complex data structures in real-world contexts.

Acknowledgments

I would like to thank my supervisor Dr. Thet Thet Hlaing (Professor), Dr. Soe Mya Mya Aye (Professor and Head), Department of Computer Studies, University of Yangon for their valuable guidelines in my research. I also would like to extend my gratitude to Dr. Wint Pa Pa Kyaw (Professor), and Dr. Khin Sandar Myint (Associate Professor), Department of Computer Studies, University of Yangon for their help given to me in the research.

References

Bingham, E.& Mannila, H. (2001). *Random projection in dimensionality reduction*. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining – KDD '01. p. 245. doi:10.1145/502512.502546. ISBN 978-1581133912. S2CID 1854295

Hinton, G. & Roweis, S. (2002). *Stochastic neighbor embedding*. Neural Information Processing Systems.

Jolliffe, I.T. & Cadima, J. (2016), *Principal component analysis: a review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 374 (2065): 20150202. Bibcode: 2016RSPTA.37450202J. doi:10.1098/ rsta.2015. 0202. PMC 4792409. PMID 26953178

Lespinats, S., Colange, B., & Dutykh, D. (2022). Nonlinear dimensionality reduction techniques: A data structure preservation approach. ISBN: 978-3-030-81025-2.

Maaten, L.V.D. & Hinton, G. (2008), *Visualizing Data Using t-SNE*. Journal of Machine Learning Research. 9: 2579–2605

McInnes, L., Healy, J. & Melville, J. (2018), *Uniform manifold approximation and projection for dimension reduction*. arXiv:1802.03426