

Exploring Dimensionality Reduction Methods for Nonlinear Data: A Performance-Based Study

Soniya Patil¹, Parinitha Oruganti², Sonal Ullegaddi³, Mrs. Pooja Chandargi⁴ and Mrs. Renuka M Ganiger⁵

¹⁻³Student 3rd year, KLE Technological University, Hubli, Karnataka, India

⁴⁻⁵Assistant Professor, KLE Technological University, Hubli, Karnataka, India

Email: 01fe22bee104@kletech.ac.in, 01fe22bee101@kletech.ac.in, 01fe22bee114@kletech.ac.in,
pooja.appanavar@kletech.ac.in, renuka.ganiger@kletech.ac.in

Abstract—Dimensionality reduction plays a pivotal role in handling high-dimensional datasets by mitigating computational inefficiency, redundancy, and overfitting while preserving critical data structure. This paper evaluates three prominent dimensionality reduction techniques—Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Autoencoders—on a breast cancer dataset sourced from the SEER program. The study compares their performance across key metrics, including computational efficiency, reconstruction error, and structural preservation.

Index Terms—(dimensionality reduction, PCA, t-SNE, Autoencoders, Computation cost, Efficiency, Comparison).

I. INTRODUCTION

Any given dataset has inherent characteristics that define its structure, behaviour and relationships among its elements. Dimensionality is one such aspect that greatly affects the data representation.

Dimensionality indicates the number of samples, features or variables available in the dataset. Each feature conforms to a measurable entity or characteristic of the data.

Real world datasets are massive in size and spread across various domains such as finance, social media, healthcare, academic literature, image processing etc. The huge number of features within these datasets, likely in thousands or millions, make them high dimensional.

High dimensional datasets bring along with them concerns and constraints as they are complicated to work with due to their sheer size, speed of data generation, data processing as well as variety of information. High-dimensional data increases computational costs for machine learning algorithms. In high-dimensional spaces, data points tend to become sparse, making it difficult to visualize and detect meaningful patterns or relationships. Datasets like these require more storage spaces and tend to have noise or redundant features that lead to overfitting in the models making it necessary to address the problems by reducing these dimensions. Dimensionality reduction directly tackles these issues by transforming data into a

lower-dimensional space without significant information loss.

These techniques can broadly be categorized into linear and non-linear methods, each with unique strengths and limitations. Classical methods like Principal Component Analysis (PCA) excel in handling linearly separable data, while modern algorithms such as t-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are adept at uncovering complex, non-linear patterns in data.

This research paper aims to provide a comprehensive analysis of various dimensionality reduction techniques and comparing their performance based on some parameters. It investigates the trade-offs between computational efficiency, scalability, interpretability, and accuracy, with the objective of identifying optimal methods for specific use cases.

II. LITERATURE REVIEW

Dimensionality reduction addresses the challenges posed by high-dimensional datasets, such as computational inefficiency and the curse of dimensionality. Techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are commonly applied to improve model performance and data interpretability.[1][2]

PCA is a widely used linear dimensionality reduction method that transforms high-dimensional data into a lower-dimensional space by identifying directions of maximum variance. Its effectiveness in medical datasets, such as liver disease detection, has been demonstrated. PCA provides computational efficiency and noise reduction, making it suitable for large-scale datasets. However, its limitations include the inability to capture non-linear relationships in data [1][3].

PCA's advantages include its unsupervised nature and ability to reduce redundant features and noise, but it can suffer from high information loss in certain scenarios[4][5]. Extensions like Distributed Parallel PCA (DP-PCA) address these limitations by improving computational efficiency and scalability, making it suitable for big data applications [4]. Unlike PCA, which is unsupervised, LDA is a supervised method that maximizes class separability. It projects data onto a subspace that optimally discriminates between classes. While effective for classification tasks, LDA struggles with datasets containing overlapping classes or limited samples [6][4].

t-SNE and Multidimensional Scaling (MDS): Both methods have been explored for their ability to preserve local and global structures in high-dimensional data. t-SNE is effective for visualization but is computationally expensive, while MDS focuses on preserving pairwise distances and is more suited for low-dimensional tasks. These methods have been applied to datasets such as Parkinson's and diabetes to enhance classification accuracy with algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM)[2][3].

t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are powerful for visualizing high-dimensional data. t-SNE excels in maintaining local structure, making it suitable for clustering and classification visualization, although it can be computationally intensive. UMAP, on the other hand, offers similar benefits with faster computation[4].

Autoencoders: Unlike PCA and t-SNE, autoencoders leverage deep learning to learn latent features and reduce dimensions through encoding-decoding mechanisms. Autoencoders outperform traditional techniques in terms of reconstruction error and robustness against noise. They have been effectively applied to large, feature-rich datasets, demonstrating superior performance in both linear and non-linear data structures[2].

RBM's(Restricted Boltzmann Machines) are another neural network-based technique used for unsupervised DR. They transform high-dimensional input into low-dimensional latent variables, preserving essential features for tasks like image recognition[6].

Several comparative studies demonstrate the strengths and weaknesses of these techniques in specific contexts. For instance, in micro-expression recognition, KPCA, t-SNE, and RBMs were evaluated using CASMEII datasets. The results highlighted that t-SNE produced superior low-dimensional representations for visualization, while KPCA excelled in classification tasks[6]. Another study using PCA and LDA on medical datasets revealed PCA's superior performance for high-dimensional datasets, while LDA was more effective for datasets with fewer features[5].

III. DATASET DESCRIPTION

The named dataset was attained from the Kaggle website. This dataset of cases with breast cancer was taken from the NCI's SEER Program, which offers data on population- grounded cancer statistics and was streamlined in November 2017. womanish cases with insinuating conduit and lobular melanoma bone cancer(foreseer primary cites recode NOS histology canons 8522/3), diagnosed between 2006 and 2010 were included in the dataset. Cases with uncertain excrescence size, delved indigenious LNs, positive indigenious LNs, and survival months of lower than one month were neglected; hence cases were ultimately included. The dataset provides a comprehensive overview of cancer cases' clinical, pathological, and demographic aspects and issues. Its primary focus appears to be determining correlations between multitudinous factors similar as excrescence features, receptor statuses, lymph knot involvement, and staging and case survival.

IV. METHODOLOGY

For the analysis, the Breast Cancer dataset has been used to evaluate and compare the performance of the following techniques:

1. Principal Component Analysis (PCA): A linear DR technique that projects data onto a lower-dimensional subspace, maximizing variance.
2. t-Distributed Stochastic Neighbor Embedding (t-SNE): A non-linear DR method optimized for visualizing high-dimensional data in two or three dimensions.
3. Autoencoders: A deep learning-based non-linear DR approach that encodes data into a compressed representation and decodes it back to reconstruct the original input.

A. Principal Component Analysis

Principal Component Analysis (PCA) is one of the most widely used techniques for dimensionality reduction, aiming to transform a dataset into a lower-dimensional space while retaining as much variance as possible. It does this by identifying the principal components, which are the directions of maximum variance in the data.

In a study[12], that involved data compression, PCA minimized storage and computational demands by eliminating redundancy and focusing on the most informative components, making it ideal for handling big data in industrial contexts.

It also serves as a valuable tool in feature engineering, where it enhances machine learning model performance by removing correlated and less significant features, ensuring more robust predictions. According to another study [13], in exploratory data analysis, PCA facilitated the visualization of high-dimensional datasets, helping researchers uncover patterns and relationships that might not have been immediately apparent.

PCA algorithm identifies the directions or principal components, that maximize the variance in the dataset. It does so by solving the eigenvalue decomposition of the covariance matrix or applying singular value decomposition (SVD).

1. Standardize the dataset.
2. Compute the covariance matrix of the data.
3. Perform eigenvalue decomposition or SVD.
4. Select the top-k principal components and project the data onto this subspace.

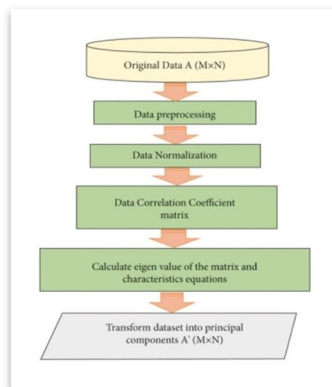


Figure 1.[14] Flowchart of PCA Algorithm

B. t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE), introduced by van der Maaten and Hinton in 2008, is a widely-used dimensionality reduction technique designed to visualize high-dimensional data in two or three dimensions. It builds upon Stochastic Neighbor Embedding (SNE) by addressing optimization challenges and resolving the "crowding problem" through the use of a Student-t distribution for low-dimensional similarities [8]. Its utility lies in revealing intricate relationships in datasets with manifold structures, as commonly observed in high-dimensional data like image features or single-cell RNA sequencing data [8][9].

In paper[8], it excelled in exploring relationships in datasets with non-linear structures, revealing insights that traditional linear methods might miss and by addressing issues like the crowding problem, t-SNE provides meaningful 2D or 3D maps even for data with manifold properties. The study [9] however, stated that although t-SNE is highly effective, it can be slow and inefficient when working with very large datasets because the

standard version does not scale well. Improved methods, like FFT-accelerated t-SNE, address this issue by making the process faster and more practical for widespread use.

t-SNE algorithm models the high-dimensional pairwise similarities using a Gaussian distribution and maps them to a low-dimensional space using a student t-distribution, minimizing the Kullback-Leibler divergence between the two distributions.

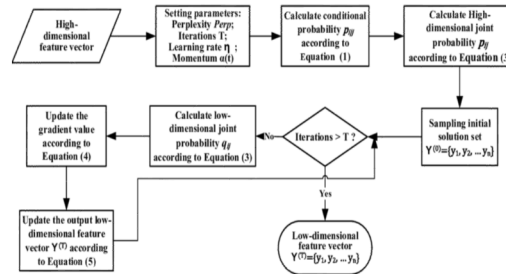


Figure 2.[15] t-SNE Algorithm

1. Compute pairwise similarities in the high-dimensional space.
2. Map these similarities to a low-dimensional space using t-distributions.
3. Iteratively optimize the embeddings using gradient descent to preserve neighborhood relationships.

C. Autoencoders

The use of autoencoders in handling high-dimensional and complex datasets has demonstrated transformative potential across various domains. In the financial sector, one study introduced a novel framework for clustering financial time series. Financial data often presents challenges due to its high dimensionality and non-linear relationships. By utilizing autoencoders as a dimensionality reduction technique, this framework compresses multi-dimensional financial metrics into lower-dimensional representations while retaining critical informational features.

This compression enabled traditional clustering algorithms, such as Agglomerative, BIRCH, and KNN, to achieve enhanced granularity and category isolation which in turn impacted predictive modelling, optimizing investment strategies, and facilitating risk management in the financial sector [10].

In medical imaging, a dual-featured autoencoder (DFAE) has been proposed for the classification of histological images of human lung tissues. This approach combined spectral and spatial feature extraction using the Persistent Homology Algorithm Toolbox (PHAT) and an image-super-resolution autoencoder.

When these compressed features are processed using a Triple Layered Convolutional Architecture (TLCA), the system achieved a classification accuracy of 96%. The method addressed key challenges in analyzing high-resolution histological images, offering advancements in diagnostic accuracy and efficiency[11].

Unlike traditional methods like PCA, autoencoders excel in capturing complex non-linear patterns, as demonstrated in their use for clustering financial time series, where they improved granularity and predictive accuracy[10]. Similarly, in histological image classification, autoencoders retained essential structural and textural information, achieving a remarkable 96% classification accuracy while mitigating overfitting risks[11]. Autoencoders flexibility across diverse domains and their capability to extract meaningful, compressed representations of high-dimensional data make them an efficient, and adaptable solution for improving the accuracy and efficiency of complex and high dimensional tasks. The algorithm follows the steps below:

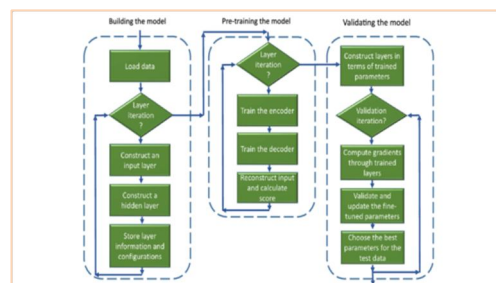


Figure 3.[15] Autoencoders Algorithm

1. Design an encoder network to compress data into a latent space.

2. Design a decoder network to reconstruct the original data from the latent space.
3. Train the network using backpropagation to minimize reconstruction error.

D. Procedure for Analysis

Dataset Preparation

- Load the Breast Cancer dataset and split it into training and testing subsets.
- Normalize or standardize the dataset to ensure compatibility with DR techniques.

Application of DR Techniques

1. PCA:

- a. Perform PCA and retain a predefined number of principal components.
- b. Record the explained variance ratio for each component.

t-SNE:

- c. Apply t-SNE with optimized hyper parameters.
- d. Project data into a 2D or 3D space for visualization.

Autoencoders

- e. Design and train an autoencoder with appropriate architecture.
- f. Use the encoder to obtain the latent representation.

Visualization

Visualize the reduced-dimensional data for each method using scatter plots in Python. Compare the spatial separability of different classes in the dataset.

Computation of Performance Metrics

Evaluate the techniques based on the following metrics:

1. Computational Efficiency: Measuring the execution time for training and transforming data.
2. Reconstruction Error: For PCA and autoencoders, compute the reconstruction error as the mean squared error between the original and reconstructed data.
3. Preservation of Data Structure: For t-SNE, evaluate how well local neighborhoods are preserved in the low-dimensional space.

Analysis

The study involved comparative analysis and evaluated PCA, t-SNE, and Autoencoders across performance metrics, including computational efficiency, reconstruction error, and cluster separability. A tabular summary presents key metrics, while bar plots or line graphs visually compare them. The trade-offs reveal PCA's superior computational efficiency, the stronger reconstruction accuracy of PCA and Autoencoders, and t-SNE's excellence in class separability for 2D visualizations.

V. RESULTS AND ANALYSIS

The linear dimensional reduction technique PCA performed exceptionally well on the dataset, reducing dimensionality by 60% while retaining 95% of the variance, demonstrating its efficiency. The reconstruction error was negligible ($\sim 5.498 \times 10^{-31}$), indicating minimal information loss. Computational efficiency was high, with a transformation time of only 0.0096 seconds.

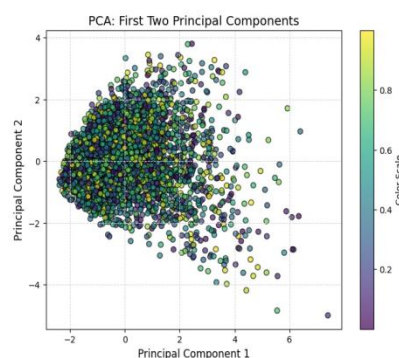


Figure 4. Visualization of data on a 2-dimensional Scatter plot.

The non-linear technique t-SNE showed high reconstruction error, poor global distance preservation, and required significant computational time (43.57 seconds), making it less efficient for large datasets or tasks needing global structure retention.

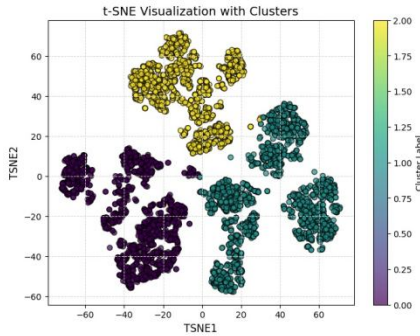


Figure 5

Auto encoders demonstrated a balanced performance in dimensionality reduction. The reconstruction error was 0.122, indicating a reasonable ability to reconstruct the original data from its latent representation. The dimensionality reduction efficiency turned out to be 0.4, reducing the data to 40% of its original space.

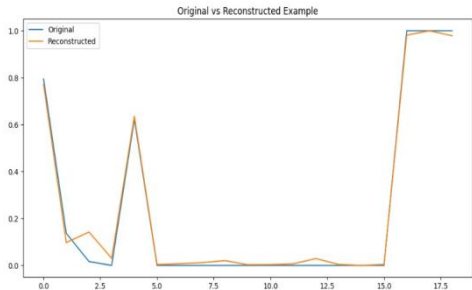


Figure 6

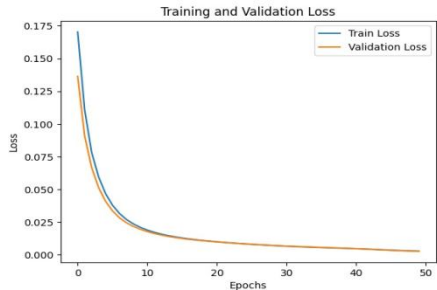


Figure 7

The correlation plot shows some discrepancies in structure preservation, which could be optimized further. The training and validation loss curves show a smooth and consistent decline, indicating that the model was trained effectively without overfitting. The computational efficiency, with a training time of 17.55 seconds, reflected good performance for this dataset.

TABLE I: COMPARISON OF DIMENSIONALITY REDUCTION TECHNIQUES

Metric	PCA	t-SNE	Autoencoders
Reconstruction Error	5.498×10^{-31} (near-zero)	55.467 (high)	0.122 (low-moderate)
Dimensionality Reduction Efficiency	0.4	243.99356 (very high)	0.4
Computational Efficiency (Time)	0.0096 seconds (fast)	43.57 seconds (slow)	17.55 seconds (moderate)
Strengths	Efficient, retains global structure	Excellent for visualization, preserves local structure	Learns complex, non-linear patterns

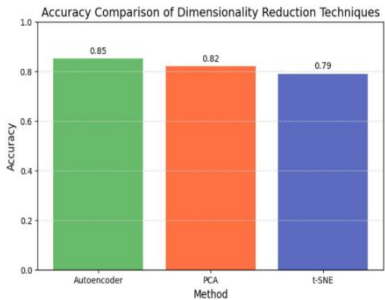


Figure 8

VI. CONCLUSION

The evaluation of dimensionality reduction techniques highlights that the choice of method is highly dependent on the dataset's characteristics. For linear datasets, Principal Component Analysis (PCA) emerges as the most effective technique due to its computational efficiency and ability to retain significant variance. It demonstrated minimal reconstruction error and high dimensionality reduction efficiency, making it suitable for large-scale, structured datasets.

For non-linear datasets, autoencoders and t-Distributed Stochastic Neighbor Embedding (t-SNE) showed distinct strengths. Autoencoders excel in capturing complex, non-linear patterns, offering a balanced trade-off between reconstruction accuracy and computational cost. They are particularly effective for datasets requiring intricate feature learning. On the other hand, t-SNE is highly valuable for visualizing non-linear relationships, particularly in clustering tasks, despite its computational intensity.

REFERENCES

- [1] M. Rai, H. Parmar, S. Singh, "Comparative Analysis of Dimensionality Reduction Techniques in Machine Learning Models for Liver Disease Detection", Proceedings of the International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2024) IEEE Xplore Part Number: CFP24OSV-ART; ISBN: 979-8-3503-7642-5.
- [2] S. Sakib, Md. A.B Siddique, Md. A. Rahman, "Performance Evaluation of t-SNE and MDS Dimensionality Reduction Techniques with KNN, ENN and SVM Classifiers", 2020 IEEE Region 10 Symposium (TENSYP), 5-7 June 2020, Dhaka, Bangladesh.
- [3] Y. Liang, X. Li, X. Huang, Y. Yao, Z. Zhang, "An Automated Data Mining Framework Using Autoencoders for Feature Extraction and Dimensionality Reduction".
- [4] F.R. Mulla, Dr. A.K. Gupta, (2022). "A Review Paper on Dimensionality Reduction Techniques"
- [5] G.T. Reddy, R. Kaluri, P.K. Reddy, K. Lakshmana, D. Rajput, G. Srivastava, T. Baker, (2020). "Analysis of Dimensionality Reduction Techniques on Big Data"
- [6] Shtino, Viola & Muca, Markela & Kapçiu, Rinela. (2024). "Dimensionality Reduction: A Comparative Review using RBM, KPCA, and t-SNE for Micro-Expressions Recognition." Volume 15 Issue 1. 370-379. 10.14569/IJACSA.2024.0150135.
- [7] <https://www.kaggle.com/datasets/reihanenamdari/breast-cancer>.
- [8] L. van der Maaten, G. Hinton, (2008). "Visualizing Data using t-SNE".
- [9] G.C. Linderman, M. Rachh, J.G. Hoskins, S. Steinerberger, Y. Kluger, (2017). "Efficient Algorithms for t-Distributed Stochastic Neighborhood Embedding".
- [10] D. G. CORTÉS, E. ONIEVA, I.P. LÓPEZ, L. TRINCHER, J. WU, (2024). "Autoencoder-Enhanced Clustering: A Dimensionality Reduction Approach to Financial Time Series".
- [11] A. Ashraf, N.M. Nawi, T. Shahzad, M. Aamir, M.A. Khan, K. Ouahada, (2024). "Dimension Reduction Using Dual-Featured Auto-Encoder for the Histological Classification of Human Lungs Tissues".
- [12] Migenda, Nico & Möller, Ralf & Schenck, Wolfram. (2019). Adaptive Dimensionality Adjustment for Online "Principal Component Analysis". 10.1007/978-3-030-33607-3_9.
- [13] B.G. Sarmina, Guo-Hua Sun, Shi-Hai Dong, (2023), "PCA and t-SNE analysis in the study of QAOA entangled and non-entangled mixing operators".
- [14] Kumar, Vinod & Biswas, Sougatamoy & Rajput, Dharmendra & Patel, Harshita & Tiwari, Basant. (2022). PCA-Based Incremental Extreme Learning Machine (PCA-IELM) for COVID-19 Patient Diagnosis Using Chest X-Ray Images. Computational Intelligence and Neuroscience. 2022. 1-17. 10.1155/2022/9107430.
- [15] Xue, Jingliang & Chen, Yingchun & Li, Ou & Li, Fei. (2020). Classification and identification of unknown network protocols based on CNN and T-SNE. Journal of Physics: Conference Series. 1617. 012071. 10.1088/1742-6596/1617/1/012071.
- [16] Yu, Ning & Pan, Yi. (2017). A deep learning method for lincRNA detection using auto-encoder algorithm. BMC Bioinformatics. 18. 10.1186/s12859-017-1922-3.