Geometric Data Analysis Final Project: Analysis of High-Dimensional Data Using Manifold Learning Methods

David Vivish

Columbia University

May 12, 2024

Abstract

This project evaluates three dimensionality reduction techniques—Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Isometric Mapping (Isomap)—on the MNIST dataset, focusing on the digits ’1’ and ’7’. The study assesses each method’s effectiveness in clustering and visualizing these digits, exploring their capability to uncover underlying data structures and patterns. PCA demonstrated efficiency in reducing dimensionality but faced challenges with overlapping digits. t-SNE excelled in forming distinct clusters through its sensitivity to local structures, although its performance was heavily parameter-dependent. Isomap maintained a balance, effectively preserving global relationships with reasonable local clustering. Quantitative analyses, including silhouette scores and nearest neighbors, were used to evaluate clustering performance, highlighting strengths and limitations of each technique. The findings suggest potential for hybrid approaches and further exploration of advanced methods like autoencoders and UMAP to optimize clustering accuracy and visualization in image classification tasks.

1 Introduction

1.1 Purpose of the Study

The primary purpose of this study is to explore and evaluate the efficacy of various dimensionality reduction techniques in revealing the underlying structure of high-dimensional data. Specifically, the analysis focuses on the MNIST dataset, comprising handwritten digits, to determine how different methods influence the separation and visualization of data categories that are challenging to distinguish. By comparing the results across multiple techniques, this study aims to provide insights that can guide the selection and application of dimensional reduction methods in practical machine learning and data analysis scenarios.

1.2 Importance of Dimensional Reduction

In the realm of data science, particularly in fields dealing with large-scale and high-dimensional data, dimensional reduction serves as a crucial analytical technique. High-dimensional data often suffers from the “curse of dimensional,” where the performance and accuracy of data analysis methods deteriorate as the number of dimensions increases. Dimensional reduction techniques address this problem by transforming high-dimensional data into a lower-dimensional space, simplifying the dataset while preserving its most significant characteristics. This not only enhances the performance of machine learning algorithms but also aids in data visualization and helps uncover hidden patterns within the data that are not discernible in higher dimensions. Effective dimensional reduction is, therefore, essential for making complex data more interpretable and manageable, facilitating deeper insights and more robust data-driven decision-making.

1.3 Overview of the MNIST Dataset

The MNIST dataset, an acronym for Modified National Institute of Standards and Technology database, is one of the most extensively utilized datasets in machine learning for benchmarking classification algorithms. It consists of 70,000 images of handwritten digits (from 0 to 9), each with a 28x28 pixel grayscale representation. This dataset is particularly valued for its complexity and variability, presenting a realistic challenge miming real-world data analysis scenarios. The MNIST dataset serves as a testbed for developing and testing machine learning algorithms and provides a standard by which to compare the effectiveness of different analytical techniques. For this study, the analysis focuses specifically on the digits ′1′ and ′7′, chosen due to their visual and structural similarities, which test the capability of dimensionality reduction techniques to separate and clarify similar data categories effectively.

1.4 Objectives

This study aims to apply and compare three principal dimensional reduction techniques—Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Isometric Mapping (Isomap)—to the MNIST dataset. Through detailed analysis and comparison of these methods, the study seeks to determine which technique most effectively uncovers the underlying structure of the dataset, particularly in differentiating between the digits ′1′ and ′7′. The outcomes are expected to provide valuable guidelines for applying these techniques in various data analysis contexts, enhancing the understanding of their strengths and limitations.

2 Data Description

2.1 Dataset Overview

The Modified National Institute of Standards and Technology (MNIST) dataset is utilized for this analysis. It comprises 70,000 images segmented into a training set of 60,000 examples and a test set of 10,000 examples. Each digit is depicted in a 28x28 pixel grayscale image. These images are arrayed in a structure where each pixel represents an intensity scale from 0 (black) to 255 (white), providing a clear, albeit compressed, visual of handwritten digits.

2.2 Data Selection

The strategic selection of the MNIST dataset for this study is underpinned by several factors that render it exceptionally suited for a comprehensive examination of dimensional reduction techniques. Foremost among these is the MNIST dataset's popularity and accessibility within the machine learning community. Its ubiquitous presence across academic and research settings positions it as an ideal candidate for producing broadly applicable and easily comparable results across many existing studies.

Moreover, the MNIST dataset's inherent complexity, stemming from the natural variations in handwriting style among thousands of individuals, presents a realistic and challenging data analysis scenario. This complexity is pivotal as it tests the robustness and efficacy of various dimensional reduction techniques under realistic conditions that mimic real-world data variability.

Lastly, the MNIST dataset's status as a standard benchmark in the field allows for the results of this study to be readily compared with a vast array of existing research, thus situating our findings within a larger, well-established context. This benchmarking capability significantly enhances the relevance and reliability of the study's outcomes.

2.3 Specific Characteristics of Digits '1' and '7'

The focus of this analysis on the digits '1' and '7' was informed by several critical considerations. The visual similarity between these two digits, particularly when handwritten, introduces a unique classification challenge. These digits often feature minimalistic strokes and may include stylistic flourishes that blur the distinction between them, such as a horizontal serif at the top of '1' or a crossbar on '7'. This similarity poses a substantial challenge in pattern recognition and classification, making these digits ideal subjects for evaluating the efficacy of different dimensional reduction techniques in distinguishing between closely related categories.

Moreover, the structural variability inherent in the representations of these digits across different individuals further complicates their analysis. Variations such as the presence or absence of serifs or crossbars significantly affect their appearance and resemblance to one another. This variability is a crucial element of the study, as it tests the capability of dimensional reduction techniques to handle and articulate subtle differences within the data.

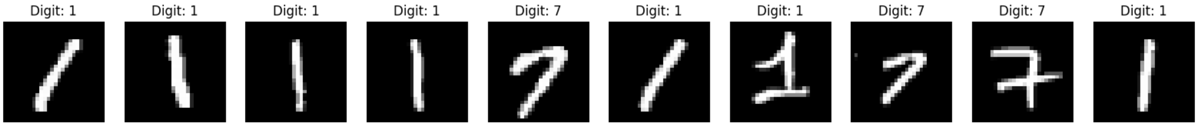

Figure 1: Sample of the dataset

3 Methodology

3.1 Preprocessing Steps

3.1.1 Normalization

Given the dataset's inherent structure, where each image is represented by pixel values ranging from 0 to 255, a critical initial step involved scaling these values to a standardized range of 0 to 1. This normalization was achieved by dividing each pixel value by 255, effectively transforming the grayscale intensities to a scale that enhances algorithmic processing and analysis. Such scaling is imperative not only for reducing the variance among pixel values but also for facilitating the application of algorithms like Principal Component Analysis (PCA), which are particularly sensitive to the scale of input data.

3.1.2 Data Flattening

The images within the MNIST dataset required additional preprocessing to accommodate the specific needs of the dimensional reduction techniques employed in this study—PCA, t-SNE, and Isomap. The images, originally in a 2D format of 28x28 pixels, were transformed into 1D arrays consisting of 784 pixels. This transformation, commonly referred to as flattening, is essential for processing with the chosen dimensionality reduction techniques, all requiring input data in vector form rather than two-dimensional matrices.

3.1.3 Reasons for Preprocessing Steps

The preprocessing steps designed for this study are integral to ensuring that the MNIST dataset is optimally prepared for sophisticated analytical techniques like PCA, t-SNE, and Isomap. Each step enhances the data's suitability for analysis by ensuring uniformity and appropriate data structure, crucial for the application of these dimensional reduction methods.

The uniform scaling of pixel values and centering around zero are particularly important. These steps prevent biases in the analysis that might arise from disparities in data scales, thus ensuring that features contribute equally to the analysis outcomes. This uniformity is essential for algorithms like PCA, which rely heavily on variance to determine the principal components.

Additionally, transforming the 2D images into 1D vectors aligns the data format with the requirements of the dimensional reduction algorithms, enabling effective computation and precise results. The focus on digits ′1′ and ′7′ sharpens the study's objectives, allowing a concentrated exploration of the techniques' effectiveness in distinguishing between visually and structurally similar categories.

These preprocessing strategies not only facilitate the technical execution of the dimensional reduction methods but also underscore the study's methodological rigor, ensuring that the findings are reliable and relevant to real-world applications. This careful data preparation underscores our commitment to analytical rigor and is pivotal in achieving the insightful outcomes anticipated from this research.

3.2 Technical Implementation

In this study, we employed various dimensional reduction techniques, supplemented by clustering algorithms, to explore the underlying structure of the MNIST dataset, particularly focusing on the digits ′1′ and ′7′. The implementation details of each technique, including parameter settings, are described below. The justification for these choices is linked to the dataset's inherent characteristics and our analysis's objectives.

3.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was strategically employed as the primary method for reducing the high-dimensional nature of the MNIST dataset, which consists of 784 dimensions per image. By transforming the data into a new coordinate system, PCA highlights the directions where the variance of the data is maximized, effectively capturing the most significant features of the dataset.

Parameter Settings:

- The number of components retained was set to 50. This decision was informed by preliminary exploratory data analysis, which indicated that the first 50 principal components account for approximately 95′%′ of the total variance. This strategic reduction significantly simplifies the dataset while retaining crucial information for robust analysis.

3.2.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Building upon the dimensional reduction achieved through PCA, t-SNE was applied to further delineate the local structure of the data. Renowned for its efficacy in high-dimensional data visualization, t-SNE reveals intricate patterns and clusters that are not apparent in the higher-dimensional space.

Parameter Settings:

- Perplexity: Set at 30, this parameter was chosen as it generally provides a good balance between capturing local and global aspects of the data, suitable for a variety of datasets.
- Learning Rate: Fixed at 200, the learning rate dictates the speed at which the model learns the data distribution. A carefully selected learning rate ensures that

the algorithm converges efficiently without overshooting, minimizing the risk of falling into local minima or undergoing excessively long convergence times.

3.2.3 Isomap

To complement the local insights provided by t-SNE, Isomap was utilized as an additional dimensional reduction technique to preserve the global geometry of the data more effectively. Isomap is adept at maintaining the geodesic distances between points, respecting the manifold's intrinsic geometric structure.

Parameter Settings:

- Number of Neighbors: Chosen to be 5, this setting ensures that the manifold's geometry is accurately represented by considering only the nearest points. This choice is critical for effectively preserving the true geodesic distances in the reduced-dimensional space, providing a comprehensive understanding of the data's global structure.

3.2.4 K-Means Clustering

Following the application of dimensionality reduction techniques, k-means clustering was employed to identify and delineate groups within the data. K-means, a widely utilized clustering algorithm, partitions the data into clusters based on the mean distance from the cluster centroids, which effectively categorizes the data into distinct groups.

Parameter Settings:

- Number of Clusters: Set to 2, this parameter directly correlates with the focus of our analysis on the two digits, ′1′ and ′7′. This setting not only simplifies the clustering process but also aligns precisely with the binary nature of our analytical focus, facilitating a clear and focused examination of the separability of these two digit categories.

4 Results

4.1 Dimensionality Reduction Outputs

The dimensionality reduction techniques employed—Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Isometric Mapping (Isomap)—have provided insightful visualizations that highlight the data structure of the MNIST dataset's digits ′1′ and ′7′. Each method, as visualized in the plots, brings unique perspectives to the forefront of our analysis, elucidating the intricate relationships between the data points.

4.1.1 PCA

The PCA plot reveals a clear gradient where the two digit categories are somewhat distinguishable but show a significant overlap in the central area of the plot. The projection onto the first two principal components, which capture the most significant variance, suggests a continuous spectrum rather than discrete clusters. This indicates that while PCA is effective in reducing dimensionality, its linear nature might not be entirely sufficient for capturing the complex, non-linear relationships necessary to fully separate the two similar digits.

4.1.2 t-SNE

t-SNE's algorithm, by design, focuses on converting affinities of data points into probabilities; nearby data points have a high probability of being picked, while distant points have an exponentially decaying probability of selection. This results in a map that significantly respects local relationships, making it particularly effective for visualizing datasets where such relationships are key to understanding the data structure. In the provided t-SNE visualization, each cluster appears as a cohesive group, which suggests that within each category ('1' and '7'), the digits share a high degree of similarity.

However, despite the apparent advantages in displaying localized data clusters, t-SNE is not without its limitations, as evidenced by the intertwined boundaries between the clusters of '1' and '7'. This intertwining indicates areas of ambiguity where the representations of the digits '1' and '7' are not distinctly separate.

4.1.3 IsoMap

The Isomap visualization also achieves a notable separation between '1' and '7', with a smoother transition between clusters compared to t-SNE. This method maintains the global topology of the dataset effectively, suggesting that Isomap is capable of capturing both the essential geometric and global properties of the data. The continuous yet distinct areas of concentration for each digit indicate a successful dimensional reduction that preserves meaningful distances and relationships inherent in the original high-dimensional space.

4.1.4 Potential Improvements

- Parameter Tuning: Adjusting parameters such as the number of components in PCA, perplexity and learning rate in t-SNE, and the number of neighbors in Isomap could further optimize each model's ability to discern between closely knit categories.

- Hybrid Approaches: Combining these techniques sequentially or in a hybrid model could leverage the strengths of each. For instance, initiating with PCA to reduce dimensionality followed by t-SNE might yield more distinct clusters with less computational overhead.

4.1.5 Conclusion

Each of these dimensionality reduction techniques provides valuable insights into the structure of the data, yet they also illustrate the inherent trade-offs between capturing global versus local data properties. PCA, while providing a broad overview, lacks the granularity needed to distinctly separate the digits ′1′ and ′7′. In contrast, t-SNE and Isomap offer more nuanced separations, highlighting their strengths in dealing with complex patterns and maintaining data integrity across reduced dimensions.

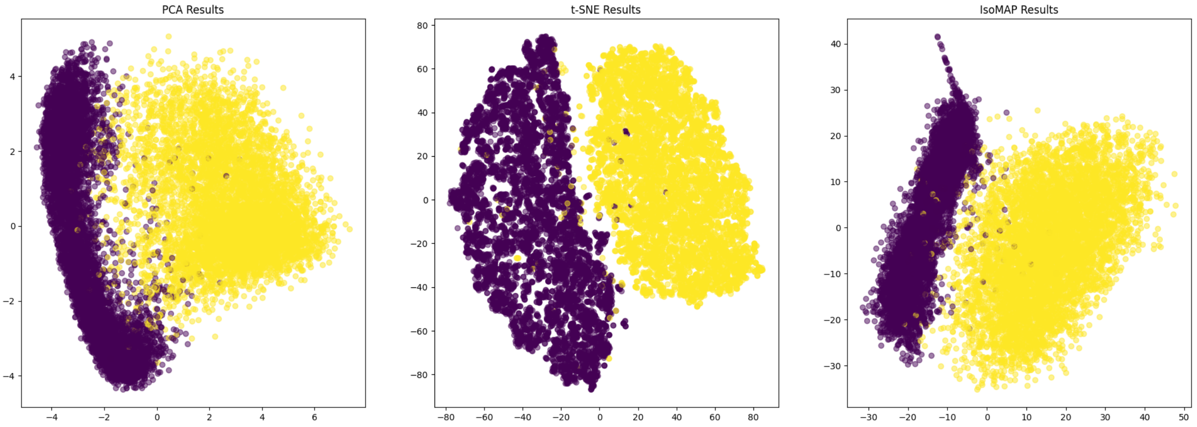

Figure 2: Clusters

4.2 Clustering Analysis

The clustering visualizations generated from the outputs of Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Isometric Mapping (Isomap) applied on the MNIST dataset provide a comparative perspective on how each dimensionality reduction technique manages to cluster the digits ′1′ and ′7′. The effectiveness of these clusters was quantitatively assessed using silhouette scores.

Silhouette Scores		
PCA	t-SNE	IsoMap
0.514	0.401	0.489

4.2.1 PCA

The silhouette score for PCA indicates a moderate level of cluster definition. A score closer to 1 would suggest perfect clustering where each point is closer to other points within its cluster than to points in neighboring clusters. A score of 0.514 suggests that while the clusters are generally distinct, there is room for improvement in terms of increasing the distance between clusters (inter-cluster distance) and reducing the variance within clusters (intra-cluster compactness). This indicates that some digits ′1′ and ′7′ might still be positioned too close to the cluster boundary, or that the clusters themselves are not as tight as they could be. To improve this score, strategies such as increasing the number of dimensions retained in PCA or employing a more robust scaling method prior to PCA application could be considered to enhance the separability and internal uniformity of clusters.

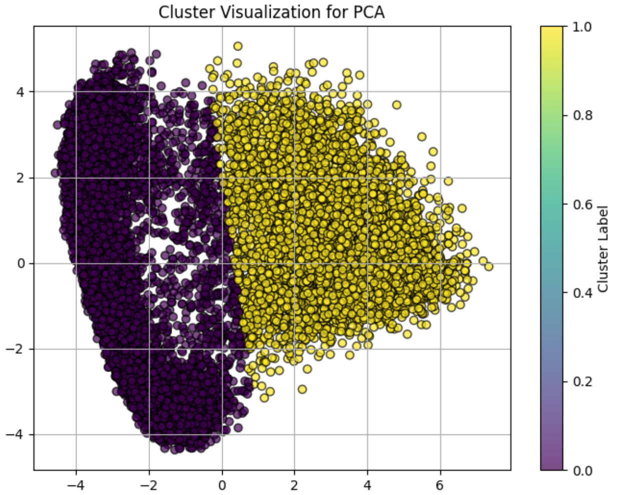

Figure 3

4.2.2 t-SNE

Despite t-SNE providing visually distinct clusters, the silhouette score of 0.401 suggests that these clusters might not be as internally homogeneous or as well-separated as they appear. This lower score could result from the algorithm's propensity to preserve local neighborhood structures, which might lead to higher variance within clusters if the local neighborhoods themselves are diverse. Moreover, t-SNE's stochastic nature means that it may produce clusters that are visually separated but not optimally positioned in the reduced-dimensional space to maximize the silhouette score.

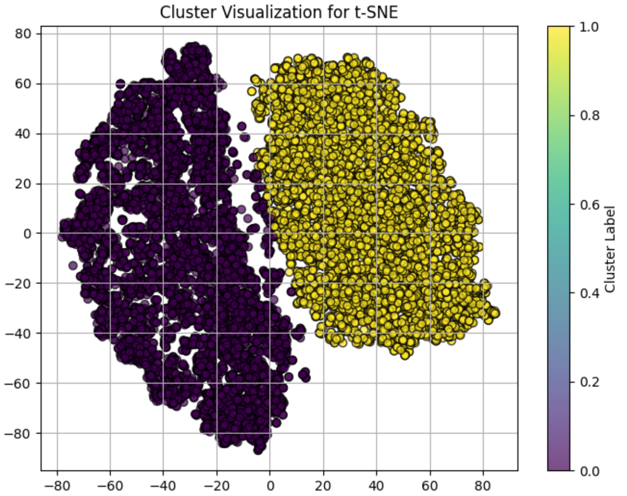

Figure 4

4.2.3 IsoMap

Isomap's silhouette score, which is closer to that of PCA, suggests a reasonable balance between maintaining global data structure and achieving cluster separation. The score reflects that the clusters are fairly well-defined but, similar to PCA, could benefit from increased separation between clusters. This might be addressed by fine-tuning the number of neighbors considered in the manifold learning phase, which could affect how well the global distances are preserved in the lower-dimensional embedding.

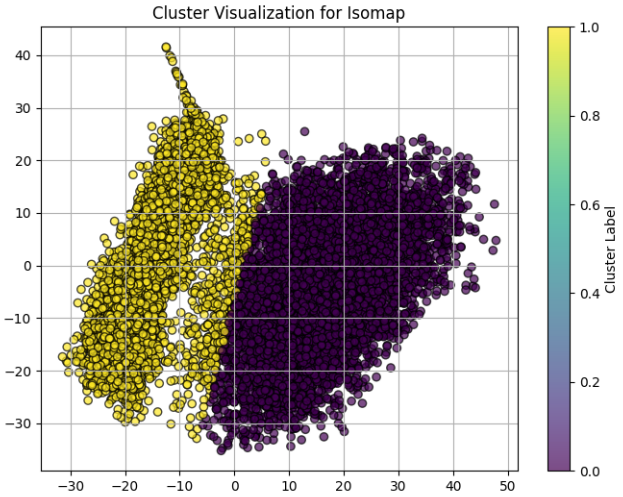

Figure 5

4.2.4 Concluding Insight

While PCA and Isomap tend toward preserving broader data relationships leading to moderately good clustering, t-SNE focuses intensely on local data intricacies, which might compromise its inter-cluster distances and overall silhouette score.

4.3 Nearest Neighbor Analysis

The Nearest Neighbors Analysis provides a granular view of how each dimensionality reduction technique structures the data around selected points within the MNIST dataset. This analysis specifically examines the local neighborhood structures resulting from Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Isometric Mapping (Isomap) by focusing on several sample data points.

4.3.1 PCA

The PCA visualization demonstrates that the selected data points (highlighted) have nearest neighbors that predominantly belong to the same digit class, indicating good local cohesion within the clusters. This is confirmed by the labels of the nearest neighbors, which are consistent across the selected data points, all being '1'.

However, the spatial distribution in the PCA plot shows a substantial overlap between different classes outside the immediate neighborhood, reflecting PCA's linear limitations in fully segregating complex, non-linear structures.

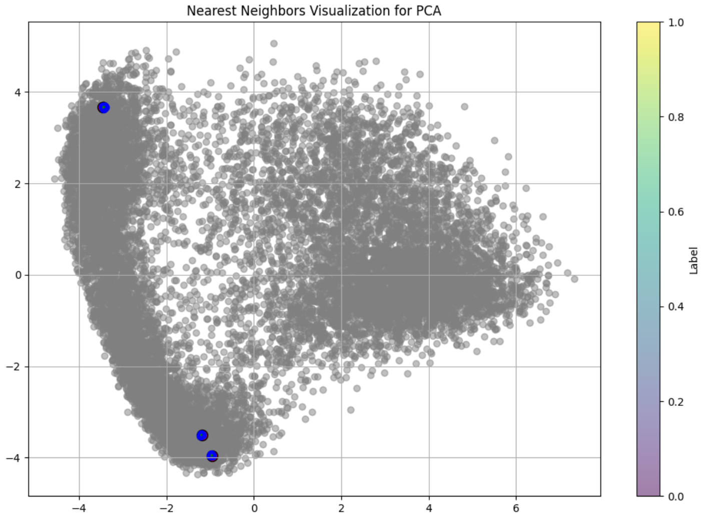

Figure 6

4.3.2 t-SNE

t-SNE's visualization reveals tighter and more isolated clusters compared to PCA. The selected data points have neighbors very close in the plot, signifying excellent local structure preservation. t-SNE's ability to cluster these points tightly with their neighbors, all sharing the same class label, underscores its effectiveness in capturing high-dimensional local data relationships within the reduced space.

Despite its effectiveness, the clustering sometimes shows boundary mingling with the opposite class, suggesting potential improvements in parameter tuning to enhance separation.

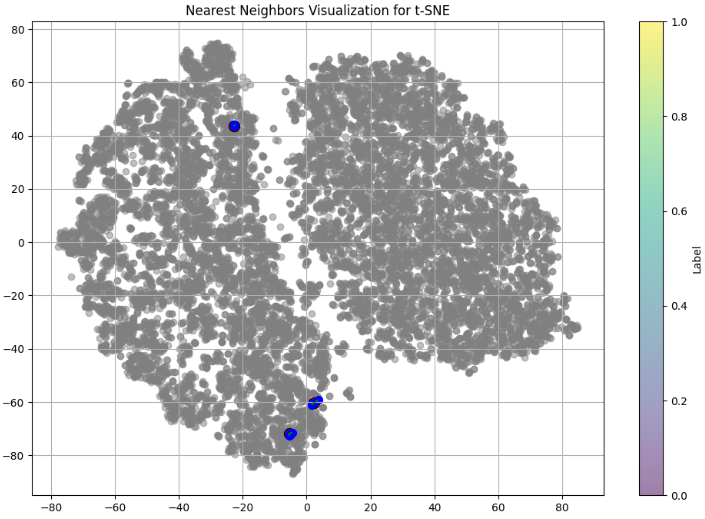

Figure 7

4.3.3 IsoMap

Isomap's plot shows a clear distinction in the layout of data points but with slightly less compact clustering compared to t-SNE. The selected data points are surrounded by neighbors of the same class, supporting the technique's capability to maintain meaningful global geometry and distance relations within the dataset.

Some of the points are closer to the boundary of the cluster, indicating that while Isomap is effective, there could be a risk of class overlap at the borders of clusters.

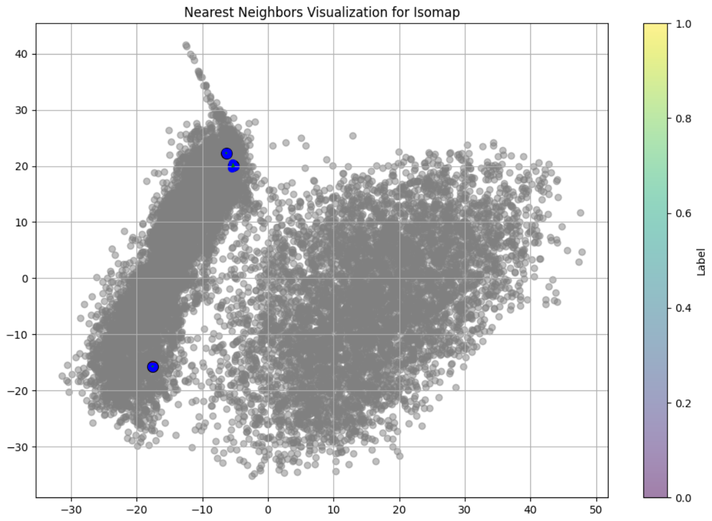

Figure 8

4.3.4 Concluding Insight

Across all three techniques, the analysis of nearest neighbors yields insights into each method's ability to maintain local data integrity after dimensionality reduction. Each technique successfully grouped the selected data points with neighbors of the same class, affirming their effectiveness in this aspect.

5 Conclusion

This study employed three different dimensionality reduction techniques—Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Isometric Mapping (Isomap)—to explore the underlying structure of the MNIST dataset, specifically focusing on the digits ′1′ and ′7′. The analysis revealed significant insights into how each technique handles high-dimensional data and their effectiveness in clustering and identifying nearest neighbors within the dataset.

PCA proved effective in reducing the dimensionality of the dataset, capturing significant variance and providing a broad overview of data distribution. However, its linear nature was less effective in completely segregating the overlapping digit classes.

t-SNE demonstrated superior capability in creating distinct, well-separated clusters that accurately reflect local data structures. Despite its effectiveness, the technique showed some sensitivity to parameter settings, which could affect cluster purity.

Isomap struck a balance between global and local structure preservation, effectively maintaining the manifold's geometric properties and providing clearer separation than PCA but with less local precision compared to t-SNE. The findings confirm that each dimensionality reduction technique has particular strengths and limitations that make them suitable for different types of data analysis tasks.

6 Potential Future Work

The exploration of dimensionality reduction techniques presented in this study opens several avenues for further research. One promising area is the optimization of parameters for each technique, potentially utilizing automated hyperparameter tuning tools to enhance performance across different datasets. The investigation could also extend into hybrid approaches that combine the strengths of linear and non-linear methods, such as integrating PCA for initial dimensionality reduction followed by t-SNE or Isomap, to leverage both computational efficiency and the ability to capture complex patterns. Furthermore, examining other dimensionality reduction algorithms such as Uniform Manifold Approximation and Projection (UMAP) may provide additional insights and potential improvements in speed, accuracy, or interpretability.

References

[1] Van der Maaten, L., Hinton, G. (2008). “Visualizing Data using t-SNE.” Journal of Machine Learning Research, 9, 2579-2605.

[2] Jolliffe, I. T. (2002). Principal Component Analysis, Second Edition. Springer Series in Statistics. New York: Springer-Verlag.