Enhancing t-SNE with PCA Initialization and Hybrid Distance Metrics

MSc Research Project
Programme Name

Srushti Wadkar
Student ID: 23201622

School of Computing
National College of Ireland

Supervisor: Prof. Christian Horn.

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Srushti Wadkar		
Student ID:	23201622		
Programme:	MSc. Data Analytics	**Year:**	2024-25
Module:	Research Practicum		
Supervisor:	Christian Horn		
Submission Due Date:	11 August 2025		
Project Title:	Enhancing t-SNE with PCA Initialization and Hybrid Distance Metrics		
Word Count:	5411	**Page Count**	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Srushti Wadkar
Date:	15 September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing t-SNE with PCA Initialization and Hybrid Distance Metrics

Srushti Wadkar
23201622

Abstract

High-dimensional data is often employed in the modern data analysis, but the complexity of this data type impairs visualization and interpretation. The t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm is a common method of dimensionality reduction and particularly effective at preserving local data structure of 2D or 3D projections. However, there are two important limitations of t-SNE: finite memory of the global structure that can cause biases in the interactions within clusters and the inability to reproduce results because of random initialisation.

When solving these limitations, this study has two specific contributions to make PCA-based initialization and a hybrid distance measure, which can be expressed as a combination of cosine similarity and Euclidean distance. Initialization of PCA also gives an initial state that t-SNE optimization can rely on, to improve repeatability and preserve the global variance structure. So as to improve the stability and interpretability of clusters, the hybrid distance measure is a trade-off between angular and magnitude similarity.

Having been implemented in Python with the help of scikit-learn, the techniques were tried out on a range of datasets, including MNIST image set and financial records of bankruptcies. The embeddings were visually inspected in a qualitative way, and silhouette scores were taken in a quantitative way. The results showed that the hybrid distance measure performed better in yielding globally consistent embeddings particularly when a complex data is under consideration, but initialization by PCA was more consistent and reproducible in its layouts.

The paper confirms and scales up the results of the earlier research by systematizing and testing these advancements to a wide variety of data domains, presenting practical, repeatable procedures to more effectively visualize high-dimensional data. The results can be useful to researchers and practitioners seeking useful low-dimensional representations of complex data that are easy to understand.

1 Introduction

High-dimensional data have been becoming more common in most fields, such as image processing, finance, natural language processing, and bioinformatics, in the data-driven society. Interpretation and visualization of structure of larger and more complex data is a key problem and dimensionality reduction techniques are very important in enabling this to be done. t-distributed Stochastic Neighbour Embedding (t-SNE) is an excellent example of a dimensionality reduction technique that has been taken up as the way of preserving local data structure as well as offering useful low dimensional visualizations.

However, the standard t-SNE has a number of limitations, which cannot be overlooked. The first one is its random initialization that leads to wildly different visual outputs after multiple runs on the same data. Second, it has difficulty in sustaining global structure, and as such, it becomes hard to understand the relationship between distant clusters. Finally, the traditional t-SNE is very computationally complex particularly in the case of large size datasets.

In order to address such limitations, it is the objective of this paper to enhance the t-SNE algorithm. Two improvements proposed:

- PCA-based Initialization: To make t-SNE deterministic and reproducible, a PCA-based form of initialization is applied where structured start points are employed instead of the random start.
- Hybrid Distance Metric: Apply a hybrid distance metrics, combination of the cosine similarity that measures direction and the Euclidean distance that measures magnitude in order to enhance the global structure.

To check the robustness of the improvements, these enhancements are used and evaluated on the two different datasets: MNIST image dataset of 784 features per sample of handwritten digits and financial bankruptcy dataset of 72 features per sample of structured numerical data.

This paper aims to offer a more-reliable, interpretable, and generalizable variation of t-SNE in practicable applications of visualizing high-dimensional data by comparing the results in multiple fields.

2 Related Work

The t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm by van der Maaten and Hinton has been one of the most popular tools in the visualization of the high-dimensional data based on the local structure preservation. However, the fact remains that the conventional t-SNE algorithms are time-consuming and sensitive to the metric. The contribution by (Linderman et al. 2019) to the optimization of t-SNE in terms of better algorithms, the Barnes-Hut t-SNE and FFT-based approximations, opened the possibilities of further experimentation involving the initialization strategies and custom distance functions. This basic realization is made against this background that there is a need to make some improvements that bring out the benefits of t-SNE and address the shortcomings of its structural and reproducibility limitations.

2.1 PCA Initialization

The most noticeable drawback of t-SNE is the fact that it is random initialization-based, and, therefore, runs cannot be embedded consistently. It has recently been suggested that Principal Component Analysis (PCA) is in a systematic and predictable way of initializing to reduce this.

(Kobak and Linderman 2021) was one of the first to demonstrate the initiation by PCA results in an enormous gain in retention of global structure in t-SNE and UMAP. This was affirmed by their empirical evidence whose results indicated that they could better interpretable and repeatable arrangement of data by starting their analysis with the first two major components. (Kobak and Berens 2019) went further and tested t-SNE on single-cell transcriptomics, finding similar results as above: the same clustering results could be achieved with PCA, as well as with no initialization of the embeddings, thus, the visualization was more reliable and scientifically sound.

(Chourasia et al. 2022) also explored initialization and kernel in t-SNE and it was determined that PCA-based initialization could be used to achieve a higher fidelity of clustering and in addition to that, it also converged much faster. Similarly, (Tränkner et al.) verified the positive attributes of the pre-processing t-SNE in real-world data using PCA, i.e., improved stability of the layout and diminished computation artifacts. The same concept has also been applied by (Cristian et al. 2024) to UMAP and they indicated that PCA-based initialisation of UMAP can improve noise cleaning and global topology in high-dimensional biological data.

The combination of PCA as a t-SNE initialization strategy can be confirmed by the collection of studies provided and, as such, it is a direct justification of the methodological approach implemented in the present research to enhance the interpretability and reproducibility of the layouts designed.

2.2 Hybrid Distance Metrics

The t-SNE has been known to maintain local structure, but it loses global connections of clusters. To take into consideration both magnitude and directional correlations, researchers have tried out the hybrid distance metrics, which combine the various similarity functions, typically the Euclidean and cosine distance.

In their comparative analysis of the methods of dimension reduction, like t-SNE, TriMap, and UMAP, (Wang et al. 2021) found that the preservation of global structure depends directly on the choice and the construction of the similarity measure. By carrying out GPU-accelerated variants of the hybrid t-SNE and comparing it to the original t-SNE and the fast t-SNE on large-scale datasets, (Meyer et al. 2022) demonstrated that it could be applied to attain a better quality of the embedding and spatial coherence.

To visualise large data, (Meyer et al. 2022) extended t-SNE to allow the use of hybrid metrics and found that by combining distance functions, more interpretable plots were obtained, where cluster boundaries were much more certain. Some studies also offered theoretical support on the effectiveness of the hybrid geometric relations in enhancing the quality of manifold learning algorithm embedding procedures such as t-SNE. Finally, (Chourasia et al. 2022) by confirming the appropriateness of the use of a custom hybrid distance matrix consisting of normalized Euclidean and cosine elements, substantiated the idea of tailoring the similarity measures to reflect more the structure of the data.

These contributions are highly relevant to the hybrid distance approach adopted in this project and are a good indication of the inclusion of hybrid distance measurements as a means of overcoming the global structure constraints of t-SNE.

3 Research Methodology

3.1.Tools and Libraries

Implementations were conducted using Python 3.10 within a Jupyter Notebook Environment. Some of the most popular libraries such as Matplotlib and Seaborn (For Visualization), NumPy and Pandas (For data handling), and Scikit-learn (For PCA, t-SNE, distance metrics and evaluation) were used to conduct data analysis and machine learning activities in the project (seaborn 0.11.1 documentation; scikit-learn 0.22.2 documentation).

3.2.Datasets

The two different data sets were selected, MNIST and a Taiwanese bankruptcy dataset, in order to evaluate the proposed enhancements. The MNIST dataset contains 60,000 grayscale images of handwritten numerals, each 28 × 28 pixels and flattened into 784 dimensions. A random sample size of 2,000 samples was chosen to conduct the experiment to ensure that the samples represent a wide variety of classes of digits because t-SNE takes a very long time to compute. On the other hand, the bankruptcy data is structured financial data with a binary label (the bankruptcy status) and 72 numerical variables which indicate the company performance.

3.3.Data Preparation

Since the bankruptcy dataset has a high-class imbalance, VIF analysis was done, and a couple of features were eliminated. Both the datasets are pre-processed and also normalized before applying algorithms. In PCA and hybrid measures, the measures are also sensitive to scale of features hence it was necessary to normalize the features so that each feature contributes the same to the calculation of the distance and similarity. Features (X) and labels (Y) were extracted and saved in order to use them in the following procedures.

3.4.Baseline Standard t-SNE Model

The two datasets were analysed using the standard t-SNE with default settings in order to have a point of reference. Euclidean distance was the metric, and it was random 2-dimensional initialization. The default perplexity of 30 was used. This baseline was made with the vision of being able to observe the limitations of using t-SNE, especially its instability and a poor global structure.

3.5.PCA Initialization

To solve the issue of the reproducibility of the t-SNE, the preliminary embeddings that were to be used as the starting point had already been measured through PCA. The two dimensions out of the high-dimension data were calculated and given as the parameters to the init parameter of TSNE instead of initializing the t-SNE layout at random parameters. PCA is a linear and deterministic procedure which ensured each run of t-SNE started at same point (misha-pyshark 2020).

3.6.Hybrid Distance Metric Application

The issue with the conventional t-SNE is that it makes use of Euclidean distance which only reflects the differences in magnitude and does not reflect directions or patterns of relationships. In order to maintain as much of the global structure as possible, hybrid distance was formulated as a combination of cosine similarity and Euclidian distance. Cosine similarity considers directions of the vectors in the high dimensional space whereas the Euclidean distance measures the absolute differences between the features (Levy et al. 2024).

During the process of implementation, the cosine similarity matrix and Euclidean distance matrix were obtained with the help of functions in Scikit-learn. The cosine distance was obtained by subtracting cosine similarity from one. Two metrices were then normalised to a [0,1] range and averaged with equal weights (α=0.5) to form combined hybrid distance matrix. It was then fed into t-SNE format with metric="precomputed" option.

3.7.Dimensional Extension

This study has extended the functionality of t-SNE to five dimensions to investigate how t-SNE affects the quality of clustering because most t-SNE visualizations are limited to two dimensions. It was only used with the MNIST dataset, due to the high-dimensional and rich structure. As the Barnes-Hut approximation cannot be used in more than three dimensions, the t-SNE algorithm was modified to have an exact value and n_components parameter was defined to be 5. The 5D output was again reduced to 3 dimensions by PCA so that even when direct display of 5D embeddings is not possible, we can examine it through 3D.

3.8.Perplexity Sensitivity Test

The perplexity is a very important parameter in t-SNE that helps to specify the number of neighbours each data point takes into consideration. The perplexity values of 5, 10, 20, 30, 40 and 50 were applied on MNIST dataset to analyse its implications. With both Euclidean and hybrid distance metrics, t-SNE was applied at each perplexity setting and the results were written as 2D and 5D.

3.9.Evaluation Strategy

The different patterns were evaluated with both quantitative and qualitative measures. The silhouette score was the principal quantitative measure of expressing the similarity of a data points of a cluster with itself and the other clusters. Higher scores reflect better clustering and separation. The silhouette_score in Scikit-learn was used to compute this.

To evaluate visual representation of the embeddings, qualitative measurement is made through a plot of the embeddings in 2D and 3D plots. Plots were interpreted in terms of clear separation of groups, cluster shapes and overlaps using colour coding of the plots based on the class labels (digit or bankruptcy status) (Awan 2023). All these evaluations allowed making an adequate comparison between the initial and enhanced t-SNE setups.

4 Design Specification

The improved t-SNE framework developed during this research is designed to be cross-domain adaptable, reproducible and modular. The purpose of the design is to systematically address the familiar limitations of the traditional t-SNE i.e. its inability to preserve global structure adequately and its instability due to random initialization. The design is composed of individual modules which can be replaced and effortlessly integrated and tested with enhancements such as PCA-based initialization, hybrid distance measures, dimensionality configuration and perplexity adjustment.

4.1.Modular Architecture

The system architecture allows each of the enhancements to be tested individually or together with other enhancements. This modularity is necessary to scale to future extensions like GPU acceleration or to other dimensionality reduction approaches, and to be able to compare the impacts of innovations on different types of data.

Key Design Modules:

- Initialization Module: The initialization module enables the toggling between random initialization and deterministic initialization with PCA.
- Distance Metric Module: It accepts a user defined hybrid distance matrix or a standard Euclidean distance.
- Dimensionality Configuration: Allows dynamic selection of techniques (exact in higher dimensions and barnes_hut in lower dimensions), 2D, 3D and 5D embedding.
- Perplexity Loop Controller: It automatically executes t-SNE with different perplexity values.
- Evaluation Module: It is graphical elements and silhouette scores-based module to measure quality of embedding

This modular framework makes it simple to connect and play around with various set-ups and thus allows experimentation to be efficient, systematic and extendable.

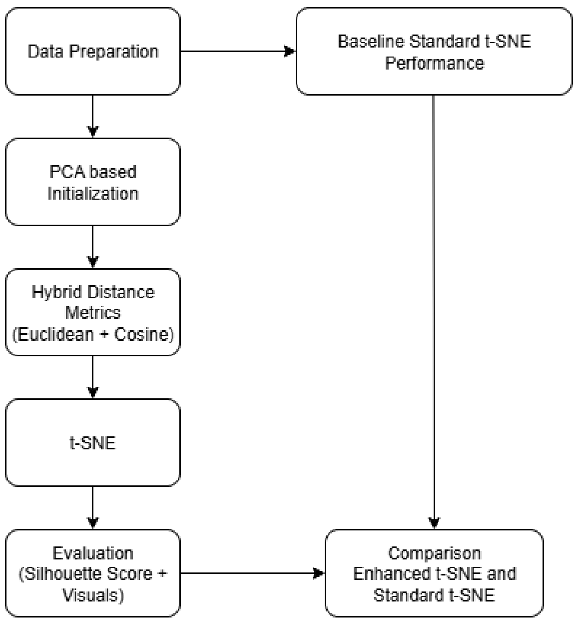


Figure 1: Project Flow

4.2.PCA Initialization Design

Traditional t-SNE is non-deterministic between runs because it starts with random initial embedding. To avoid this problem, there is the development of a PCA-based initialization module to promote stability and repeatability.

In this architecture, the first high-dimensional data is subjected to PCA in order to generate the top principal components. These factors provide t-SNE with an orderly initial layout and they record the directions of maximum variance. The results of the PCA are directly passed to the init parameter of the t-SNE which replaces randomness with a deterministic starting point. The technique ensures proper interpretation and reduces the variability of the results, which is essential to analytical reporting and scientific research.

The design of PCA allows the setting up of a large number of output dimensions (2D, 3D or 5D) and this makes it flexible in analyzing complex datasets such as MNIST or simpler structured data such as bankruptcy data.

4.3.Hybrid Distance Matrix Design

The architecture incorporates a hybrid distance metric component which combines the cosine-based and Euclidean distances to promote global structure preservation. The goal is to consider both directional alignment (where cosine similarity is used) and spatial proximity (where Euclidean distance is used).

In the process, two matrices of pairwise distances are computed: one of cosine similarities (transformed to distances using 1-similarity) and one of Euclidean distances. The last hybrid matrix is composed through equal weighting of the two matrices which have been normalized to the range between [0, 1].

$$Hybrid\ Distance = 0.5 \times Euclidean_{norm} + 0.5 \times (1 - Cosine_{norm})$$

This hybrid matrix replaces the internal distance computation step by using t-SNE metric="precomputed" option. The method can be applied to data of other forms and is fully interoperable with the already existing Scikit-learn functionality. This is a very effective design aspect in datasets such as financial risk analysis where linkages of global clusters are very important.

4.4.Dimensionality Configuration Design

The method allows the dimensionality of the t-SNE embeddings to be dynamically defined in order to perform larger testing. Depending on the goal of analysis or visualization, the outputs may be 2D, 3D, or 5D.

- The faster barnes_hut method is used to speed up computing both 2D and 3D embeddings.
- Due to the algorithmic shortcomings of Barnes-Hut with higher dimensions, the exact method is selected to be used in 5D embeddings.

The design enables all combinations (e.g. 5D embedding with PCA init and hybrid distance) to be tested with minimal code change due to the smooth integration with the initialization and distance modules.

This dimensional flexibility is needed not only in the case of data that is visually rich but also in the case of more abstract structured data where hidden patterns can emerge more clearly in higher-dimensional embeddings.

4.5.Perplexity Experiment Controller

The perplexity value in t-SNE affects the effective number of neighbors that each point considers. The design integrates a loop controller that will run t-SNE over a set series of perplexity values (e.g. 5 to 50) to perform a systematic assessment of the influence of this factor.

The controller:

- Accepts parameters such as initialization and distance.
- Goes through the list of the perplexity values.

- Saves intermediate results and silhouette scores.
- Generates line graphs to compare the performances between perplexities.

This aspect of design helps pinpoint the most optimal perplexity configurations of different datasets and arrangements and is crucial to understanding the strength of embeddings.

5 Implementation

The implementation step was done using Python 3.10 in Jupyter Notebook and translated the proposed enhancements into a modular pipeline. Notable libraries were Matplotlib and Seaborn to visualize, NumPy and Pandas to preprocess and Scikit-learn to do t-SNE, PCA and evaluation. The basic enhancements were incorporated into reusable functions to enable flexibility and reproducibility, and the system could be tested on the MNIST and Bankruptcy databases.

MNIST data set with 2,000 randomly drawn 784-dimensional vectors of 28x28 grayscale images in CSV format was loaded. The Taiwanese Bankruptcy dataset comprised 72 financial characteristics of which the target was binary indicating the bankruptcy status. In order to have consistent input in PCA and computing distances, the datasets were zero-mean and unit-variance normalized with 'StandardScaler'. With the help of parameterized functions, feature and label arrays were split and routed into the t-SNE pipeline.

A baseline was created using standard t-SNE with 'n_components=2', 'init='random'' and 'metric='euclidean''. Different runs led to different embeddings because of random initialization. Overlaps of clusters could be seen in the visual results and notably in the Bankruptcy dataset. The silhouette scores were used to measure cluster compactness.

PCA was employed to project the data into 2D or 5D regions depending on the t-SNE output setting so as to enhance reproducibility. The PCA output was fed to t-SNE with the help of the 'init' argument. Our deterministic approach provided consistent embeddings in terms of silhouette scores because it increased silhouette scores very slightly but steadily (scikit-learn developers; misha-pyshark 2020; Awan 2023).

To preserve local and global structure, normalized Euclidean distances and cosine distances (1 - similarity) were averaged to give a hybrid distance matrix and scaled to [0, 1], after which 'metric='precomputed'' was used to feed the matrix into t-SNE. The hybrid structure improved the layout continuity of the Bankruptcy dataset and generated more separated clusters of MNIST (HugoLasticot 2016; Contributor 2024; Levy et al. 2024).

t-SNE was configured to have 'n_components=5' and 'method='exact'' to have more rich embeddings. PCA was used both to initialize and to project the 5D output back to 3D to visualize it. This arrangement gave the best silhouette scores which indicates better clustering (3D scatterplot - Matplotlib 3.7.1 documentation).

The level of perplexity is different in 2D and 5D systems [5, 10, 20, 30, 40, 50]. Stability was measured with the help of silhouette ratings. The results showed that 5D embeddings were insensitive to perplexity but 2D was (Preserving global structure - openTSNE 1.0.0 documentation).

Since this was a modular system, all parameters of input, methods of initialization, and output dimensions could be varied arbitrarily. In the evaluation, scatter plots and silhouette scores were employed to give a qualitative and quantitative assessment of the quality of the embedding (Awan 2023; seaborn.scatterplot - seaborn 0.11.1 documentation).

6 Evaluation and Discussion

In this section, the outcomes of using various configurations of the t-SNE algorithm on two different datasets ‘Taiwanese Bankruptcy’ dataset and ‘MNIST’ handwritten digit dataset are presented. Besides qualitative observations in which visual inspection was suitable, the quality of each configuration was measured with the help of silhouette score, a quantitative measure of clustering quality. The experiments were performed to examine the impact of the initialization method (random and PCA), distance measure (Euclidean and hybrid), perplexity settings (5 to 50) and dimensionality of the embedding space (2D and 5D).

6.1 Baseline Performance: Standard t-SNE

Both datasets did not give reproducible results with default t-SNE using Euclidean distance and random initialization, and here we can observe the variation across runs of the same dataset. The baseline silhouette scores were however surprisingly high which means that there was initial cluster separation but not consistent in structure.

The traditional t-SNE setting of MNIST dataset led to a silhouette score of about 0.188 but due to stochastic initialization, the cluster shapes were not uniform when observed visually across trials. Similarly, the default t-SNE demonstrated that the silhouette score of the Bankruptcy dataset was about - 0.007, which indicates a low level of cluster formation. However, its reliability in terms of analysis was limited because it was not reproducible.

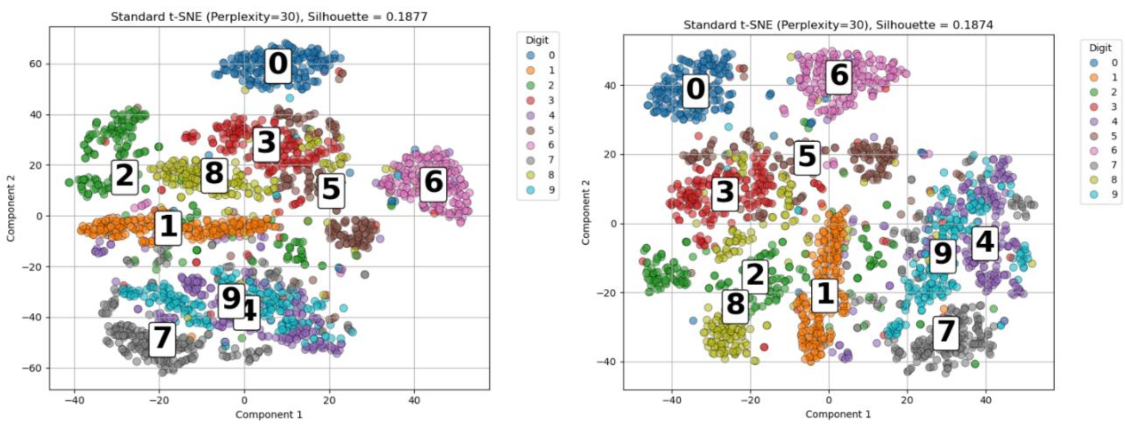


Figure 2: Baseline t-SNE Visuals of MNIST Dataset.

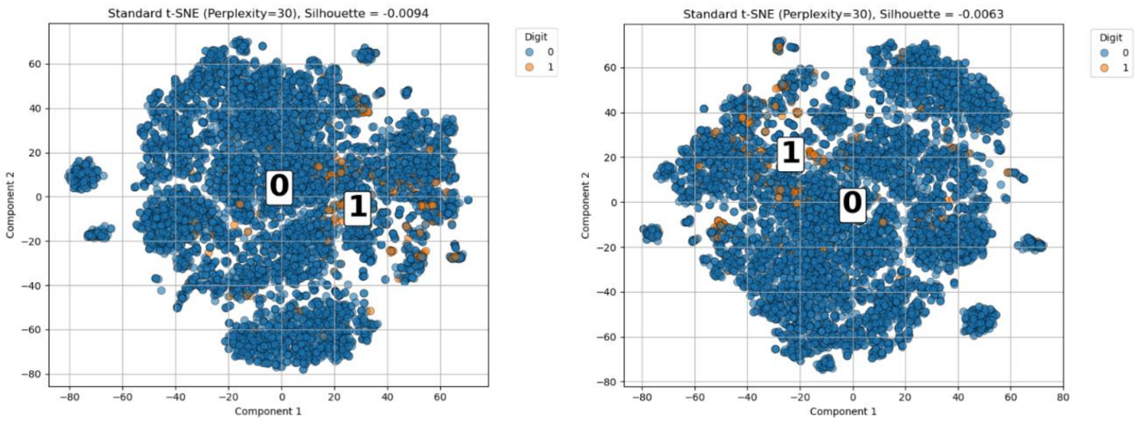


Figure 3: Baseline t-SNE Visuals of Bankruptcy Dataset.

6.2 PCA Initialization Effects

The results could be replicated between runs, and the spatial configuration of the clusters was more consistent when PCA was used as the initialisation method. Even though the 2D embeddings in the MNIST dataset were visually stable when initialized with PCA, they appeared cramped and compact, making it harder to pick out overlapping clusters in a visual manner. The disadvantage of this is that high-dimensional data, like photographs, can only be represented in two dimensions.

Nevertheless, PCA initialization was more favourable to the lower-dimensional, binary-labelled Bankruptcy dataset in terms of layout stability. The visual clarity and repeatability were much higher, although the silhouette score when using PCA alone was slightly less (-0.0166).

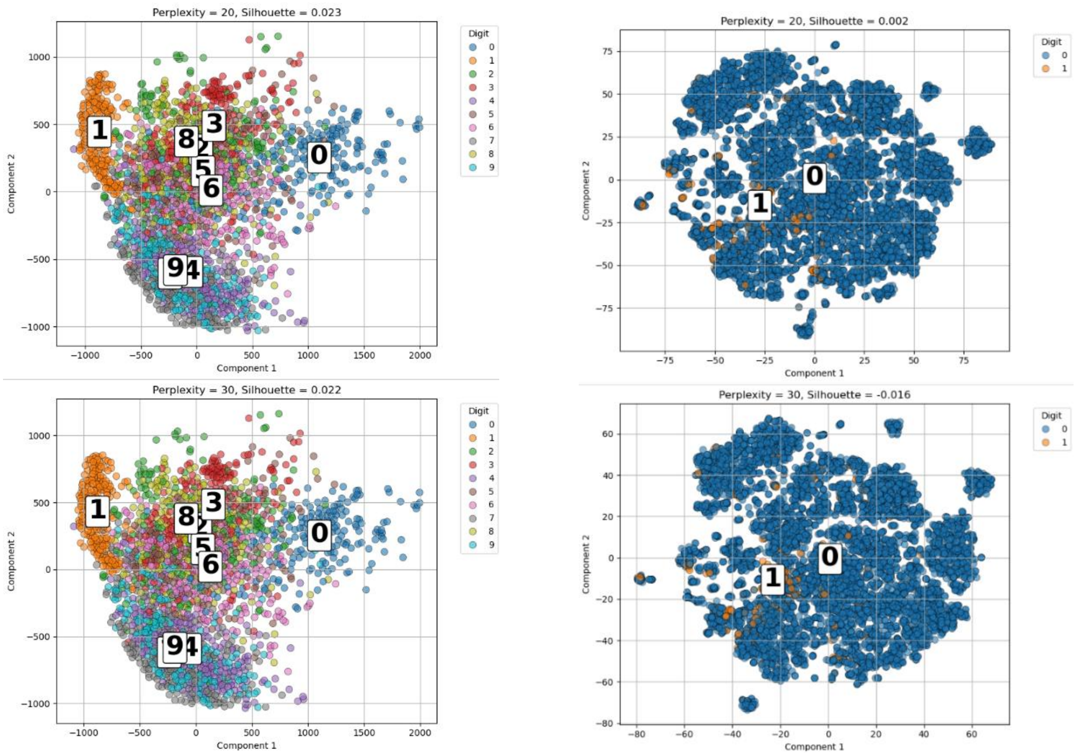


Figure 4: MNIST (Left) and Bankruptcy (Right) Data Visuals with PCA Initialization

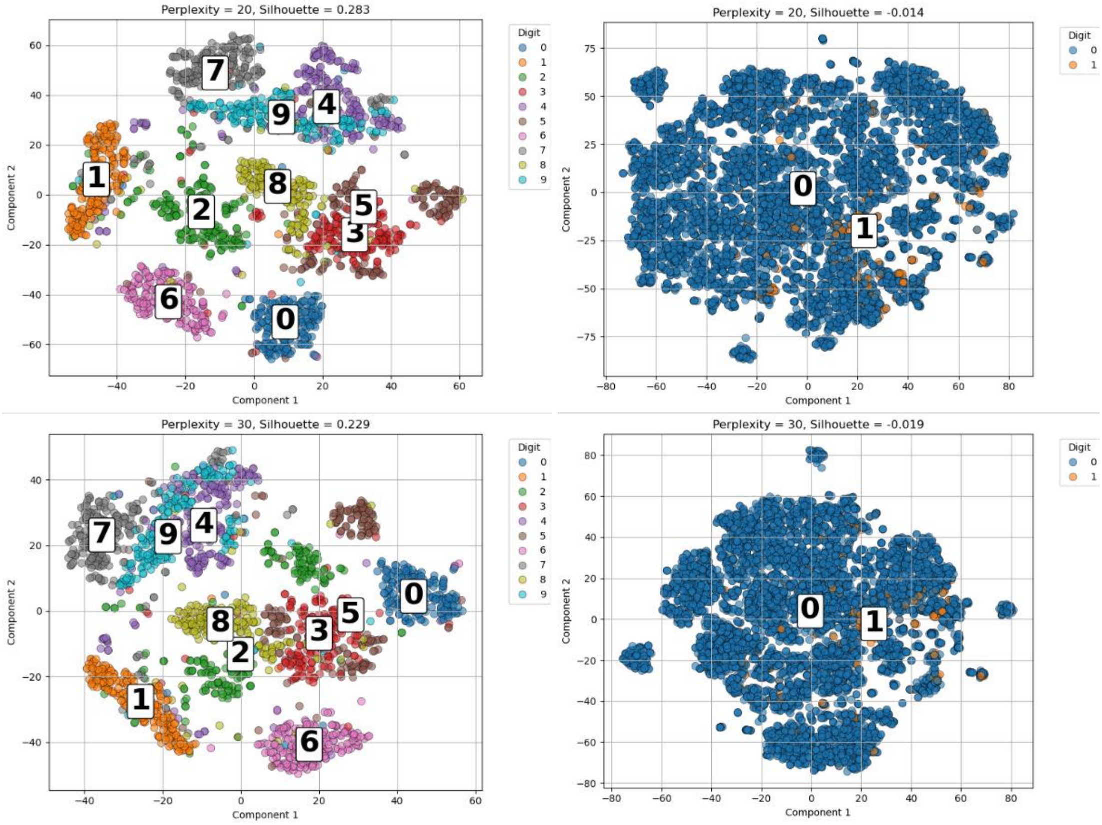


Figure 5: MNIST (Left) and Bankruptcy (Right) Data Visuals without PCA Initialization

6.3 Hybrid Distance Metrics Effects

The hybrid distance measure which is the combination of normalized Euclidean distance and cosine distance was tested with and without PCA initialization. Its goal was to maintain local and global connections in the embedded environment.

Bankruptcy Dataset Results:

With PCA Initialization (Figure 6):

- The highest silhouette score (0.0025) was found on Perplexity 20 and the other scores were -0.0035 to -0.0202.
- This shows that the clustering structure of the data was still bad even though there was visual improvement due to PCA and hybrid distance.

Without PCA Initialization (Figure 7):

- All silhouette scores were negative with the bigger the value the worse it is. At the highest perplexity of 5, the best score was 0.0010.
- Although the overall score was negative and less it was improved compared to standard t-SNE score.

```
Perplexity: 5, Silhouette Score: -0.0029
Perplexity: 10, Silhouette Score: -0.0035
Perplexity: 20, Silhouette Score: 0.0025
Perplexity: 30, Silhouette Score: -0.0161
Perplexity: 40, Silhouette Score: -0.0202
Perplexity: 50, Silhouette Score: -0.0099
```

Figure 6: Bankruptcy Data Silhouette Score with PCA Initialization on diff. Perplexity

```
Perplexity: 5, Silhouette Score: 0.0010
Perplexity: 10, Silhouette Score: -0.0100
Perplexity: 20, Silhouette Score: -0.0141
Perplexity: 30, Silhouette Score: -0.0192
Perplexity: 40, Silhouette Score: -0.0221
Perplexity: 50, Silhouette Score: -0.0098
```

Figure 7: Bankruptcy Data Silhouette Score without PCA Initialization on diff. Perplexity

MNIST Dataset Results:

With PCA Initialization (2D) (Figure 8):

- The highest silhouette scores occurred at the value of perplexity = 10 (0.0231), with all values of silhouette scores ranged closely between 0.0215 to 0.0231.
- These results indicate that there has been a slight decrease in cluster separation, and there was a cleaner aesthetic on layouts compared to the baseline.

With PCA Initialization (5D) (Figure 9):

- In all the perplexity settings, the silhouette score remained at 0.0734.
- This demonstrates that high dimensional embedding of visual data is robust and efficient, especially when combined with hybrid distance and PCA initialization.

Without PCA Initialization (Figure 10):

- The surprisingly high scores were in the silhouette:
  At perplexity = 20 the score was 0.2846, the highest.
  With higher perplexity the scores went down (0.1783 with perplexity = 50).
- This means that in certain setups, random initialization can lead to favourable cluster separation even at the cost of reproducibility.

```
Perplexity: 5, Silhouette Score: 0.0218
Perplexity: 10, Silhouette Score: 0.0229
Perplexity: 20, Silhouette Score: 0.0228
Perplexity: 30, Silhouette Score: 0.0225
Perplexity: 40, Silhouette Score: 0.0219
Perplexity: 50, Silhouette Score: 0.0214
```

Figure 8: MNIST Data Silhouette Score with PCA Initialization on diff. Perplexity

```
Running t-SNE with perplexity = 5 ...
Silhouette Score: 0.0734

Running t-SNE with perplexity = 10 ...
Silhouette Score: 0.0734

Running t-SNE with perplexity = 20 ...
Silhouette Score: 0.0734

Running t-SNE with perplexity = 30 ...
Silhouette Score: 0.0734

Running t-SNE with perplexity = 40 ...
Silhouette Score: 0.0734

Running t-SNE with perplexity = 50 ...
Silhouette Score: 0.0734
```

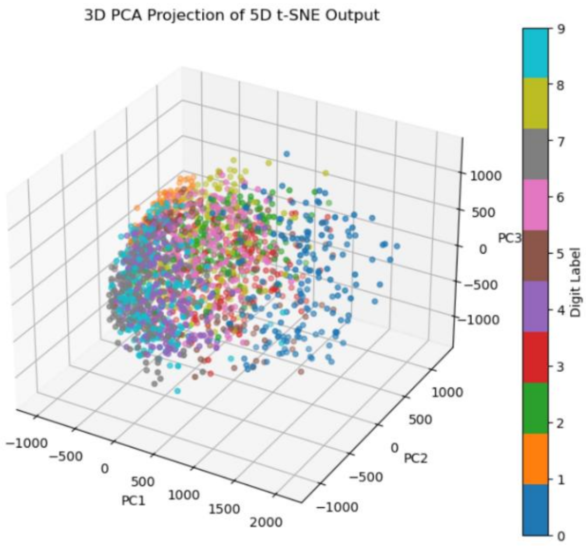
Figure 9: MNIST Data Silhouette Score with PCA Initialization (5D) on diff. Perplexity

```
Perplexity: 5, Silhouette Score: 0.2422
Perplexity: 10, Silhouette Score: 0.2773
Perplexity: 20, Silhouette Score: 0.2835
Perplexity: 30, Silhouette Score: 0.2292
Perplexity: 40, Silhouette Score: 0.1869
Perplexity: 50, Silhouette Score: 0.1809
```

Figure 10: MNIST Data Silhouette Score without PCA Initialization on diff. Perplexity

6.4 Dimensionality Analysis

The MNIST experiments pointed clear trend emerged, 5D embeddings were performing better than 2D by a consistent score in the silhouette scores. In all cases of perplexity, all 5D configurations that applied PCA, and hybrid distance had a consistent silhouette score of 0.0734. This is to show the significance of exploring higher-dimensional embeddings when analysing more complex data, like handwritten numbers. The 2D embeddings were interpretable visually but were compact and frequently led to overlapping clusters because they are compact.




```
Silhouette Score (5D t-SNE): 0.0734
```

Figure 11: 3D Visual and Silhouette Score of 5D t-SNE output on MNIST Data

6.5 Perplexity Sensitivity Analysis

In all tests, perplexity had more impact on 2D embeddings. The best 2D silhouette score of MNIST was Perplexity = 10 (0.231), with slight decreases at both lower and higher levels. The best result with PCA and hybrid distance on the Bankruptcy dataset was observed at Perplexity = 20 (0.0025), but the total values remained negative.

Conversely, the differences in perplexity were not observed to influence 5D embeddings, all runs yielded identical MNIST score (0.0734). This study found that embeddings in higher dimensions are less sensitive to neighbourhood size variations, which makes them a reliable choice in scenarios where it is difficult or costly to tune perplexity.

6.6 Dataset Specific Observations

The PCA and hybrid distance enhancements were positively impacted more on the MNIST dataset. The improvements yielded consistent clusters especially in 5D due to its inherent multi-class, high-dimensional nature. The t-SNE could achieve high silhouette scores with particular perplexity settings when PCA not used.

In contrast, the lower dimensional and two-class Bankruptcy dataset performed much better in terms of layout consistency but not in terms of numerical improvement. Although in some situations silhouette scores were still low or negative, it was important to initialize PCA to improve the visual clarity.

7 Conclusion and Future Work

The aim of this paper was to enhance the performance and interpretability of t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm by addressing three of its most widely known limitations: inability to preserve global structure and reproducibility. To test two major modifications to the standard t-SNE pipeline that were applied, including PCA-based deterministic initialization and hybrid distance metric, two fundamentally different datasets namely MNIST digit dataset and a Taiwanese financial bankruptcy dataset were utilized.

The outcomes of the experimental tests that have been conducted revealed that the PCA-based initialization significantly increased the replicability of the t-SNE embedding of the two datasets. However, the output embeddings seemed to be too compact whenever the PCA was performed on a 2D or 3D initialization space and ended up in clusters which were not distinguishable visually. This was the most notable in the example of the MNIST dataset where due to the dense 2D structure, putting cluster boundaries reliably was impossible. These reduced images suggest that PCA initialization may be more effective in the case of low-dimensional small-sized data, but it becomes problematic with high-dimensional data when viewed in two or three dimensions only.

Silhouette scores of the two datasets were improved when a global structure preserving distance measure was used. The three setups, which were subjected to the experiments, are two-dimensional embeddings initialized by PCA, five-dimensional embeddings initialized by PCA, and no PCA. For the MNIST dataset, PCA in 2D showed more stable clusters but silhouette scores did not improve compared to the standard t-SNE. In contrast, the combination of hybrid distance and PCA displayed a minor increase in silhouette score on the Bankruptcy dataset, nevertheless, the variance of the score over the perplexity parameters revealed sensitivity suggested due to the imbalance of the dataset classes.

Interestingly, even the hybrid distance settings which were not initialized with PCA did better than the standard t-SNE when it comes to silhouette scores on all the datasets. However, these came at the cost of non-reproducibility.

The advantage of 5D embeddings on MNIST dataset was an excellent discovery; under all the perplexity settings, the silhouette score reached steadily to 0.0734. It offered a more sensible and understandable framework with reduced variation as compared to the highest score of conventional t-SNE.

In conclusion:

- PCA initialization improves reproducibility but compresses visual clarity in low dimensional outputs.

- To ensure better retention of the structure, hybrid distance works best when dealing with high-dimensional embeddings.
- The best results in terms of silhouette improvements are achieved using hybrid distance without PCA but it is not repeatable.

It is possible to continue this research in the future in many ways. One of them is using Approximate Nearest Neighbour methods such as FAISS or Annoy to improve scalability when working with large datasets. In future, there is a possibility that the enhanced t-SNE framework can be compared to other dimensionality reduction techniques, such as TriMap, PaCMAP or UMAP. Finally, an assessment of the method on high dimensional, non-visual data could reveal the extent of its applicability.

Acknowledgement

I would like to give my gratitude to my supervisor, Prof. Christian Horn who gave critical advice, inspiration, and provided myself with a great deal of support throughout this research. His encouraging constructive criticism and wise advice was of great use to me when I was struggling to overcome all the challenges and refine my work to its final form.
I would like to express my gratitude to my parents, Mrs. and Mr. Shrikant Wadkar and my brother Sarthak Wadkar, who always supported me, understood, and believed in me and my abilities, and this was the foundation of my academic life. Without their eternal support and love nothing of this sort would have been achieved.

References

2.2. Manifold learning — scikit-learn 0.22.2 documentation. [no date]. Available at: https://scikit-learn.org/stable/modules/manifold.html.

3D scatterplot — Matplotlib 3.7.1 documentation. [no date]. Available at: https://matplotlib.org/stable/gallery/mplot3d/scatter3d.html.

Awan, A.A. 2023. Python t-SNE with Matplotlib. Available at: https://www.datacamp.com/tutorial/introduction-t-sne.

Chourasia, P., Ali, S. and Patterson, M. 2022. Informative Initialization and Kernel Selection Improves t-SNE for Biological Sequences. *2021 IEEE International Conference on Big Data (Big Data)*, pp. 101–106. Available at: https://ieeexplore.ieee.org/document/10020217.

Contributor, Stats.S. 2024. Dimensionality reduction and precomputed distance matrix. Available at: https://stats.stackexchange.com/questions/644469/dimensionality-reduction-and-precomputed-distance-matrix.

Cristian, P.-M. et al. 2024. Diffusion on PCA-UMAP Manifold: The Impact of Data Structure Preservation to Denoise High-Dimensional Single-Cell RNA Sequencing Data. *Biology* 13, p. 512. Available at: https://www.mdpi.com/2079-7737/13/7/512.

HugoLasticot. 2016. Cosine similarity TSNE in sklearn.manifold. Available at: https://stackoverflow.com/questions/36545434/cosine-similarity-tsne-in-sklearn-manifold.

Kobak, D. and Berens, P. 2019. The art of using t-SNE for single-cell transcriptomics. *Nature Communications* 10. Available at: https://www.nature.com/articles/s41467-019-13056-x.

Kobak, D. and Linderman, G.C. 2021. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology* 39, pp. 156–157. doi: 10.1038/s41587-020-00809-z.

Levy, A., Shalom, B. and Chalamish, M. 2024. A Guide to Similarity Measures. doi: 10.48550/arXiv.2408.07706.

Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S. and Kluger, Y. 2019. Fast Interpolation-based t-SNE for Improved Visualization of Single-Cell RNA-Seq Data. *Nature methods* 16, pp. 243–245. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6402590/.

Meyer, B.H., Trinidad, A. and Zola. 2022. Global and local structure preserving GPU t-SNE methods for large-scale applications. *Expert systems with applications* 201, p. 116918. doi: 10.1016/j.eswa.2022.116918.

misha-pyshark. 2020. Initialization of tSNE with PCA, allow for “auto” instead of “random.” Available at: https://github.com/scikit-learn/scikit-learn/issues/16986.

Preserving global structure — openTSNE 1.0.0 documentation. 2015. Available at: https://opentsne.readthedocs.io/en/latest/examples/03_preserving_global_structure/03_preserving_global_structure.html.

scikit-learn developers. 2014. sklearn.manifold.TSNE — scikit-learn 0.21.3 documentation. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.
seaborn.scatterplot — seaborn 0.11.1 documentation. [no date]. Available at: https://seaborn.pydata.org/generated/seaborn.scatterplot.html.

Tränkner, S. [no date]. Data Visualization with t-SNE in Theory and Practice. Available at: https://ins.uni-bonn.de/media/public/publication-media/Bachelorarbeit_Solveig_Traenkner.pdf.

Wang, Y., Huang, H., Rudin, C. and Shaposhnik, Y. 2021. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. *Journal of Machine Learning Research* 22, pp. 1–73. Available at: http://jmlr.org/papers/v22/20-1061.html.