Comparing Optimal Transport-based and Traditional Manifold Learning Algorithms in Image Classification

Full Name: Phuong Trinh

Report Submission Date: May 2, 2024

1. Introduction:

This report presents a comprehensive analysis of the effectiveness of Optimal Transport-based and traditional manifold learning algorithms in image classification tasks. We explore whether Optimal Transport-based manifold learning enhances the accuracy and efficiency of image classification compared to traditional methods. Our experiments span various datasets, each offering unique insights into the application and benefits of manifold learning techniques. These datasets include Fashion MNIST, Handwritten MNIST, Dog Breed Classification, and Coil-100, each processed with tailored approaches to meet their specific challenges.

2. Literature Review:

Manifold learning techniques are pivotal in reducing high-dimensional data to lower-dimensional spaces while preserving intrinsic geometrical properties, making them particularly suitable for image classification tasks. Traditional methods like MDS, Isomap, t-SNE, LLE and spectral embeddings focus on capturing global or local data structures but may not fully encapsulate the manifold's complexity. Recent literature suggests that Optimal Transport-based methods, which consider the data as distributions and seek to minimize transportation costs between these distributions, could offer enhanced performance by better aligning data in a meaningful embedding space. Despite the promising applications of these advanced techniques, direct comparisons in supervised learning settings remain limited, underscoring a significant gap this research aims to fill.

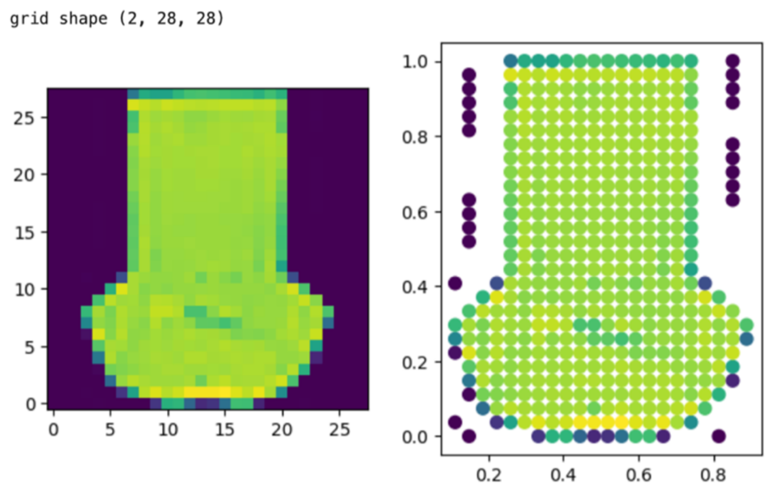
3. Data Understanding:

- Fashion MNIST: Fashion MNIST comprises 60,000 grayscale images categorized into 10 fashion items. This dataset serves as a benchmark for evaluating manifold learning techniques in a controlled environment, enabling the assessment of embedding effectiveness across distinct fashion categories.

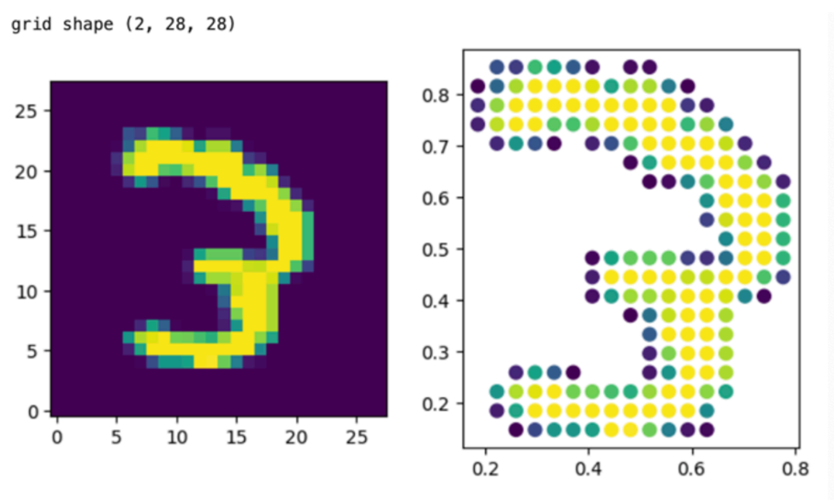
- Handwritten MNIST: Handwritten MNIST consists of 70,000 grayscale images representing digits from 0 to 9. It serves as a standard dataset for evaluating the algorithms' ability to preserve class separations post-embedding, crucial for accurate digit recognition.
- Dog Breed Classification: The Dog Breed Classification dataset consists of a diverse collection of dog images categorized into different breeds. This dataset presents challenges due to variations in image sizes, resolutions, and backgrounds, making it ideal for evaluating the robustness of manifold learning techniques in handling complex classification tasks involving high intra-class and low inter-class variance.
- Coil-100: Coil-100 contains 7,200 color images depicting 100 different objects captured from various angles. This dataset poses challenges due to the three-dimensional nature of the objects, requiring manifold learning algorithms to effectively capture and represent object viewpoints in lower-dimensional embedding spaces.

4. Data Preprocessing:

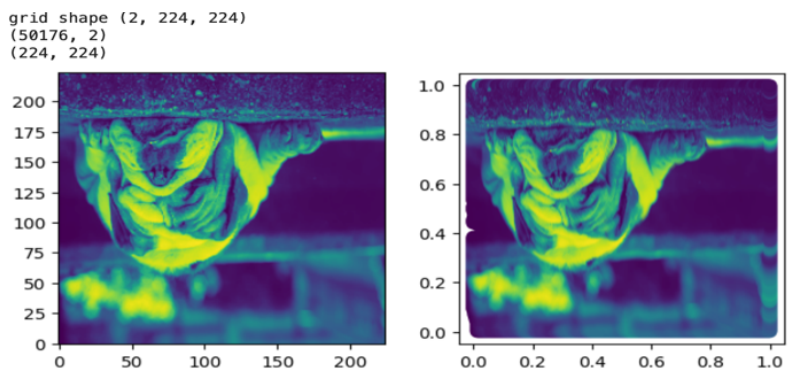
- Fashion MNIST: To prepare the Fashion MNIST dataset for analysis, images were standardized to a uniform size and normalized to ensure consistent pixel values across all samples. These preprocessing steps are essential for mitigating variations in image characteristics and facilitating effective manifold learning. Additionally, the images were converted into point cloud representations using the Dr. Hamm function to prepare them for Wassmap embeddings. A sample size of n = 1000-1500 was selected from the Fashion MNIST dataset for further analysis.



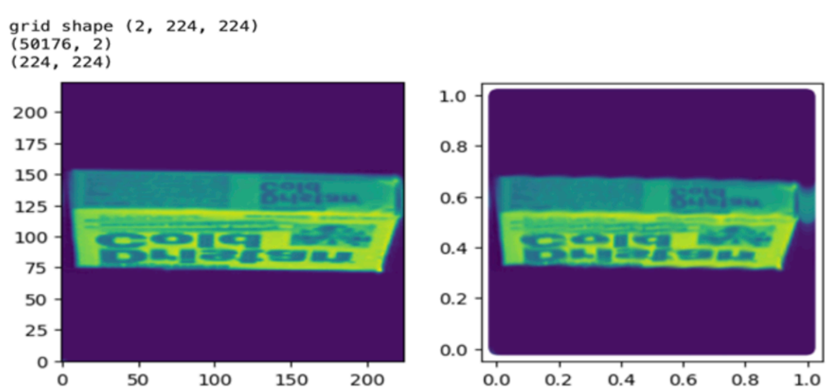
- Handwritten MNIST: Similar to the preprocessing of Fashion MNIST, the Handwritten MNIST dataset underwent resizing to uniform dimensions and normalization of pixel values. These steps are vital for maintaining data integrity and facilitating accurate classification of handwritten digits. The images were then transformed into point cloud representations using the Dr. Hamm function to facilitate the application of Wassmap embeddings. A sample size of n = 1000-1500 was chosen from the Handwritten MNIST dataset for embedding analysis.



- Dog Breed Classification: Preprocessing of the Dog Breed Classification dataset involved resizing images to a standardized format and applying techniques to address potential data imbalances across different breed categories. Additionally, normalization of pixel values was performed to ensure uniformity in image representations by convert all the images to grayscale, facilitating the effective application of manifold learning algorithms. To prepare for Wassmap embeddings, 5-10 images were selected from each class within the Dog Breed dataset and transformed into point cloud representations using the Dr. Hamm function.



- Coil-100: Prior to preprocessing, images were organized by object number for labeling clarity. To simplify the data and reduce computational complexity, the images were converted to grayscale, removing color information while retaining essential object features. Subsequently, the images were resized to a standardized size of 224x224 pixels to ensure uniformity across the dataset. This resizing step was essential for preparing the dataset for manifold learning analysis and subsequent machine learning model training, facilitating consistent and efficient processing. Furthermore, to facilitate the application of Wassmap embeddings, 5-10 images were chosen from each object class within the Coil-100 dataset and transformed into point cloud representations using the Dr. Hamm function.



5. Methodology:

This study employs a comparative analysis design to assess the effectiveness of various manifold learning techniques, including traditional embeddings and advanced Optimal Transport-based embeddings (Wassmap), in improving image classification accuracy across multiple datasets: Dog Breed Classification, COIL-100, Fashion-MNIST, and Handwritten MNIST.

Manifold Learning Techniques:

The embeddings are applied in this study include Multidimensional Scaling (MDS), Isometric Mapping (IsoMap), t-Distributed Stochastic Neighbor Embedding (t-SNE), Locally Linear Embedding (LLE), and Spectral Embedding. These are applied using a Euclidean metric to facilitate the exploration of the data's intrinsic geometric structure. After applying on the Euclidean metric, we conducted the same embeddings on Wasserstein distance metric (Wassmap) to compare the accuracy score and t-test performance.

Classification Algorithms:

Classification is performed using Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest, and Multinomial Logistic Regression. The effectiveness of these models is evaluated based on accuracy scores obtained from embedded and non-embedded data.

Statistical Analysis:

To rigorously test the effectiveness of these embeddings, we conduct two-sample t-tests comparing the mean accuracy scores of models using embeddings versus those without embeddings. The hypotheses tested are as follows:

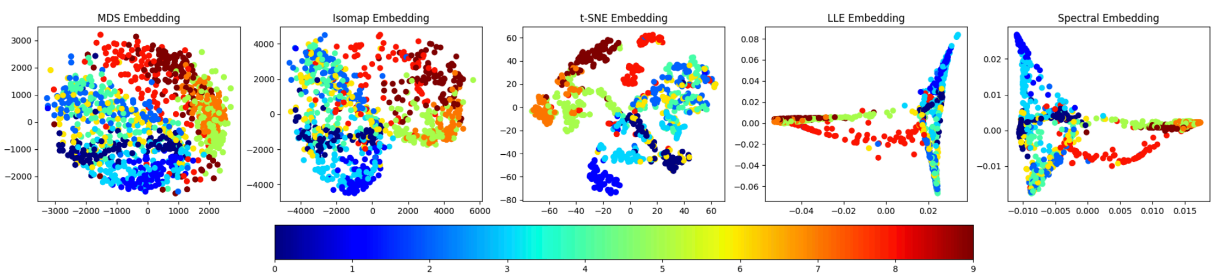
Null Hypothesis (H_0): The performance of classification models using embeddings is less than or equal to those without embeddings ($\mu_{embeddings} \leq \mu_{no\ embeddings}$)

Alternative Hypothesis (H_a): The performance of classification models using embeddings is better than those without embeddings $\left(\mu_{embeddings} > \mu_{no\ embeddings}\right)$

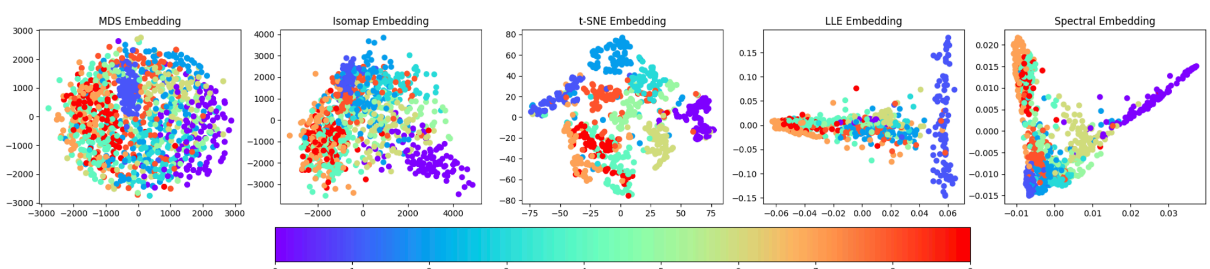
The purpose of the t-test is to determine if the use of manifold learning techniques statistically improves the classification accuracy, providing evidence for or against the added value of embeddings in the classification process.

6. Results:

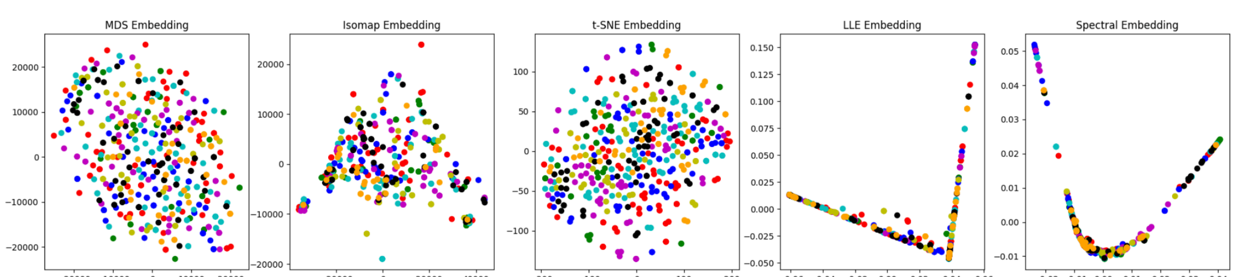
- Euclidean Embeddings Visualizations:
    - Fashion MNIST:



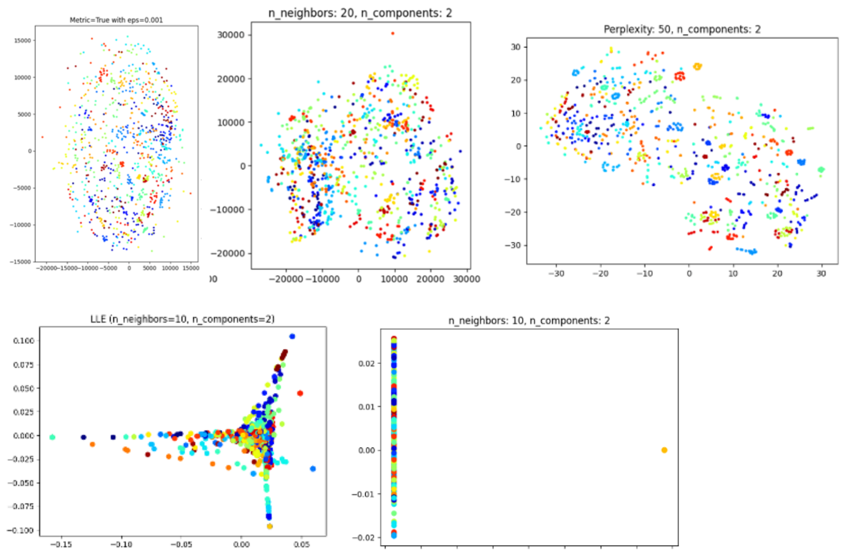
-
    - Handwritten MNIST:



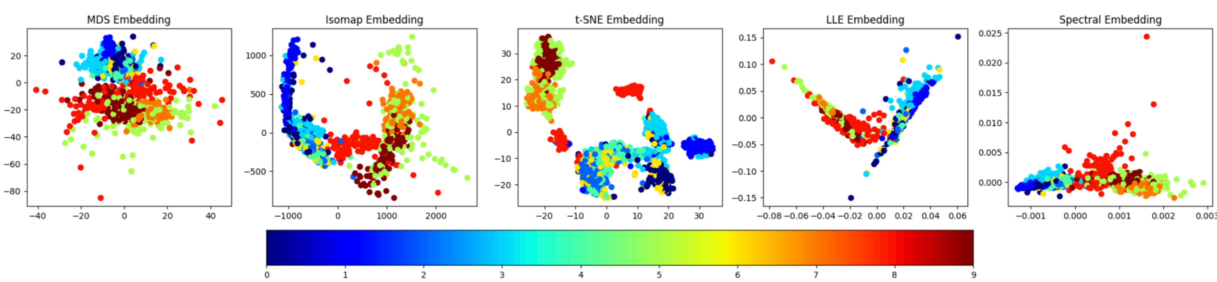
-
    - Dog Breed:



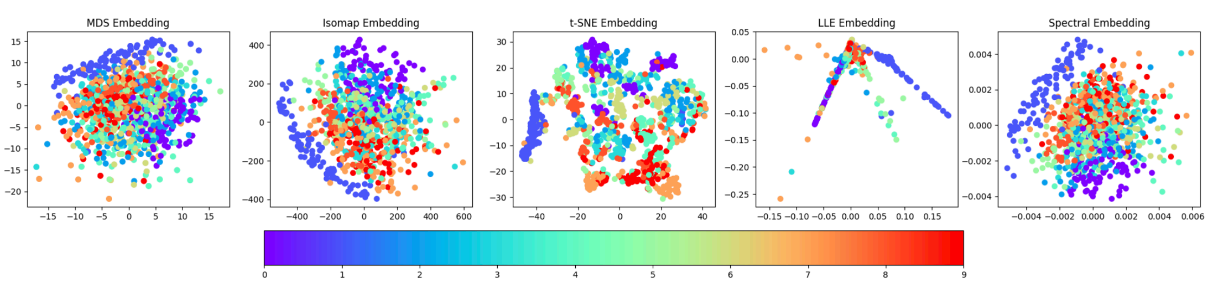
-
    - Coil-100:



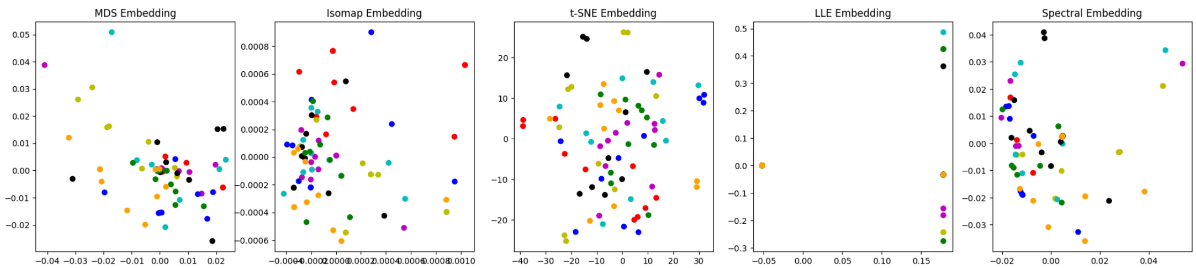
- Wassmap Embeddings Visualization:
  - Fashion MNIST:



  - Handwritten MNIST:



  - Dog Breed:



Because of time-consuming, we do not have enough time to finish Wassmap embeddings on Coil-100 dataset.

Accuracy Score and T-test performance on the Euclidean Embeddings:

- Fashion MNIST
  - Accuracy Score:

```
combined_df_10 = pd.concat([no_embedding_df, mds_df_10, isomap_df_10, tsne_df, lle_df_10, spectral_df_10])
combined_df_10
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.40 ± 0.00	0.76 ± 0.00	0.72 ± 0.00	0.68 ± 0.00	0.82 ± 0.00	0.74 ± 0.00	0.76 ± 0.00	0.80 ± 0.00
MDS	0.71 ± 0.01	0.69 ± 0.01	0.71 ± 0.00	0.71 ± 0.01	NaN	NaN	0.69 ± 0.01	0.72 ± 0.01
Isomap	0.68 ± 0.01	0.69 ± 0.01	0.70 ± 0.02	0.71 ± 0.02	NaN	NaN	0.71 ± 0.01	0.71 ± 0.02
t-SNE	0.58 ± 0.07	0.71 ± 0.03	0.72 ± 0.04	0.72 ± 0.03	0.60 ± 0.09	0.60 ± 0.08	0.73 ± 0.03	0.59 ± 0.08
LLE	0.58 ± 0.08	0.67 ± 0.01	0.70 ± 0.02	0.69 ± 0.02	NaN	NaN	0.69 ± 0.03	0.51 ± 0.09
Spectral Embedding	0.64 ± 0.01	0.65 ± 0.03	0.68 ± 0.03	0.69 ± 0.02	NaN	NaN	0.69 ± 0.02	0.31 ± 0.12

```
combined_df_100 = pd.concat([no_embedding_df, mds_df_100, isomap_df_100, tsne_df, lle_df_100, spectral_df_100])
combined_df_100
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.40 ± 0.00	0.76 ± 0.00	0.72 ± 0.00	0.68 ± 0.00	0.82 ± 0.00	0.74 ± 0.00	0.76 ± 0.00	0.80 ± 0.00
MDS	0.72 ± 0.01	0.69 ± 0.01	0.72 ± 0.02	0.70 ± 0.01	0.73 ± 0.01	0.73 ± 0.01	0.71 ± 0.01	0.68 ± 0.01
Isomap	0.73 ± 0.02	0.70 ± 0.01	0.69 ± 0.01	0.70 ± 0.02	0.74 ± 0.01	0.73 ± 0.00	0.72 ± 0.01	0.68 ± 0.02
t-SNE	0.58 ± 0.07	0.71 ± 0.03	0.72 ± 0.04	0.72 ± 0.03	0.60 ± 0.09	0.60 ± 0.08	0.73 ± 0.03	0.59 ± 0.08
LLE	0.75 ± 0.03	0.70 ± 0.03	0.72 ± 0.02	0.73 ± 0.02	0.51 ± 0.03	0.74 ± 0.02	0.74 ± 0.01	0.66 ± 0.02
Spectral Embedding	0.73 ± 0.02	0.69 ± 0.01	0.72 ± 0.02	0.72 ± 0.01	0.12 ± 0.00	0.72 ± 0.01	0.73 ± 0.01	0.34 ± 0.14

  - T-test performance:

```
One-Sided T-Test Results for 10 Components:
MDS: p-value = 0.6141
  No significant difference, fail to reject H0
Isomap: p-value = 0.5837
  No significant difference, fail to reject H0
t-SNE: p-value = 0.4325
  No significant difference, fail to reject H0
LLE: p-value = 0.2551
  No significant difference, fail to reject H0
Spectral Embedding: p-value = 0.1943
  No significant difference, fail to reject H0
```

```
One-Sided T-Test Results for 100 Components:
MDS: p-value = 0.5000
  No significant difference, fail to reject H0
Isomap: p-value = 0.5101
  No significant difference, fail to reject H0
t-SNE: p-value = 0.1658
  No significant difference, fail to reject H0
LLE: p-value = 0.3859
  No significant difference, fail to reject H0
Spectral Embedding: p-value = 0.1282
  No significant difference, fail to reject H0
```

- Handwritten MNIST
  - Accuracy Score:

```
combined_df_10 = pd.concat([no_embedding_df, mds_df_10, isomap_df_10, tsne_df, lle_df_10, spectral_df_10])
combined_df_10
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	1.00 ± 0.00	1.00 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	1.00 ± 0.00	0.97 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
MDS	0.73 ± 0.01	0.77 ± 0.00	0.75 ± 0.03	0.76 ± 0.00	0.73 ± 0.02	0.80 ± 0.01	0.75 ± 0.01	0.74 ± 0.04
Isomap	0.77 ± 0.02	0.75 ± 0.01	0.75 ± 0.01	0.76 ± 0.01	0.76 ± 0.01	0.81 ± 0.02	0.80 ± 0.03	0.75 ± 0.01
t-SNE	0.43 ± 0.15	0.79 ± 0.02	0.72 ± 0.04	0.71 ± 0.08	0.51 ± 0.16	0.64 ± 0.18	0.79 ± 0.03	0.45 ± 0.16
LLE	0.59 ± 0.06	0.70 ± 0.03	0.67 ± 0.03	0.65 ± 0.05	0.08 ± 0.01	0.64 ± 0.08	0.72 ± 0.02	0.39 ± 0.03
Spectral Embedding	0.70 ± 0.06	0.79 ± 0.02	0.81 ± 0.03	0.79 ± 0.04	0.07 ± 0.00	0.79 ± 0.06	0.81 ± 0.03	0.16 ± 0.07

```
combined_df_100 = pd.concat([no_embedding_df, mds_df_100, isomap_df_100, tsne_df, lle_df_100, spectral_df_100])
combined_df_100
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	1.00 ± 0.00	1.00 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	1.00 ± 0.00	0.97 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
MDS	0.73 ± 0.07	0.73 ± 0.03	0.72 ± 0.06	0.72 ± 0.04	0.81 ± 0.05	0.84 ± 0.02	0.72 ± 0.06	0.78 ± 0.08
Isomap	0.74 ± 0.01	0.71 ± 0.02	0.67 ± 0.03	0.67 ± 0.04	0.79 ± 0.03	0.84 ± 0.01	0.78 ± 0.01	0.78 ± 0.02
t-SNE	0.43 ± 0.15	0.79 ± 0.02	0.72 ± 0.04	0.71 ± 0.08	0.51 ± 0.16	0.64 ± 0.18	0.79 ± 0.03	0.45 ± 0.16
LLE	0.84 ± 0.02	0.75 ± 0.00	0.74 ± 0.01	0.75 ± 0.02	0.41 ± 0.00	0.82 ± 0.02	0.81 ± 0.01	0.65 ± 0.02
Spectral Embedding	0.70 ± 0.06	0.79 ± 0.02	0.81 ± 0.03	0.79 ± 0.04	0.07 ± 0.00	0.79 ± 0.06	0.82 ± 0.02	0.16 ± 0.07

  - T-test performance:

```
One-Sided T-Test Results for 10 Components:
MDS: p-value = 1.0000
  No significant difference, fail to reject H0
Isomap: p-value = 1.0000
  No significant difference, fail to reject H0
t-SNE: p-value = 0.9999
  No significant difference, fail to reject H0
LLE: p-value = 0.9996
  No significant difference, fail to reject H0
Spectral Embedding: p-value = 0.9935
  No significant difference, fail to reject H0
```

```
One-Sided T-Test Results for 100 Components:
MDS: p-value = 1.0000
  No significant difference, fail to reject H0
Isomap: p-value = 1.0000
  No significant difference, fail to reject H0
t-SNE: p-value = 0.9999
  No significant difference, fail to reject H0
LLE: p-value = 0.9995
  No significant difference, fail to reject H0
Spectral Embedding: p-value = 0.9933
  No significant difference, fail to reject H0
```

- Dog Breeds:
  - Accuracy Score:

```
combined_df_10 = pd.concat([no_embedding_df, mds_df_10, isomap_df_10, tsne_df_10, lle_df_10, spectral_df_10])
combined_df_10
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.61 ± 0.00	0.70 ± 0.00	0.47 ± 0.00	0.33 ± 0.00	0.76 ± 0.00	0.72 ± 0.00	0.73 ± 0.00	0.76 ± 0.00
MDS	0.28 ± 0.02	0.68 ± 0.01	0.47 ± 0.01	0.32 ± 0.00	NaN	NaN	0.70 ± 0.02	0.29 ± 0.01
Isomap	0.26 ± 0.02	0.69 ± 0.03	0.45 ± 0.02	0.33 ± 0.05	NaN	NaN	0.71 ± 0.02	0.26 ± 0.05
t-SNE	0.20 ± 0.03	0.72 ± 0.02	0.50 ± 0.01	0.38 ± 0.01	0.19 ± 0.00	0.25 ± 0.01	0.60 ± 0.03	0.20 ± 0.03
LLE	0.22 ± 0.03	0.65 ± 0.05	0.46 ± 0.07	0.34 ± 0.09	NaN	NaN	0.62 ± 0.07	0.16 ± 0.01
Spectral Embedding	0.23 ± 0.02	0.67 ± 0.04	0.45 ± 0.01	0.33 ± 0.01	NaN	NaN	0.65 ± 0.07	0.14 ± 0.00

```
combined_df_100 = pd.concat([no_embedding_df, mds_df_100, isomap_df_100, tsne_df_10, lle_df_100, spectral_df_100])
combined_df_100
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.61 ± 0.00	0.70 ± 0.00	0.47 ± 0.00	0.33 ± 0.00	0.76 ± 0.00	0.72 ± 0.00	0.73 ± 0.00	0.76 ± 0.00
MDS	0.51 ± 0.01	0.69 ± 0.01	0.47 ± 0.02	0.32 ± 0.01	0.69 ± 0.02	0.67 ± 0.01	0.70 ± 0.02	0.57 ± 0.02
Isomap	0.54 ± 0.03	0.70 ± 0.03	0.46 ± 0.02	0.30 ± 0.02	0.69 ± 0.01	0.60 ± 0.07	0.70 ± 0.02	0.67 ± 0.01
t-SNE	0.20 ± 0.03	0.72 ± 0.02	0.50 ± 0.01	0.38 ± 0.01	0.19 ± 0.00	0.25 ± 0.01	0.60 ± 0.03	0.20 ± 0.03
LLE	0.63 ± 0.01	0.70 ± 0.02	0.47 ± 0.03	0.33 ± 0.07	0.26 ± 0.06	0.63 ± 0.02	0.68 ± 0.02	0.45 ± 0.03
Spectral Embedding	0.61 ± 0.03	0.71 ± 0.02	0.49 ± 0.02	0.37 ± 0.03	0.14 ± 0.00	0.64 ± 0.03	0.71 ± 0.02	0.19 ± 0.08

  - T-test performance:

```
One-Sided T-Test Results for 10 Components:
MDS: p-value = 0.8991
  No significant difference, fail to reject H0
Isomap: p-value = 0.9007
  No significant difference, fail to reject H0
t-SNE: p-value = 0.9174
  No significant difference, fail to reject H0
LLE: p-value = 0.9463
  No significant difference, fail to reject H0
Spectral Embedding: p-value = 0.9362
  No significant difference, fail to reject H0
```

```
One-Sided T-Test Results for 100 Components:
MDS: p-value = 0.7758
  No significant difference, fail to reject H0
Isomap: p-value = 0.7518
  No significant difference, fail to reject H0
t-SNE: p-value = 0.9923
  No significant difference, fail to reject H0
LLE: p-value = 0.9138
  No significant difference, fail to reject H0
Spectral Embedding: p-value = 0.9285
  No significant difference, fail to reject H0
```

- Coil-100:
  - Accuracy Score:

```
combined_df_10 = pd.concat([no_embedding, mds_df_10, isomap_df_10, tsne_df, lle_df_10, spectral_df_10])
combined_df_10
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.75 ± 0.00	0.81 ± 0.00	0.69 ± 0.00	0.58 ± 0.00	0.84 ± 0.00	0.58 ± 0.00	0.80 ± 0.00	0.77 ± 0.00
MDS	0.64 ± 0.00	0.69 ± 0.00	0.62 ± 0.00	0.58 ± 0.00	0.72 ± 0.00	0.64 ± 0.00	0.63 ± 0.00	0.57 ± 0.00
Isomap	0.54 ± 0.03	0.68 ± 0.03	0.61 ± 0.01	0.56 ± 0.01	0.69 ± 0.02	0.49 ± 0.02	0.67 ± 0.01	0.57 ± 0.02
t-SNE	0.41 ± 0.07	0.78 ± 0.01	0.70 ± 0.02	0.63 ± 0.02	0.58 ± 0.04	0.34 ± 0.08	0.73 ± 0.05	0.45 ± 0.05
LLE	0.43 ± 0.06	0.67 ± 0.01	0.61 ± 0.02	0.57 ± 0.02	0.01 ± 0.00	0.41 ± 0.07	0.57 ± 0.03	0.03 ± 0.01
Spectral Embedding	0.22 ± 0.17	0.64 ± 0.03	0.60 ± 0.01	0.56 ± 0.03	0.01 ± 0.00	0.29 ± 0.02	0.64 ± 0.02	0.01 ± 0.00

```
combined_df_100 = pd.concat([no_embedding, mds_df_100, isomap_df_10, tsne_df, lle_df_100, spectral_df_100])
combined_df_100
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.75 ± 0.00	0.81 ± 0.00	0.69 ± 0.00	0.58 ± 0.00	0.84 ± 0.00	0.58 ± 0.00	0.80 ± 0.00	0.77 ± 0.00
MDS	0.66 ± 0.00	0.68 ± 0.00	0.58 ± 0.00	0.54 ± 0.00	0.76 ± 0.00	0.62 ± 0.00	0.62 ± 0.01	0.67 ± 0.00
Isomap	0.54 ± 0.03	0.68 ± 0.03	0.61 ± 0.01	0.56 ± 0.01	0.69 ± 0.02	0.49 ± 0.02	0.67 ± 0.01	0.57 ± 0.02
t-SNE	0.41 ± 0.07	0.78 ± 0.01	0.70 ± 0.02	0.63 ± 0.02	0.58 ± 0.04	0.34 ± 0.08	0.73 ± 0.05	0.45 ± 0.05
LLE	0.43 ± 0.06	0.67 ± 0.01	0.61 ± 0.02	0.57 ± 0.02	0.01 ± 0.00	0.41 ± 0.07	0.56 ± 0.04	0.03 ± 0.01
Spectral Embedding	0.36 ± 0.31	0.72 ± 0.01	0.67 ± 0.02	0.61 ± 0.02	0.01 ± 0.00	0.65 ± 0.03	0.71 ± 0.02	0.03 ± 0.01

  - T-test performance:

```
One-Sided T-Test Results for 10 Components:
MDS: p-value = 0.9774
  No significant difference, fail to reject H0
Isomap: p-value = 0.9931
  No significant difference, fail to reject H0
t-SNE: p-value = 0.9768
  No significant difference, fail to reject H0
LLE: p-value = 0.9949
  No significant difference, fail to reject H0
Spectral Embedding: p-value = 0.9963
  No significant difference, fail to reject H0
```

```
One-Sided T-Test Results for 100 Components:
MDS: p-value = 0.9661
  No significant difference, fail to reject H0
Isomap: p-value = 0.9931
  No significant difference, fail to reject H0
t-SNE: p-value = 0.9768
  No significant difference, fail to reject H0
LLE: p-value = 0.9950
  No significant difference, fail to reject H0
Spectral Embedding: p-value = 0.9759
  No significant difference, fail to reject H0
```

Accuracy Score and T-test performance on the Wassmap Embeddings:

- Fashion MNIST:
  - Accuracy Score:

```python
combined_df = pd.concat([no_embedding_df, mds_df, isomap_df, tsne_df, lle_df, spectral_df])
combined_df
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.59 ± 0.00	0.76 ± 0.00	0.75 ± 0.00	0.79 ± 0.00	0.78 ± 0.00	0.79 ± 0.00	0.82 ± 0.00	0.74 ± 0.00
MDS	0.66 ± 0.01	0.63 ± 0.01	0.65 ± 0.00	0.67 ± 0.01	0.67 ± 0.01	0.69 ± 0.00	0.69 ± 0.02	0.67 ± 0.01
Isomap	0.66 ± 0.00	0.70 ± 0.00	0.69 ± 0.01	0.71 ± 0.01	0.70 ± 0.01	0.73 ± 0.01	0.71 ± 0.01	0.69 ± 0.02
t-SNE	0.53 ± 0.05	0.63 ± 0.01	0.63 ± 0.01	0.64 ± 0.01	0.59 ± 0.04	0.68 ± 0.02	0.65 ± 0.01	0.55 ± 0.04
LLE	0.43 ± 0.10	0.58 ± 0.01	0.57 ± 0.01	0.59 ± 0.01	0.10 ± 0.01	0.58 ± 0.02	0.61 ± 0.02	0.35 ± 0.15
Spectral Embedding	0.58 ± 0.08	0.56 ± 0.02	0.58 ± 0.03	0.61 ± 0.04	0.09 ± 0.00	0.63 ± 0.07	0.59 ± 0.04	0.33 ± 0.10

```python
combined_df_100 = pd.concat([no_embedding_df, mds_df_100, isomap_df_100, tsne_df, lle_df_100, spectral_df_100])
combined_df_100
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.59 ± 0.00	0.76 ± 0.00	0.75 ± 0.00	0.79 ± 0.00	0.78 ± 0.00	0.79 ± 0.00	0.82 ± 0.00	0.74 ± 0.00
MDS	0.68 ± 0.02	0.63 ± 0.00	0.64 ± 0.00	0.67 ± 0.00	0.63 ± 0.01	0.68 ± 0.00	0.72 ± 0.00	0.60 ± 0.02
Isomap	0.72 ± 0.00	0.72 ± 0.01	0.72 ± 0.00	0.72 ± 0.01	0.76 ± 0.01	0.75 ± 0.01	0.76 ± 0.01	0.76 ± 0.01
t-SNE	0.53 ± 0.05	0.63 ± 0.01	0.63 ± 0.01	0.64 ± 0.01	0.59 ± 0.04	0.68 ± 0.02	0.65 ± 0.01	0.55 ± 0.04
LLE	0.67 ± 0.02	0.61 ± 0.03	0.61 ± 0.03	0.63 ± 0.02	0.38 ± 0.04	0.68 ± 0.01	0.65 ± 0.02	0.63 ± 0.01
Spectral Embedding	0.69 ± 0.02	0.59 ± 0.02	0.61 ± 0.02	0.64 ± 0.02	0.09 ± 0.00	0.68 ± 0.02	0.64 ± 0.01	0.44 ± 0.14

- T-test performance:

```
One-Sided T-Test Results for 100 Components:
MDS: p-value = 0.0031
  Significant difference at p < 0.05, reject H0
Isomap: p-value = 0.3050
  No significant difference, fail to reject H0
t-SNE: p-value = 0.0003
  Significant difference at p < 0.05, reject H0
LLE: p-value = 0.0021
  Significant difference at p < 0.05, reject H0
Spectral Embedding: p-value = 0.0120
  Significant difference at p < 0.05, reject H0
```

```
One-Sided T-Test Results for 10 Components:
MDS: p-value = 0.0051
  Significant difference at p < 0.05, reject H0
Isomap: p-value = 0.0356
  Significant difference at p < 0.05, reject H0
t-SNE: p-value = 0.0003
  Significant difference at p < 0.05, reject H0
LLE: p-value = 0.0013
  Significant difference at p < 0.05, reject H0
Spectral Embedding: p-value = 0.0030
  Significant difference at p < 0.05, reject H0
```

- Handwritten MNIST:
  - Accuracy Score:

```python
combined_df = pd.concat([no_embedding_df, mds_df_10, isomap_df_10, tsne_df, lle_df_10, spectral_df_10])
combined_df
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	1.00 ± 0.00	1.00 ± 0.00	0.92 ± 0.00	0.90 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
MDS	0.72 ± 0.02	0.63 ± 0.01	0.66 ± 0.01	0.68 ± 0.01	0.72 ± 0.02	0.75 ± 0.01	0.70 ± 0.02	0.72 ± 0.01
Isomap	0.84 ± 0.00	0.85 ± 0.01	0.85 ± 0.01	0.87 ± 0.00	0.85 ± 0.01	0.88 ± 0.00	0.85 ± 0.01	0.86 ± 0.01
t-SNE	0.44 ± 0.07	0.69 ± 0.01	0.68 ± 0.01	0.66 ± 0.01	0.47 ± 0.09	0.60 ± 0.05	0.71 ± 0.01	0.44 ± 0.08
LLE	0.49 ± 0.09	0.60 ± 0.04	0.62 ± 0.03	0.64 ± 0.01	0.14 ± 0.00	0.59 ± 0.04	0.65 ± 0.04	0.33 ± 0.10
Spectral Embedding	0.58 ± 0.09	0.59 ± 0.04	0.61 ± 0.02	0.64 ± 0.02	0.14 ± 0.00	0.64 ± 0.05	0.64 ± 0.02	0.23 ± 0.03

```python
combined_df_100 = pd.concat([no_embedding_df, mds_df_100, isomap_df_100, tsne_df, lle_df_100, spectral_df_100])
combined_df_100
```

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	1.00 ± 0.00	1.00 ± 0.00	0.92 ± 0.00	0.90 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
MDS	0.76 ± 0.02	0.63 ± 0.02	0.67 ± 0.01	0.70 ± 0.00	0.78 ± 0.00	0.76 ± 0.00	0.72 ± 0.01	0.75 ± 0.02
Isomap	0.88 ± 0.02	0.84 ± 0.01	0.84 ± 0.01	0.85 ± 0.01	0.88 ± 0.01	0.92 ± 0.01	0.87 ± 0.01	0.89 ± 0.01
t-SNE	0.44 ± 0.07	0.69 ± 0.01	0.68 ± 0.01	0.66 ± 0.01	0.47 ± 0.09	0.60 ± 0.05	0.71 ± 0.01	0.44 ± 0.08
LLE	0.72 ± 0.04	0.68 ± 0.02	0.68 ± 0.02	0.69 ± 0.03	0.34 ± 0.01	0.72 ± 0.04	0.72 ± 0.03	0.65 ± 0.04
Spectral Embedding	0.71 ± 0.01	0.62 ± 0.04	0.65 ± 0.02	0.68 ± 0.01	0.14 ± 0.00	0.69 ± 0.01	0.70 ± 0.02	0.32 ± 0.10

- T-test performance:

```
One-Sided T-Test Results for 100 Components:
MDS: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
Isomap: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
t-SNE: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
LLE: p-value = 0.0001
  Significant difference at p < 0.05, reject H0
Spectral Embedding: p-value = 0.0004
  Significant difference at p < 0.05, reject H0
```

```
One-Sided T-Test Results for 10 Components:
MDS: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
Isomap: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
t-SNE: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
LLE: p-value = 0.0001
  Significant difference at p < 0.05, reject H0
Spectral Embedding: p-value = 0.0001
  Significant difference at p < 0.05, reject H0
```

- Dog Breeds:
  - Accuracy Score:

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.58 ± 0.00	0.74 ± 0.00	0.49 ± 0.00	0.33 ± 0.00	0.70 ± 0.00	0.61 ± 0.00	0.71 ± 0.00	0.70 ± 0.00
MDS	0.22 ± 0.05	0.19 ± 0.09	0.07 ± 0.02	0.08 ± 0.07	0.04 ± 0.00	0.10 ± 0.02	0.18 ± 0.07	0.04 ± 0.00
Isomap	0.17 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.00 ± 0.00	0.04 ± 0.00	0.08 ± 0.00	0.14 ± 0.02	0.04 ± 0.00
t-SNE	0.25 ± 0.00	0.21 ± 0.00	0.08 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.08 ± 0.00	0.08 ± 0.04	0.04 ± 0.00
LLE	0.17 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.00 ± 0.00	0.04 ± 0.00	0.08 ± 0.00	0.10 ± 0.05	0.04 ± 0.00
Spectral Embedding	0.17 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.00 ± 0.00	0.04 ± 0.00	0.08 ± 0.00	0.09 ± 0.05	0.04 ± 0.00

	LDA	KNN (K=1)	KNN (K=3)	KNN (K=5)	SVM Linear	SVM RBF	Random Forest	Multinomial Logistic Regression
No Embedding	0.58 ± 0.00	0.74 ± 0.00	0.49 ± 0.00	0.33 ± 0.00	0.70 ± 0.00	0.61 ± 0.00	0.71 ± 0.00	0.70 ± 0.00
MDS	0.14 ± 0.05	0.25 ± 0.07	0.08 ± 0.03	0.12 ± 0.03	0.04 ± 0.00	0.08 ± 0.00	0.17 ± 0.07	0.04 ± 0.00
Isomap	0.25 ± 0.00	0.21 ± 0.00	0.08 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.08 ± 0.00	0.10 ± 0.05	0.04 ± 0.00
t-SNE	0.25 ± 0.00	0.21 ± 0.00	0.08 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.08 ± 0.00	0.08 ± 0.04	0.04 ± 0.00
LLE	0.25 ± 0.00	0.21 ± 0.00	0.08 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.08 ± 0.00	0.08 ± 0.03	0.04 ± 0.00
Spectral Embedding	0.25 ± 0.00	0.21 ± 0.00	0.08 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.08 ± 0.00	0.07 ± 0.03	0.04 ± 0.00

  - T-test performance:

```
One-Sided T-Test Results for 10 Components:
MDS: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
Isomap: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
t-SNE: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
LLE: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
Spectral Embedding: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
```

```
One-Sided T-Test Results for 100 Components:
MDS: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
Isomap: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
t-SNE: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
LLE: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
Spectral Embedding: p-value = 0.0000
  Significant difference at p < 0.05, reject H0
```

7. Discussion:

Interpretation of Results

The results of this study provide strong support for the hypothesis that Optimal Transport-based embeddings, particularly Wassmap embeddings utilizing the Wasserstein distance metric, enhance the performance of classification models. Notably, the Wassmap embeddings demonstrated a statistically significant improvement over traditional Euclidean metric embeddings across several datasets. This aligns with the initial hypothesis that leveraging the geometric properties of data distribution through Optimal Transport theory can yield better classification outcomes.

Link to Challenge and Hypotheses

Our findings are well-aligned with the theoretical benefits proposed by Optimal Transport theory. The theory suggests that capturing the intrinsic geometry of the data distribution more effectively should enhance learning algorithms' performance. The success of Wassmap embeddings in our experiments confirms this, contrasting with the traditional Euclidean embeddings where the t-test results frequently failed to reject the null hypothesis (H_0) indicating no substantial improvement over non-embedded models.

Implications

The superior performance of Optimal Transport-based embeddings suggests their potential for broader application in practical scenarios, particularly in fields requiring precise data modeling and high classification accuracy, such as medical imaging or facial recognition technologies.

These techniques may offer a significant advantage in handling complex datasets where traditional methods struggle to capture the underlying data structure.

Challenge and Problem Solving:

The process of converting large (224x224) images into point clouds for computing Wasserstein distances presented significant computational challenges, requiring several days to complete. Although the performance on these larger images did not match that observed with the MNIST datasets, it remained effective for statistical analysis via t-test. An attempt to expedite this process involved implementing cupy code to leverage GPU acceleration. However, this led to out-of-memory issues on our server due to limitations with CUDA. Addressing these challenges is crucial for future experiments, as resolving them will allow us to utilize larger sample sizes and potentially enhance the robustness and effectiveness of our results.

Future Work and Enhanced Statistical Validation:

- Completion of Wassmap Embeddings: The immediate next step is to finish applying Wassmap embeddings to the Coil-100 dataset.
- Multiple Trials on Classification Methods: To robustly test the effectiveness of various embedding techniques, we plan to conduct multiple trials across all embedding methods.
- Statistical Testing on Multiple Trials: Comprehensive statistical testing, including repeated trials, will provide a more reliable analysis of the embeddings' performance.
- Direct Comparison via T-Tests: Perform t-tests to compare Wassmap embeddings directly against Euclidean embeddings using data from these trials. This will help quantify the specific advantages of Wassmap embeddings over traditional methods.

8. Conclusion:

The visualizations and statistical tests conducted as part of this study underscore the effectiveness of manifold learning embeddings, particularly those based on Optimal Transport theory, in improving classification accuracy. The evident success of Wassmap embeddings over Euclidean metric embeddings highlights the potential of advanced manifold learning techniques to enhance the analytical capabilities of machine learning models. The results of one-sided t-tests confirm that these embedding methods statistically enhance classification accuracy, allowing us to reject the null hypothesis and affirm the performance improvement afforded by these techniques. Furthermore, Wassmap embeddings demonstrate superior clustering and consistency, evidenced by higher classification accuracies and more effective visualizations. These findings support the effectiveness of Optimal Transport-based techniques across similar datasets, underscoring their potential in advanced machine learning applications and supporting the continued exploration and adoption of these innovative methods in broader research and application contexts.

9. References:

Dr. Hamm's Wassmap function: https://colab.research.google.com/drive/1SumlVniLQ-qhOmBC473gSrhXGLue_14C

Datasets:

Coil-100: https://www.kaggle.com/datasets/jessicali9530/coil100

Dog Breeds: https://www.kaggle.com/datasets/mohamedchahed/dog-breeds

Fashion-MNIST: https://www.kaggle.com/datasets/zalando-research/fashionmnist

Handwritten-MNIST: https://www.kaggle.com/datasets/dillsunnyb11/digit-recognizer