Dimension Reduction for Classifying Medical Conditions using Ocular Data

Shane Lee
Dept. of Computer Science
Dept. of Mathematics
Yale University
shane.lee@yale.edu

Deborah Olorunisola
Dept. of Statistics and Data Science
Yale University
debbie.olorunisola@yale.edu

Linh Pham
Dept. of Statistics and Data Science
Yale University
linh.pham@yale.edu

***Abstract*—Eye doctors rely on visual appraisal to identify a host of ocular diseases. This paper seeks to use the Ocular Disease Recognition dataset from Kaggle in order to classify medical conditions using image data. By using SVD, PCA, and Laplacian Eigenmaps, we perform dimension reduction techniques to examine the usability of this dataset. From our analysis, we reveal that there are patterns and features that correctly classify ocular images for medical conditions, but the eye images alone are not sufficient for classification.**

I. Introduction

In hopes of creating comprehensive medical AI resources to assist medical professionals, Shanggong Medical Technology and Peking University collected over 10,000 images of eyes, along with identifying characteristics of the patient and their associated medical conditions [3]. The eye photographs are specifically of the fundus, which is the back inner wall of the eye. Specifically, the photographs are of the color fundus, which is a standard area used for diagnoses or any abnormalities. Because the eye can be a powerful non-invasive checkup, this project seeks to provide a deeper understanding of this dataset. This project applies three dimension reduction techniques in order to isolate patterns of interest. We then fine tune an image classification model to predict related medical conditions based on eye images.

The dataset includes the patient's age, sex, and the associated left and right eye images. Each eye is a 512 x 512 pixel image, but in order to ensure pairing, both eyes were merged into one 512 x 1024 image. Furthermore, trained human readers classified patients into eight labels according to their associated medical conditions: Normal, Diabetes, Glaucoma, Cataract, Age related Macular Degeneration, Hypertension, Pathological Myopia, and Other diseases/abnormalities. Each one of these columns is a binary representation; where 1 indicates the patient had this medical condition, and 0 means they did not. While most patients have only one medical condition, there were a few that had multiple. To make this prediction model simpler, patients with more than one condition were thrown out. The final count of patients left after preprocessing is 2,589.

The grayscale images were processed using the Python imaging library PIL (Pillow), flattened into long one-dimensional vectors, and assembled into a matrix where each row represented a patient.

II. Methods

Three dimension reduction techniques are applied to each 512 x 1024 image in order to examine patterns or trends: Singular Value Decomposition (SVD), Principal Component Analysis (PCA), and Laplacian Eigenmaps (LE) [4]. SVD helps create a lower dimension approximation of the data matrix. Including three channels of red, green and blue, each image is actually of size 512 x 1024 x 3. This lower dimension matrix helps isolate the most important variables in each image. In this case, SVD isolates which sections of the eyes are most relevant. Lastly, PCA utilizes the eigenvalues and eigenspaces in order to capture the variance of data's distribution. With PCA, this paper hopes to examine patterns in the Euclidean space as well. LE is another form of dimension reduction, but instead it relies on spectral techniques. LE assumes the data exists in a high dimensional Euclidean space, but samples from some lower dimension manifold. With LE, this project hopes to reveal any non-linear patterns.

III. Singular Value Decomposition

Singular Value Decomposition (SVD) was applied to this matrix using the randomized_svd function from the sklearn library, which efficiently approximates the top components in large datasets. For the top twenty singular values and vectors in particular, each patient's image was then projected onto these singular directions to generate twenty different coefficient distributions. These distributions were visually inspected for signs of clustering by medical condition. A linear classification model was then trained using the 100 SVD coefficients as input. To avoid overfitting, the SVD was first repeated on a portion of the initial 2,589 images dedicated as the training set, and the remaining images were left for the test set.

Figure 1 shows the magnitudes of the top 100 singular values.

There is a sharp drop-off after the first singular value: from 13571 to 1514. The 100th singular value is 77. Since our data set contains 2589 images, which bounds the rank of the data

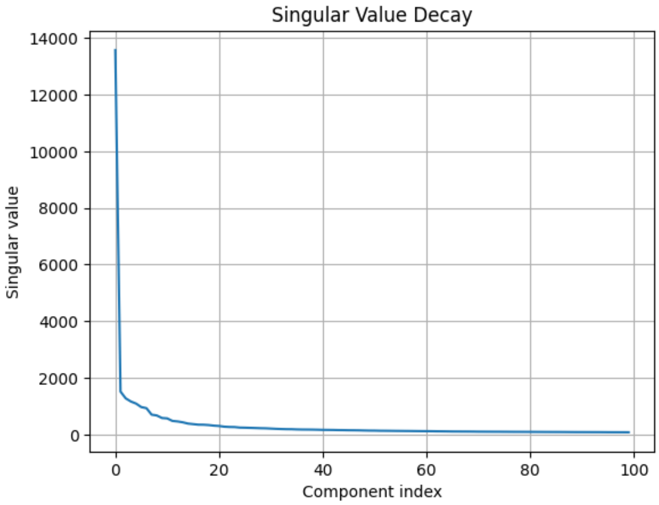

Fig. 1. Singular Values of SVD of Eye Images

matrix, we can place a lower bound on the explained variance ratio of the top 100 singular values. Note that

$$\begin{aligned}\sum_{i=1}^{r}\sigma_i^2 = \sum_{i=1}^{2589}\sigma_i^2 = \sum_{i=1}^{100}\sigma_i^2 + \sum_{i=101}^{2589}\sigma_i^2 \quad (1)\\ \leq \sum_{i=1}^{100}\sigma_i^2 + 2489\cdot 77^2 \quad (2)\end{aligned}$$

since $\sigma_i \leq \sigma_{100} \approx 77$ for all $i \geq 100$. Thus,

$$\frac{\sum_{i=1}^{100}\sigma_i^2}{\sum_{i=1}^{r}\sigma_i^2} \geq \frac{\sum_{i=1}^{100}\sigma_i^2}{\sum_{i=1}^{100}\sigma_i^2 + 2489\cdot 77^2} \quad (3)$$

The sum of squares of the first 100 singular values is 197099040, thus our lower bound for the explained variance ratio is

$$\frac{197099040}{197099040 + 2489\cdot 77^2} \approx 0.93. \quad (4)$$

This means that the top 100 singular values capture at least 93% of the total variance in the images data set, suggesting that it has a strong low-rank structure, allowing substantial reduction of dimensionality with minimal loss of information. Thus, we can proceed with our analysis of the SVD results.

Figure 2 shows the distributions of the first singular vector coefficient in the singular vector decomposition of each of the images across the varying conditions.

It appears that all of the groups have more or less the same distribution for the top 1 coefficient. Condition G (Glaucoma) appears to have a lower median than the other groups, though the entire distribution lies within the span of the other distributions so there are no signs of distinct clustering.

Unfortunately, the same conclusion can be made about the top 20 singular vector coefficients; all eight distributions overlap significantly, so differentiation by cluster is not possible.

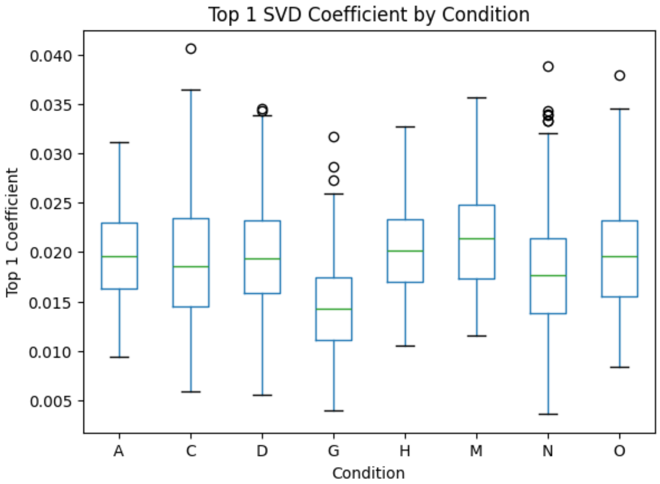

Fig. 2. 1st SV Coefficients Across Conditions

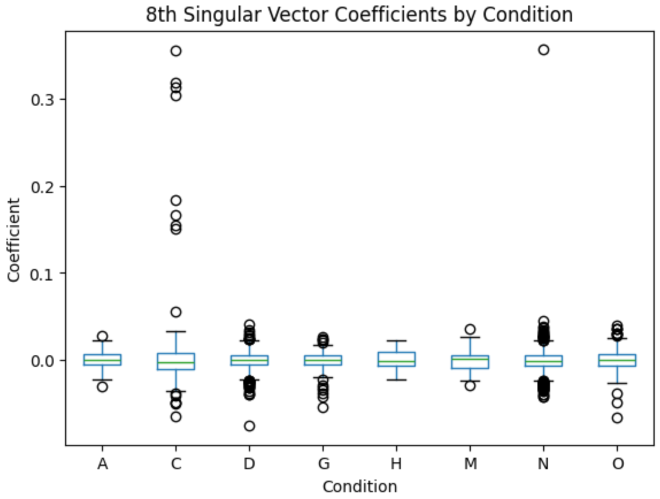

Fig. 3. 8th SV Coefficients Across Conditions

The coefficients for the 8th singular vector, shown in Figure 3, has an interesting pattern, however.

While the distributions do still overlap, the condition C (Cataract) has huge outliers compared to the other groups. Similar cataract outliers can be seen in some of the other singular vector coefficient distributions, though the 8th is the most extreme. This suggests that classification of the 7 conditions based on SVD seems difficult due to the lack of clustering, it may still be possible for cataract alone. Now that we are working with a binary problem, we use a logistic regression to model the presence of cataracts.

We first redo the SVD on a smaller training set of 2071 images, then perform the logistic regression model. Then, after extracting the singular vector coefficients of the test set of 518 images by projecting them onto the right singular vector matrix V, we run them through the model. The resulting ROC curve can be seen in Figure 4.

The area under the curve is 0.83, which suggests that the

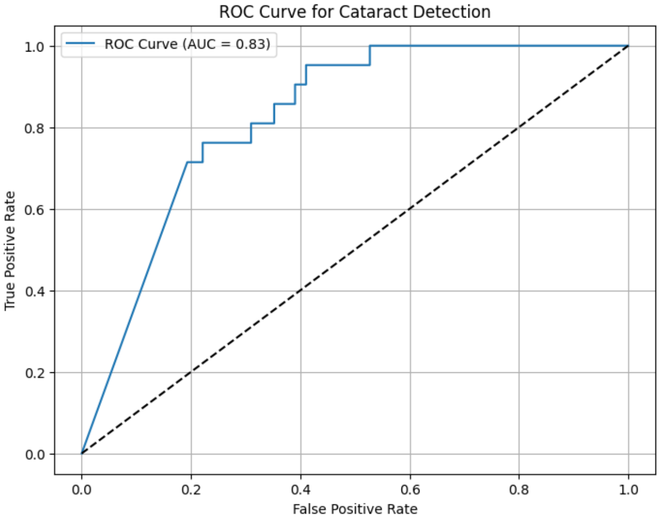

Fig. 4. ROC Curve of SVD-Based Logistic Classifier

model is good at differentiating cataract eyes from non-cataract eyes. The resulting confusion matrix at a threshold of $p = 0.5$ supports this conclusion as well (Figure 10).

The false positive rate is $\frac{207}{497} \approx 0.42$, and the false negative rate is $\frac{1}{21} \approx 0.05$. This model is quite accurate when the patient does have cataracts but not so much when they don't.

IV. Principal Components Analysis

In order to examine any linear patterns in the dataset, we applied a Principal Components Analysis (PCA) reduction technique shown in Figure 6. We can see that the dots generally appear to form three large strokes. Despite this clear structure, none of the conditions are distinguishable from one another. We then examined all combinations between two diseases to see if there anymore apparent clustering, but even that showed that the data generally grouped together. In Figure 5, we show the embeddings for just Glaucoma (G) and Age-related Macular Degeneration (A). The dots for A appear to exist a little above G but the overlap prevents us from drawing any real conclusions. All other combinations of medical diagnoses overlap in similar ways.

A. Multi-Layer Perceptron

We decided to train a Multi-Layer Perceptron (MLP) [2] to improve the embeddings and improve the clustering of the data. The MLP consisted of a two fully connected linear layers and a RELU function. The first linear layer took in the input dimension and had a hidden dimension size of 128. With several tests, this proved the best hidden dimension size. A RELU was applied to the result and another linear layer reduced the size to 8, which was our class size. Results are shown in Figure 7.

While Figure 7 does not show the data clustering by medical diagnosis, it does suggest there is some other metric which cleanly separates the ocular data. There are clearly two groups forming, and even a smaller third one between the two major groups. This suggests that medical diagnosis may not be a good differentiating indicator for fundus images.

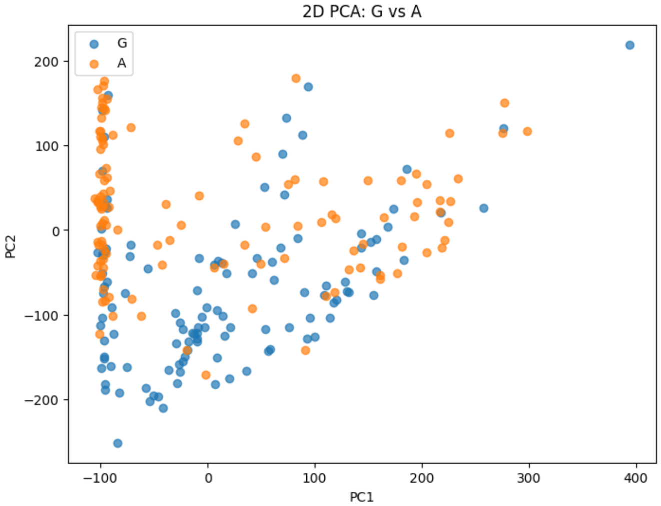

Fig. 5. Comparison between Glaucoma (G) and Age related Macular Degeneration (A)

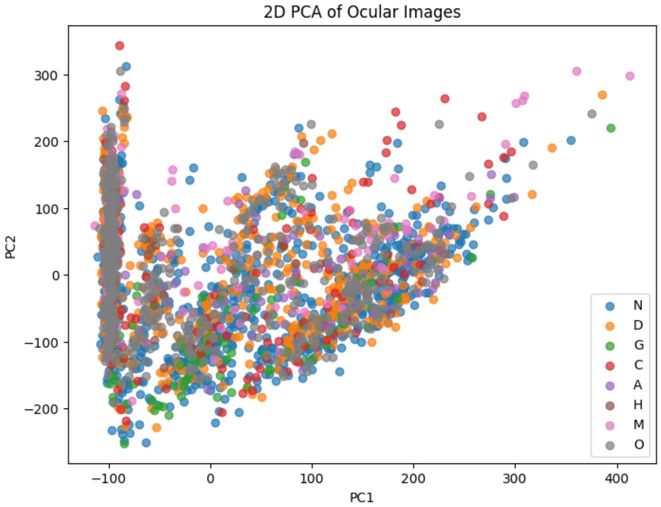

Fig. 6. Original PCA embeddings

V. Laclacian Eigenmaps

The implementation of laplacian eigenmaps here is based on the work of Belkin and Nigoyi [1]. Laplacian eigenmaps envision the vector of features (in this case, our image vectors) as nodes in a graph. Similar to PCA, we take advantage of the eigenvalues to understand where the most important variation is in the data, but instead of using the eigenvectors of the daa matrix, we use the eigenvectors of the graph Laplacian:

$$\mathbf{x}^\top L\mathbf{x} = \sum_{i,j}(x_i - x_j)^2 w_{ij}$$

This method is expected to be an improvement over SVD and PCA because it can account for nonlinear structure in the

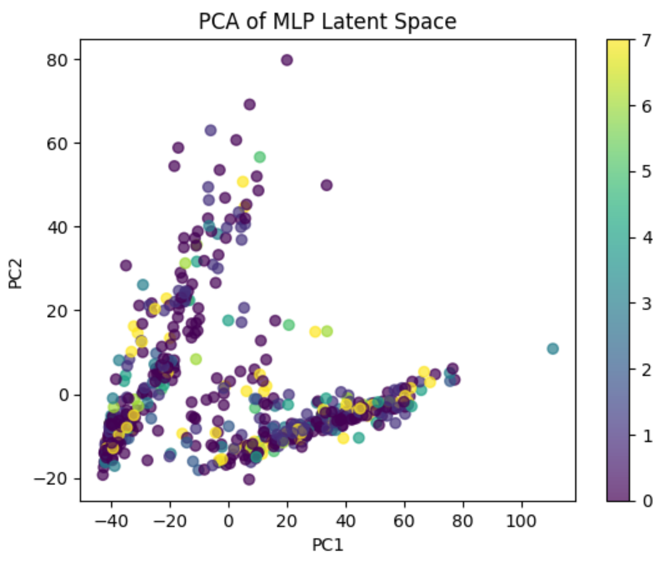

Fig. 7. After MLP Training

data. Thus, we hope that eyes with the same ocular disease are close to one another in n-dimensional space. For this paper, we look at the 2 and 3-dimensional solutions, which come from us minimizing the graph Laplacian on the black-and-white images.

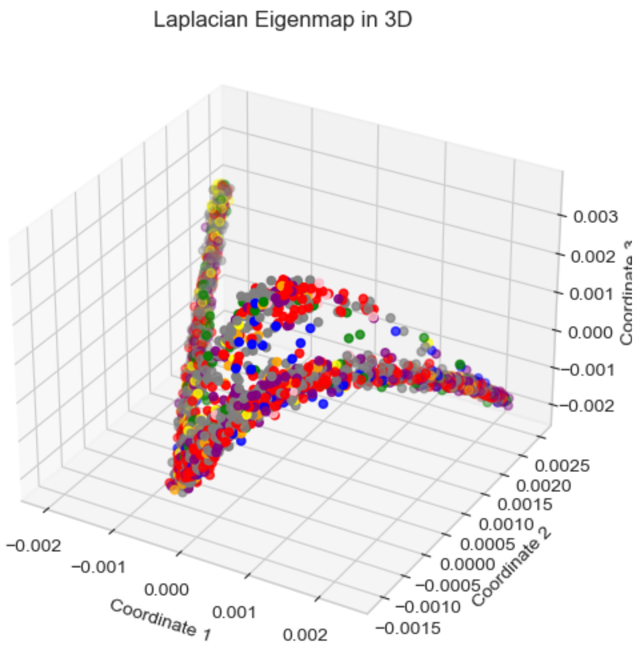

Fig. 8. 3D visualization of the Laplacian Eigenmap Embeddings

Given that the there are 8 categories for ocular disease, it is understandable that this method did not reliably separate the diseases in the linear transformation methods. There is likely information being projected into the null space; in other words, the data likely has some higher dimensional structure. Initially, we looked at the 3-dimensional case, shown in Figure 9. We see that there is a separate ”arm” that extends in the third dimension, which corroborates the 2D visualization of the the PCA. This arm is not visible in the 2D visualization of the LE embeddings. Instead, we see a parabolic shape, with all of the observations directly overlapping with one another. When we mapped the data to 8-dimensional space, then projected it back into 3D space with t-distributed stochastic neighbor embedding (t-SNE)[1], we see some separation of the “Other Diseases” images, but the other ocular conditions overlap otherwise. Referencing the singular is worth noting that the 77-dimensional LE embeddings (chosen after referencing the SVD drop-off) projected into 3D with t-SNE created an overdetermined system. There was neither a discernible pattern in the data overall nor in the pairs plots.

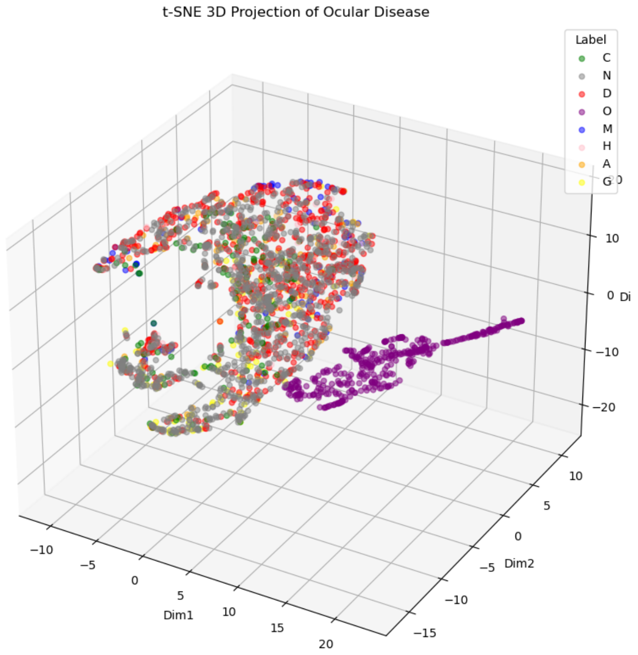

Fig. 9. t-SNE Visualization of 8D Laplacian Embeddings

Even with this limitation, we can see that some conditions are distinct from one another. Figure 11 in the Appendix shows some of these pairs. We see distinctness in myopia (M) versus glaucoma (G), myopia (M) versus age-related macular degeneration (A) and other diseases (O) versus myopia (M). Based on the visualization, a binary logistic regression based on these embeddings could be tuned to have high sensitivity (more false positive) for the lower instance conditions.

VI. Conclusion and Future Work

Laplacian eigenmaps did the best job of creating embeddings which can be used to distinguish different ocular diseases from one another. This is unsurprising, as there

[1] t-SNE approximates spectral clustering. It uses probabilities for a Gaussian kernel that determines (based on similarity) where the data points lie in n-dimensional space.

8 categories to distinguish from, so we would expect that anything less than 8 dimensions to be rank-deficient. Still, there are some conditions which are distinguishable from one another in pairs that may be useful if they are normally hard to visually distinguish.

There are many ways to improve this project. One major area of improvement would be to label the eyes *individually* by condition, rather than using the one-hot vector for two eyes. If one eye has a condition and the other exhibits normal features, then the image vector that the dimension reduction techniques learn from will associate the image with normal eyes, making clustering more difficult. Another improvement to this project would be using the color images for all clustering techniques. We were limited by our computing capacity, so we used greyscale images to conduct most of the analysis.

References

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[2] G.E.Hinton and R.R.Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):502–504, 2006.

[3] Larxel. Ocular disease recognition, 2020.

[4] Gilbert Strang. *Linear Algebra and Learning from Data.* Wellesley-Cambridge Press, Wellesley, MA, 2019.

Appendix

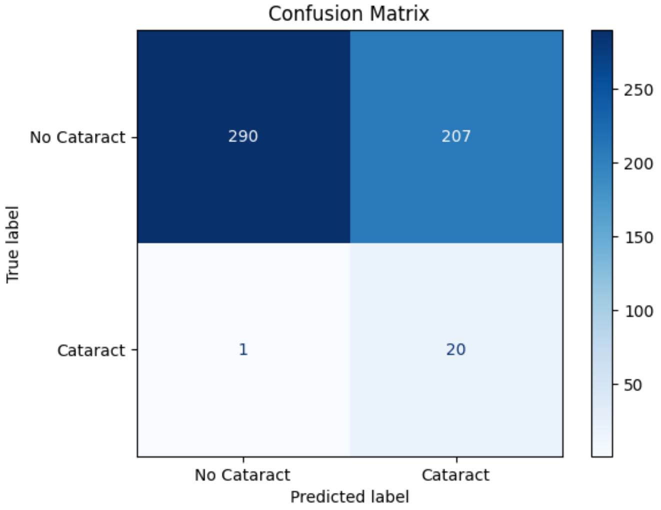

Fig. 10. Confusion Matrix of SVD-Based Logistic Classifier

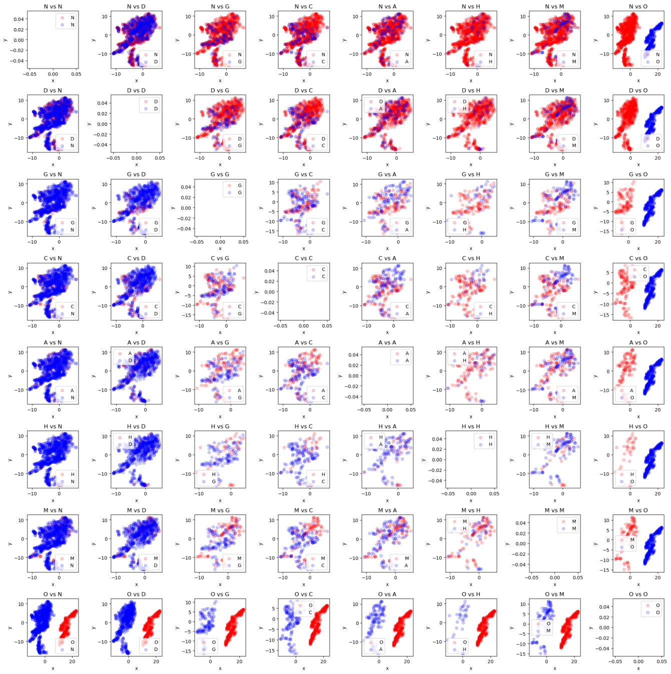

Fig. 11. Pairs Plot of the Laplacian Eigenmap Image Embeddings