An Artificial Intelligence Method for Phenotyping of OCT Scans Using Unsupervised and Self-supervised Deep Learning

Saber Kazeminasab[1], **Sayuri Sekimitsu**[2], **Mojtaba Fazli**[1,3], **Mohammad Eslami**[1], **Min Shi**[1], **Yu Tian**[1], **Yan Luo**[1], **Mengyu Wang**[1], **Tobias Elze**[1], **and Nazlee Zebardast**[4,✉]

[1]Harvard Ophthalmology AI lab, Schepens Eye Research Institute of Massachusetts Eye and Ear, Harvard Medical School, Harvard University, Boston, MA, USA
[2]Tufts University School of Medicine, Boston, MA, USA
[3]Novartis Institute for Biomedical Research (NIBR), Emeryville, San Francisco, CA, USA
[4]Department of Ophthalmology, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA

Artificial intelligence (AI) has been increasingly used to analyze optical coherence tomography (OCT) images to better understand physiology and genetic architecture of ophthalmic diseases. However, to date, research has been limited by the inability to transfer OCT phenotypes from one dataset to another. In this work, we propose a new AI method for phenotyping and clustering of OCT-derived retinal layer thicknesses using unsupervised and self-supervised methods in a large clinical dataset using glaucoma as a model disease and subsequently transfer our phenotypes to a large biobank. The model includes a deep learning model, manifold learning, and a Gaussian mixture model. We also propose a correlation analysis for the performance evaluation of our model based on Pearson correlation coefficients. Our model was able to identify clinically meaningful OCT phenotypes and successfully transfer phenotypes from one dataset to another. Overall, our results will contribute to stronger research methodologies for future research in OCT imaging biomarkers, augment testing of OCT phenotypes in multiple datasets, and ultimately improve our understanding of pathophysiology and genetic architecture of ocular diseases.

Image Phenotyping for OCT Images |Deep Unsupervised Learning | Correlation Analysis | Clustering | Self-supervised Learning in Feature Extraction | Dimensionality Reduction.

Correspondence: ***Nazlee Zebardast MD MS, Phone: (617) 573-3203; Email: Nazlee_Zebardast@meei.harvard.edu***

Introduction

Advances in artificial intelligence (AI), machine learning (ML), and deep learning (DL) have allowed for more precise analysis of medical imaging with the goals of better understanding disease pathogenesis and predicting disease burden. In particular, image phenotyping has enabled the extraction of complex patterns and subtle characteristics from medical imaging data, thereby unlocking novel insights into disease mechanisms and paving the way towards personalized diagnostic and therapeutic approaches. The majority of the medical images are unlabeled as the labeling task is expensive and time-consuming. Hence, methods like unsupervised learning or self-supervised learning are used for pattern recognition. Unsupervised learning employs machine learning algorithms to identify features of unlabeled datasets. These datasets are analyzed and clustered without the necessity of explicit labels or annotations. This allows the model to autonomously recognize and learn patterns within the input data. Through these methods, similarities and differences between data features can be discerned directly from the data itself. Techniques such as non-parametric instance discrimination (1), DeepCluster (2), unsupervised deep embedding (3), autoencoders (4), and deep adaptive image clustering (5) serve as examples of unsupervised learning for clustering tasks. Self-supervised learning extracts the feature space by designing a proxy task derived from the data itself, which is a noted limitation of this approach (6–10). This challenge, mainly the design of the proxy task in self-supervised learning, is predominantly addressed by contrastive learning (11). The objective in contrastive learning is to optimize a function for similar pairs in contrast to dissimilar pairs (12–21). The use of self-supervised learning on medical images is well studied in the literature (10, 22–26). Prior work has demonstrated the utility of image phenotyping in various ophthalmic applications such as glaucoma assessment (27), retinal disease diagnosis (28), genome-wide analysis study (GWAS) for vertical cup to disc ratio (VCDR) (29), as well as classification of age-related macular degeneration (AMD), (30) and geographic atrophy (31). Technological advances in AI and their applications to medical imaging have been fueled by the advent of large-scale biobank studies with multimodal data (e.g., the UK Biobank) (32). However, to date, image analysis has largely been limited by the lack of data annotation by experts (33) and limited transferability from one dataset to another. These limitations have consequences in settings where datasets lack the information needed for specific studies. Improving the ability to transfer phenotypes from dataset to dataset may allow for more studies tying imaging phenotypes to genetic studies; for example, making it possible to apply a model trained on a large clinical dataset without genetic information to a population-based biobank study with limited clinical data but a robust genetic repository. Optical coherence tomography (OCT) is a non-invasive imaging technique which segments retinal layers, is used in the screening, diagnosis, and management of multiple ophthalmic disease including glaucoma. Glaucoma, a leading cause of irreversible vision loss and blindness worldwide, is a complex multifactorial disease with high degrees of structural and functional variation. Primary open-angle glaucoma (POAG), the most common subtype of glaucoma, is highly heritable (34). However, to date, POAG pathophysiology and genetic architecture remains incompletely explained. Clinicians often struggle with predicting which patients may progress to more severe disease. This is likely due significant structural and functional variation in glaucoma. While OCT has emerged as a valuable tool in clinical and research contexts, contributing significantly to our understanding of glaucoma (35), current image analysis methods do not capture disease heterogeneity. We propose that unsupervised and semi-supervised learning algorithms can used to identify OCT-based endophenotypes that reflect subtypes of glaucomatous optic neuropathy. The purpose of this study therefore is to use glaucoma as a model disease to describe a new AI method for phenotyping and clustering of OCT-derived retinal layer thicknesses using unsupervised and self-supervised methods. We also describe a new method for transferring and testing OCT phenotypes from one dataset to another. Our results will enhance future research in OCT imaging biomarkers, augment testing of OCT phenotypes in multiple datasets, and ultimately improve our understanding of pathophysiology and possibly genetic architecture of ocular diseases.

Methods

Massachusetts Eye and Ear and the UK Biobank. We used data from Massachusetts Eye and Ear Infirmary (MEE), a large tertiary-care center in the United States, and the UK Biobank (UKBB), a prospective cohort study of UK residents. The MEE dataset is a large collection of macular OCTs and associated clinical data from a diverse patient population with glaucoma. Individuals with glaucoma were identified by the International Classification of Diseases, Ninth or Tenth Revision (ICD 9/10) diagnosis code for glaucoma (ICD9: 365.x, ICD10: H40.x). The UKBB includes detailed genotypic and phenotypic information on over 500,000 participants from across the United Kingdom, including ophthalmic testing and OCT imaging on a subset of participants (32).

In the MEE dataset, we used data from subset of patient with Cirrus OCT scans of the macula (Zeiss, Inc, Germany). Three-dimensional macular volume scans (128 or 200 B-scans, with a configuration of 512 or 200 horizontal A-scans in a 6x6-mm raster pattern) were obtained. Images and OCT thickness layers were reconstructed using Zeiss's automated boundary segmentations. This algorithm operates by analyzing the intensity and reflectance patterns of the incoming OCT scans. This algorithm uses complex image processing techniques, such as edge detection and pattern recognition, to identify and precisely delineate the boundaries between various retinal layers (36). In the UKBB, spectral-domain OCT scans of the macula were obtained on a subset of participants using Topcon 3D OCT 1000 Mk2 (Topcon, Inc, Japan). Three-dimensional macular volume scans were obtained (512 horizontal A-scans/B-scans; 128 B-scans in a 6x6-mm raster pattern). All OCT images were stored in .fda image files and the Topcon Advanced Boundary Segmentation (TABS) algorithm was used to automatically segment all scans. TABS uses dual-scale gradient information to allow for automated segmentation of the inner and outer retinal boundaries and retinal sublayers. For the MEE dataset, poor-quality images were identified. Missing pixels were imputed by the MissForest algorithm in Python (37). Also, the images with more than 5% missing values were excluded. The OCT scans in the UKBB have all pixels available.

A. Unsupervised Artificial Intelligence Model. Our AI model is comprised of a DL model, manifold learning for dimensionality reduction, Gaussian Mixture Model (GMM) for clustering and defining components coefficients for each image, and a visualization tool that samples representative images for each cluster (**Figure 1**). The input for the models included (1) RNFL thickness maps and (2) GCC thickness maps. Unsupervised deep learning (DL) models developed in this work include a deep autoencoder and an encoder trained using unsupervised representation method. All models were trained on the MEE dataset and tested on both the MEE and UKBB datasets.

A.1. Deep Learning Models. We use two deep learning model in our study:

- Unsupervised Autoencoder based Visual Feature Extraction
- Encoder based on Self-Supervised Representation Learning

Unsupervised Autoencoder based Visual Feature Extraction Since the scans do not have labels, we use unsupervised learning as one method for feature extraction. An autoencoder is a type of DL model used for It consists of an encoder and a decoder and works based on unsupervised learning. The encoder takes input data and maps it to a lower-dimensional latent space representation, while the decoder takes the latent space representation and reconstructs the original input. The autoencoder is trained by minimizing the difference between the input and the reconstructed output, using techniques such as backpropagation and stochastic gradient descent (38). **Figure 2** shows the schematic of the autoencoder we used in our work and **Supplemental Table S1** shows its parameters. The autoencoder model was implemented in the Keras framework.

Encoder based on Self-Supervised Representation Learning Another way of feature extraction of the dataset is self-supervised learning. We use momentum contrast (MOCO) (12) in which a feature model (e.g., an encoder) is trained based on a self-supervised learning method with contrastive learning as the proxy task. In this method, a dynamic dictionary for contrastive unsupervised learning is built using a queue and moving-averaged encoder. In MOCO, two random transformations of the same image are fed into two branches in which each branch includes an encoder and a projection

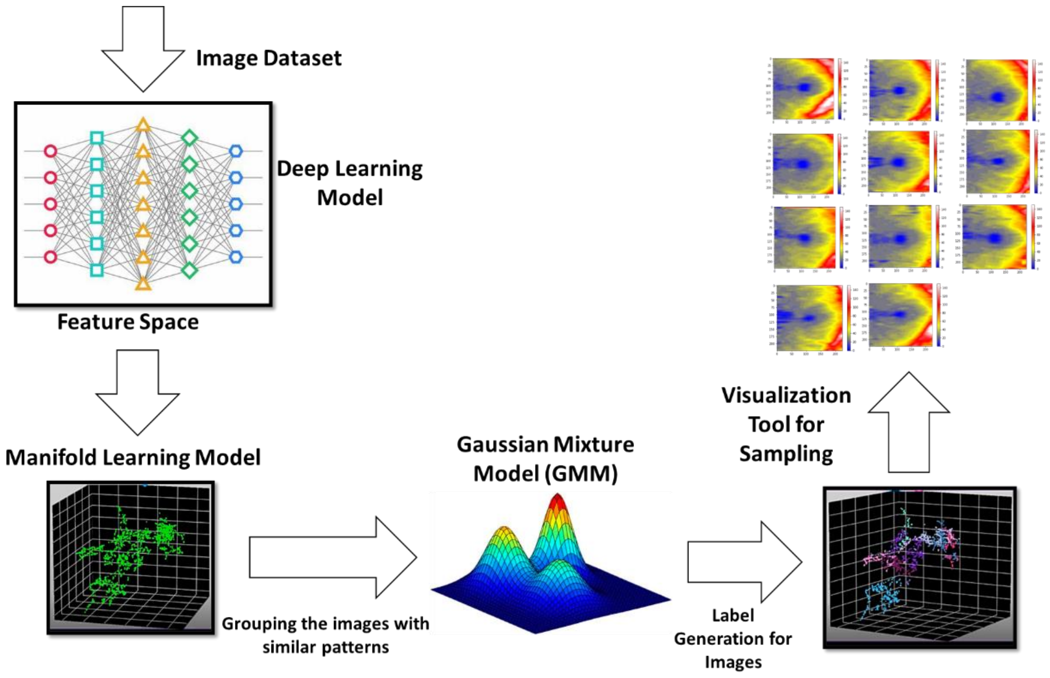


Image Dataset
Deep Learning Model
Feature Space
Manifold Learning Model
Grouping the images with similar patterns
Gaussian Mixture Model (GMM)
Label Generation for Images
Visualization Tool for Sampling

Fig. 1. The proposed artificial intelligence (AI) model for macular OCT retinal layer thickness map phenotyping.

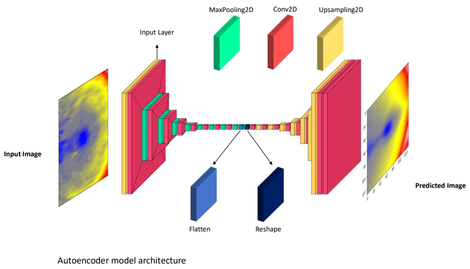


Fig. 2. The autoencoder model for feature extraction of retina layers with unsupervised learning method.

head that is a two-layer perceptron. The encoder in the lower branch is only optimized during training and the encoder in the upper branch is considered a moving average of the encoder in the lower branch.

For training the encoder, let's denote two transformations, x' and x'' from the input image x. The outputs of the upper and lower branches are denoted as $u^+ = \mathscr{M}(\mathscr{F}(x'; \theta'_{\mathscr{F}}); \theta'_{\mathscr{M}})$ and $u = \mathscr{M}(\mathscr{F}(x; \theta_{\mathscr{F}}); \theta_{\mathscr{M}})$, respectively. $\theta'_{\mathscr{F}}$ and $\theta'_{\mathscr{M}}$ are the parameters of the feature model and projection head of the upper branch whereas $\theta_{\mathscr{F}}$ and $\theta_{\mathscr{M}}$ are the parameters for feature model and projection head of the lower branch. The loss function for this model is as follows:

$$\mathscr{L} = -\log \frac{e^{u^T u^+/\tau}}{\sum_{i=1}^{b} e^{u^T u_i^-/\tau} + e^{u^T u^+/\tau}} \qquad (1)$$

Where b is the batch size, and u^- is the output of lower branch for any image other than the image x. τ is the softmax temperature and is set to 0.2. The parameters of the models are updated in each iteration as $\theta'_{\mathscr{F}} \leftarrow \mu\theta'_{\mathscr{F}} + (1-\mu)\theta_{\mathscr{F}}$ and $\theta'_{\mathscr{M}} \leftarrow \mu\theta'_{\mathscr{M}} + (1-\mu)\theta_{\mathscr{M}}$. μ is momentum coefficient and is set to 0.999. After training, the feature model and a projection head are accessible that transform the input image to a feature vector. S2 shows the parameters of the models and the hyperparameters for training based on the MOCO.

A.2. *Manifold Learning for Dimensionality Reduction.* Given that the features extracted from the previous DL models are high-dimensional, we used uniform manifold approximation and projection (UMAP) to project outputs from the previous model to a lower dimensional space (39, 40). The UMAP algorithm is a non-linear dimensionality reduction technique (41). It provides a unique representation of high-dimensional data, with underlying assumptions that the data is distributed uniformly on Riemannian manifold with a metric that is locally constant, and that the manifold is locally connected. Based on these assumptions, the manifold for the data is modeled with fuzzy topological structure in which the embedding (low dimensional representation) is constructed with the closest possible equivalent fuzzy topological structure (41). In our experiments, the latent space of 16,384 elements for the output of the autoencoder DL model and 128 elements for the encoder DL model were embedded into three elements for each data sample.

A.3. *Number of Clusters Determination.* As our dataset was unlabeled, we used the Bayesian information criterion (BIC) (42) and Akaike information criterion (AIC) for optimal selection of the number of phenotype clusters (43). BIC and AIC are commonly used statistical measures for model selection. Both BIC and AIC are based on the likelihood function of the model and include a penalty term to account for the number of parameters in the model. In general, models with lower AIC or BIC values are preferred over models with higher values, as they are considered to be more parsimonious. We calculated the AIC/BIC scores for the models with 2-20 clusters for each image type and DL model.

A.4. Gaussian Mixture Model (GMM) for Clustering. The Gaussian mixture model (GMM) is a statistical model that represents the probability distribution of a random variable as a weighted sum of Gaussian distributions (44). The probability density function (PDF) in one dimension for a Gaussian Distribution is expressed as:

$$G(X|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \textbf{(2)}$$

Where μ is the mean and σ^2 is the variance of the distribution. Analogous to one-dimensional distribution, the PDF for a multivariate Gaussian distribution is given as:

$$G(X|\mu,\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)} \quad \textbf{(3)}$$

In Eq. 3, μ is d-dimensional vector indicating the mean of the distribution and $\sum$ is a $d \times d$ covariance matrix. Assume we have K clusters, we compute the μ and $\sum$ for each cluster. If $K = 1$, they are estimated by the maximum likelihood method. However, for $K > 1$, we have:

$$p(X) = \sum_{k=1}^{K} \pi_k G(X|\mu_k, \Sigma_k) \quad \textbf{(4)}$$

π_k is the mixing coefficient for the k^{th} distribution. More details about this method can be found in (45). In this work, we fit the GMM model on one embedding space and either predict the same embedding space or a different one. Also, we calculate the probability for each image in a particular cluster. The cluster label is assigned based on the maximum probability of the image in a cluster.

Results

Study Population. In the MEE dataset, we identified 18,985 images of 8,323 patients with a diagnosis code for glaucoma. Of the 8,323 patients, 43.81% were men with an average age of 61.63 +/- 15.98 years. The race/ethnicity distribution of the study population is as follows: 57.47% White, 16.11% Black, 7.27% Asian, 0.16% American Indian or Alaska Native, 0.11% Native Hawaiian or Other Pacific Islander, and 13.81% Hispanic.

In the UKBB dataset, we identified 86,115 images of 47,908 participants. Of the 47,908 participants, 45.8% were male with an average age of 56.39 +/- 8.08 years. The ancestry distribution of the study population is as follows: 91.3% European ancestry, 3.95% African ancestry, 3.47% South Asian ancestry, 0.83% East Asian ancestry, and 0.43% American ancestry.

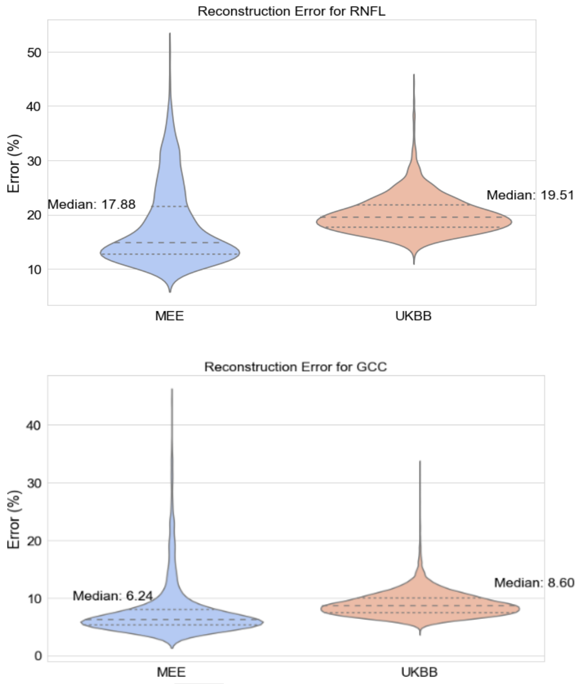


Fig. 3. Reconstruction error for the autoencoders that are trained on GCC and RNFL images of Mass Eye and Ear (MEE) dataset.

Autoencoder Model. The reconstruction error of trained autoencoders were computed (**Figure 3**), based on the comparison of each pixel of the original image and the reconstructed image. We calculated the error for each image and computed the mean error across all images. For MEE, the reconstruction error for RNFL and GCC was 18% and 6%, respectively. For UKBB, the reconstruction error for RNFL and GCC was 20% and 9%, respectively (**Figure 3**). The similar reconstruction errors for MEE and UKBB indicate that the autoencoder was able to successfully learn RNFL and GCC patterns identified in the MEE dataset and map them to the UKBB dataset.

We applied UMAP to the predictions of the autoencoder to reduce the dimensionality of the data samples. The parameters of the UMAP model are listed in **Supplemental Table S3**. AIC and BIC scores for embedding were calculated and three candidates for the optimal number of clusters are considered. The optimal number of clusters was found to be 11 for GCC and 9 for RNFL. This was selected based on a correlation analysis between images within and between clusters.

Different numbers of clusters were tested; the number of clusters with the highest correlation within images of a clusters compared to correlation of images between clusters were chosen (**Figure 4**). The applied GMM on the embeddings is shown in **Supplemental Figure S1**. Labels for the output of the GMM models were created. The labels are shown in the embedding space of MEE dataset for GCC and RNFL images in **Figure 5**. Representative sample images from each cluster are shown in **Figure 5.** The weight of the patterns for each dataset is shown in **Figure 6**. As seen in **Figure 6**, some patterns seen in MEE were not seen when the model was tested in the UKBB (RNFL – pattern 8, GCC – patterns 1, 5).

Encoder Model. As the encoder directly maps the images to feature space, reconstruction errors for the encoder model are not reported.

Similar to the autoencoder model, we applied UMAP on the predictions of the encoder and calculated AIC/BIC scores. The optimal number of clusters was found to be 11 for both GCC and RNFL based on AIC/BIC criteria and the number of clusters with highest correlation (**Figure** 7). The applied GMM on the embeddings is shown in **Supplemental Figure** S2. The labels and representative sample images from each cluster are shown in **Figure** 8. The weight of patterns in each dataset is shown in **Figure** 9. Similar to the results of the autoencoder model, some patterns seen in MEE were not observed when tested in the UKBB dataset (**Figure** 9). For GCC images, patterns 1, 2, 6, 8, 9, and 10 were not observed in the UKBB.

Unsupervised Performance Analysis. We used AIC/BIC scores to identify the optimal number of clusters in our data. Despite the utility of this strategy, we acknowledge its limitations, especially in the context of uncertainty related to the exact number of patterns and the lack of a defined ground truth for evaluating performance. To address these challenges, we introduce a new method for performance assessment that does not require supervision.

To this end, we employed a correlation analysis approach to examine the correlation within and between clusters. This was carried out for images in both pixel and feature space, allowing us to better understand the similarities and differences among images within a single cluster as well as between images in distinct clusters.

We first performed a correlation analysis between images in pixel space. To this aim, we used the normalized cross-correlation method in which the correlation is calculated based on **Equation** 5:

$$r_{fg} = \frac{\sum_{x,y}[f(x,y)-\bar{f}][g(x,y)-\bar{g}]}{\sqrt{\sum_{x,y}[f(x,y)-\bar{f}]^2 \sum_{x,y}[g(x,y)-\bar{g}]^2}} \quad \textbf{(5)}$$

where r_{fg}, $\bar{f}$ and $\bar{g}$ are the correlation, mean values of images f and g, respectively. however as this method was susceptible to noise, we did not find meaningful results (**Figure** 10). In **Figure** 10a, each cell represents the average correlation between all images of patter i: in rows, and pattern j: in columns for the RNFL images of MEE dataset for encoder model. As shown, there is no discernible pattern in pixel space correlation. Also **Figure** 10b shows an example of the failure of normalized-cross correlation for correlation analysis; two images that are visually more similar have less correlation (based on normalized-cross correlation) than two images that are less similar visually.

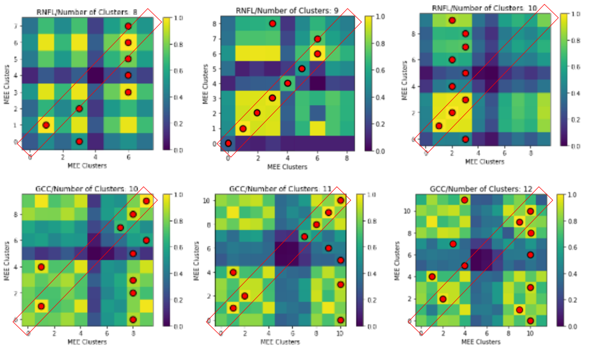


Fig. 4. The correlation analysis between the images within a cluster and the images in different clusters for different numbers of clusters for the autoencoder model.

We then performed the correlation analysis in the feature space. The correlation analysis in feature space is Pearson correlation (**Equation** 6):

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i-\bar{x})^2)(\sum_{i=1}^{n}(y_i-\bar{y})^2)}} \quad \textbf{(6)}$$

where:

- r is the correlation,
- x_i and y_i are the i-th elements of variables X and Y,
- $\bar{x}$ and $\bar{y}$ are the means of variables X and Y,
- n is the number of data points.

The results of the correlation analysis in feature space are shown in **Figures** 11 and 12 . The value for each cell is the average correlation between all images within a cluster and the images in different clusters. Most diagonal cells are brighter than other cells which means the images within a cluster have a higher correlation than the images in different clusters.

Clinical Associations. To assess the validity of our RNFL and GCC phenotypes, we identified associations between phenotypes and clinical characteristics within the MEE dataset. Correlations between phenotype coefficients and clinical characteristics (IOP, SE, VF MD, CDR, POAG PRS, number of glaucoma medications, and glaucoma severity) can be seen in **Figure** 13.

For the autoencoder model trained on RNFL thicknesses, phenotype 6 had the highest proportion of severe glaucoma diagnosis codes (ICD 10 diagnosis codes ending in 3 or 4; 49.18%), highest number of average glaucoma medications (1.57 +/- 0.12), highest mean CDR (0.694 +/- 0.237), highest proportion of IOP greater than 24 mmHg (9.99%) and lowest average VF MD (-12.57 +/- 10.1). Phenotype 4 had the highest proportion of highly myopic eyes (SE < -6, 10.73%). For the MOCO model trained on RNFL thickness, phenotype 7 had the highest proportion of severe glaucoma diagnosis codes (41.14%), highest number of average glaucoma medications (1.84 +/- 0.13), highest mean CDR (0.655 +/- 0.220), and lowest average VF MD (-8.278 +/- 7.28). Phenotype 10 had the highest proportion of highly myopic eyes (7.59%).

For the autoencoder model trained on GCC thicknesses, phenotype 4 had the highest proportion of severe glaucoma diagnosis codes (39.93%), highest mean CDR (0.654 +/- 0.229),

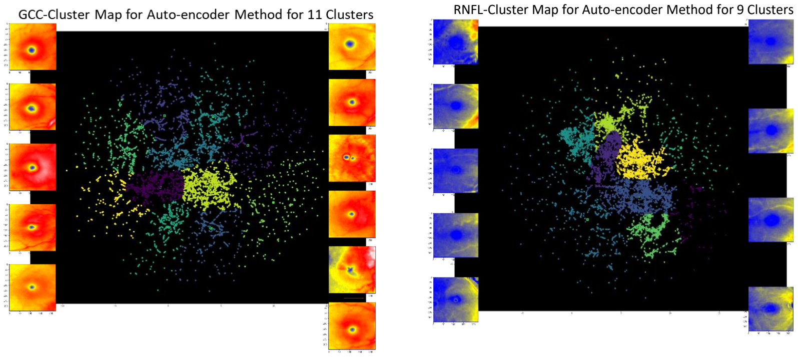


Fig. 5. Labels of data points in the embedding space of GCC and RNFL images of autoencoder model.

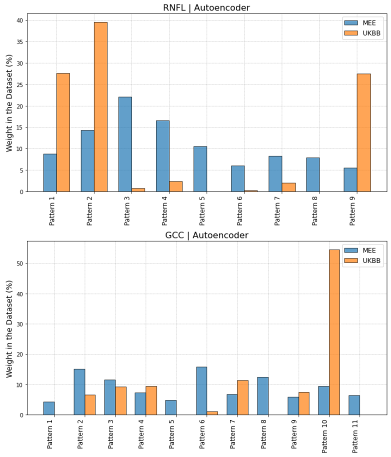


Fig. 6. The weight of patterns in the datasets for the autoencoder model

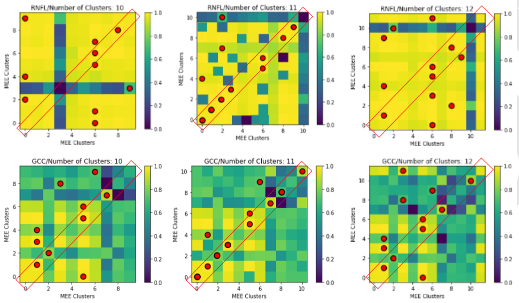


Fig. 7. Labels of the data point in the embedding space of GCC images of MOCO model.

and lowest average VF MD (-9.445 +/- 8.819), while phenotype 9 had the highest number of average glaucoma medications (1.87 +/- 0.11). Phenotype 1 had the highest proportion of eyes with high myopia (7.73%).

For the MOCO model trained on GCC thicknesses, phenotype 4 had the highest proportion of severe glaucoma diagnosis codes (48.42%), highest mean CDR (0.687 +/- 0.221), highest proportion of IOP greater than 24 mmHg (8.41%), lowest average VF MD (-10.95 +/- 9.72) and highest proportion of high myopia (6.85%). Phenotype 7 had the highest number of average glaucoma medications (1.38 +/- 0.11)

Discussion

Here we demonstrate that unsupervised deep learning can be used to identify imaging phenotypes which have meaningful clinical correlations. We employed a deep learning model, manifold learning, and GMM, and introduced a correlation analysis technique to evaluate the performance of unsupervised phenotyping. Importantly, our new method demonstrated the ability to transfer phenotypes derived from one dataset to another. This work reinforces the potential utility of deep learning models for better understanding of disease pathogenesis and prediction of disease progression.

We demonstrated that our model can be used to identify meaningful patterns of RNFL and GCC thickness amongst patients with glaucoma. In our correlation analysis, we showed that overall, images within one cluster are more correlated than images between different clusters. While pattern detection has been reported based on qualitative analysis in the literature (46) we propose a fully automated pattern recognition in glaucomatous images with a quantitative evaluation method. We also found that some images in clusters had higher correlation with images in other clusters; this may be because GMM assigns labels to each data point based on the highest probability. In some cases, the highest probability pattern may be fairly low (e.g., less than 50%), which means the image may be between multiple clusters. Additionally, GMM assumes that data is distributed in a Gaussian manner, which may not be applicable to real-world datasets such as ours, resulting in imperfect clustering performance. Furthermore, GMM may not perform well with high-dimensional data; as dimensionality increases, the volume of the space increases rapidly such that the available data become sparse making the Gaussian assumption less applicable. Therefore, our model must balance the performance of the GMM and the loss of information due to dataset dimensionality compression.

We compared the use of an autoencoder model and an encoder model based on self-supervised learning; we found that the correlations of the encoder model were higher, in average, compared to those of the autoencoder. This suggests that, while both autoencoders and encoders have demonstrated significant promise in feature extraction tasks, encoders may have an advantage for identifying more meaningful and relevant features. This may be because an encoder

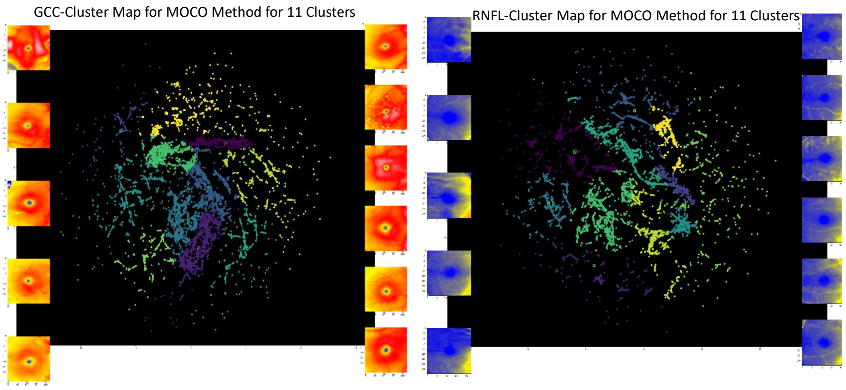


Fig. 8. Labels of the data point in the embedding space of GCC and RNFL images of MOCO model.

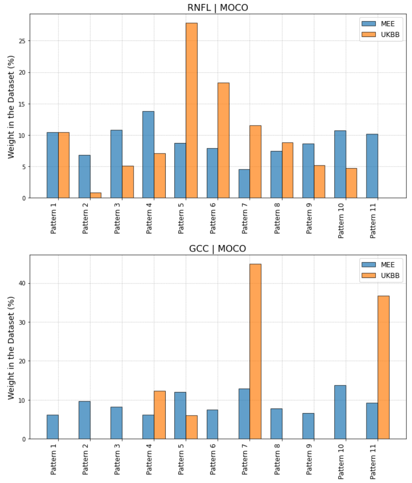
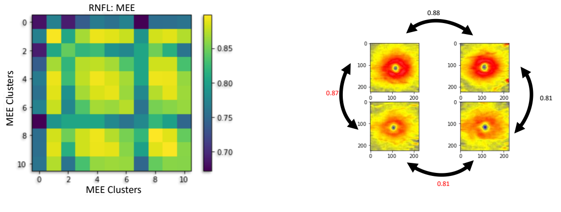


Fig. 9. The weight of patterns in the datasets for the encoder model.

focuses on mapping the high-dimensional input data into a lower-dimensional latent space and does not need to reconstruct the original data. This process may result in a more focused, information-rich feature extraction. The encoder concentrates exclusively on data compression, thus preserving more significant patterns in the data and discarding noise. Thus, in certain scenarios, a more simplified or targeted approach like an encoder could outperform more complex models such as autoencoders, further highlighting the importance of problem-specific model selection in machine learning applications.

Importantly, in this work, we demonstrate the ability to successfully transfer OCT phenotypes from one dataset to another with similar reconstruction errors (for autoencoder model) and good correlation of phenotypes (both autoencoder and encoder) in the feature space. Disparate datasets with limited genetic or clinical information pose significant challenges to cross disciplinary and translational research. The ability to transfer phenotypes from one dataset to another will allow for the ability to train and validate phenotypes in a dataset rich with clinical information and apply these phenotypes to a dataset with a genetic repository, thereby potentially increasing the utility of OCT phenotypes to discover novel disease associated genomic loci.



(a) **(b)**

Fig. 10. Correlation analysis in pixel space with normalized-cross correlation analysis for RNFL images of the UKBB dataset. Each cell in (a) shows the averaged correlation between the images in each cluster. (b) Example of failure of normalized cross correlation between images. The bottom row shows two similar images from the same pattern that are less correlated (i.e., 0.81) compared to the left column which are images from different patterns (i.e., 0.87). .

We also demonstrated that our phenotypes have meaningful clinical correlations, including variations in ocular parameters used to assess glaucoma severity such as CDR, visual field mean deviation, and IOP. Specifically, each model (autoencoder vs. MOCO) using different input retinal layers (RNFL vs. GCC) was able to identify a phenotype that correlated with more severe disease, including a more severe ICD code, greater number of glaucoma medications, higher CDR, and lower VF mean deviation. Furthermore, in correlation analysis, several RNFL phenotype coefficients were correlated with SE, a measure for myopia. This may indicate the ability for our models to distinguish RNFL thickness patterns that are affected by myopia; myopia can be associated with "tilting" of the disc due to a longer axial length as is a clinical confounder when assessing patients for glaucoma.

Our study has many notable strengths, including introducing a fully automated pattern recognition using retinal layers obtained from OCT scans, evaluation metrics for unsupervised performance, end to end clinical relevance investigation using the patterns in OCT scans, and transfer capability from one dataset to another. In addition, our MEE dataset size is one of the largest in the U.S.However, retinal thickness layers extracted from OCT images are subject to significant noise, including speckle noise (caused by the interference of multiple scattered light waves), electronic noise (associated with the electronic components of the OCT system, including the detector and analog-to-digital converter), and background noise (due to ambient light and system optics). In order to best correct for noise, we filtered poor quality images and imputed missing pixels. Furthermore, by employing UMAP,

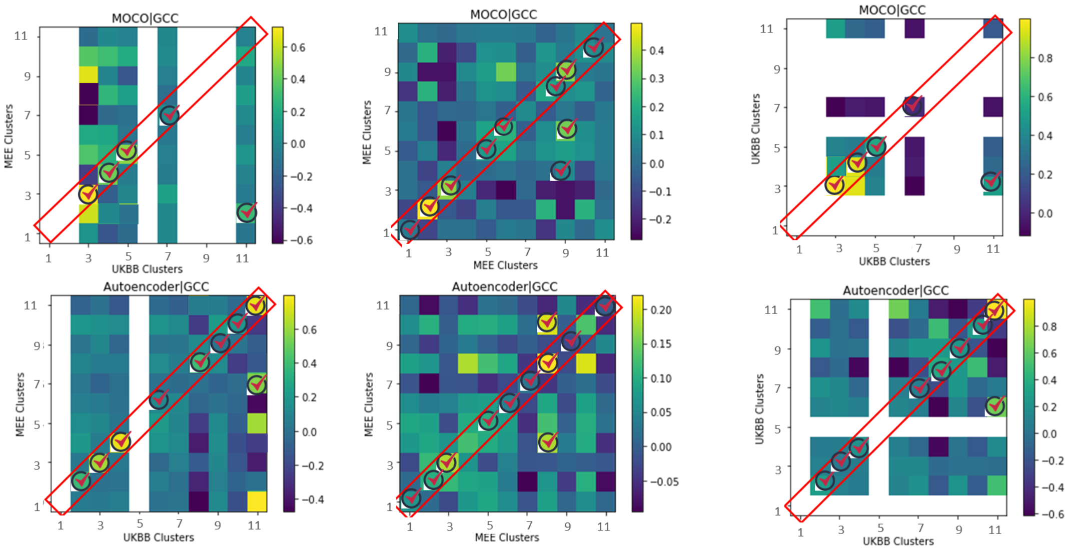


Fig. 11. Correlation analysis between the images in feature space with the Pearson correlation analysis for GCC images of MEE and UKBB dataset for both autoencoder and encoder methods.

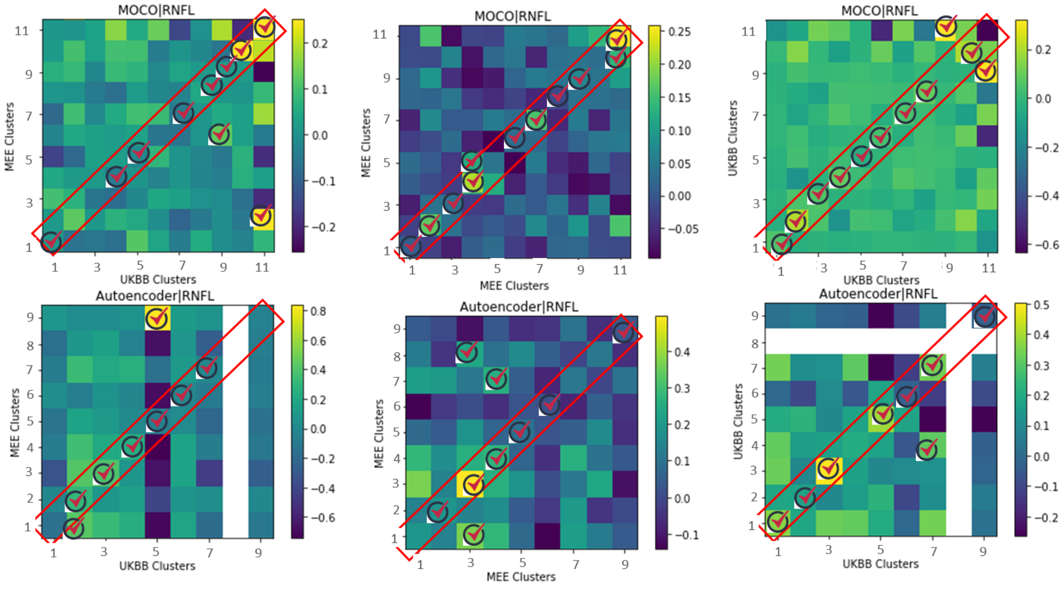


Fig. 12. Correlation analysis between the images in feature space with the Pearson correlation analysis for RNFL images of MEE and UKBB dataset for both autoencoder and encoder methods.

we were able to isolate the essential structure of the data and discard aspects of the data that are likely to be noise. UMAP maintains the topological structure of high-dimensional data in lower-dimensional representations, which ensures that the inherent relationships among data points are preserved while the random fluctuations, which often constitute noise, are minimized (47).

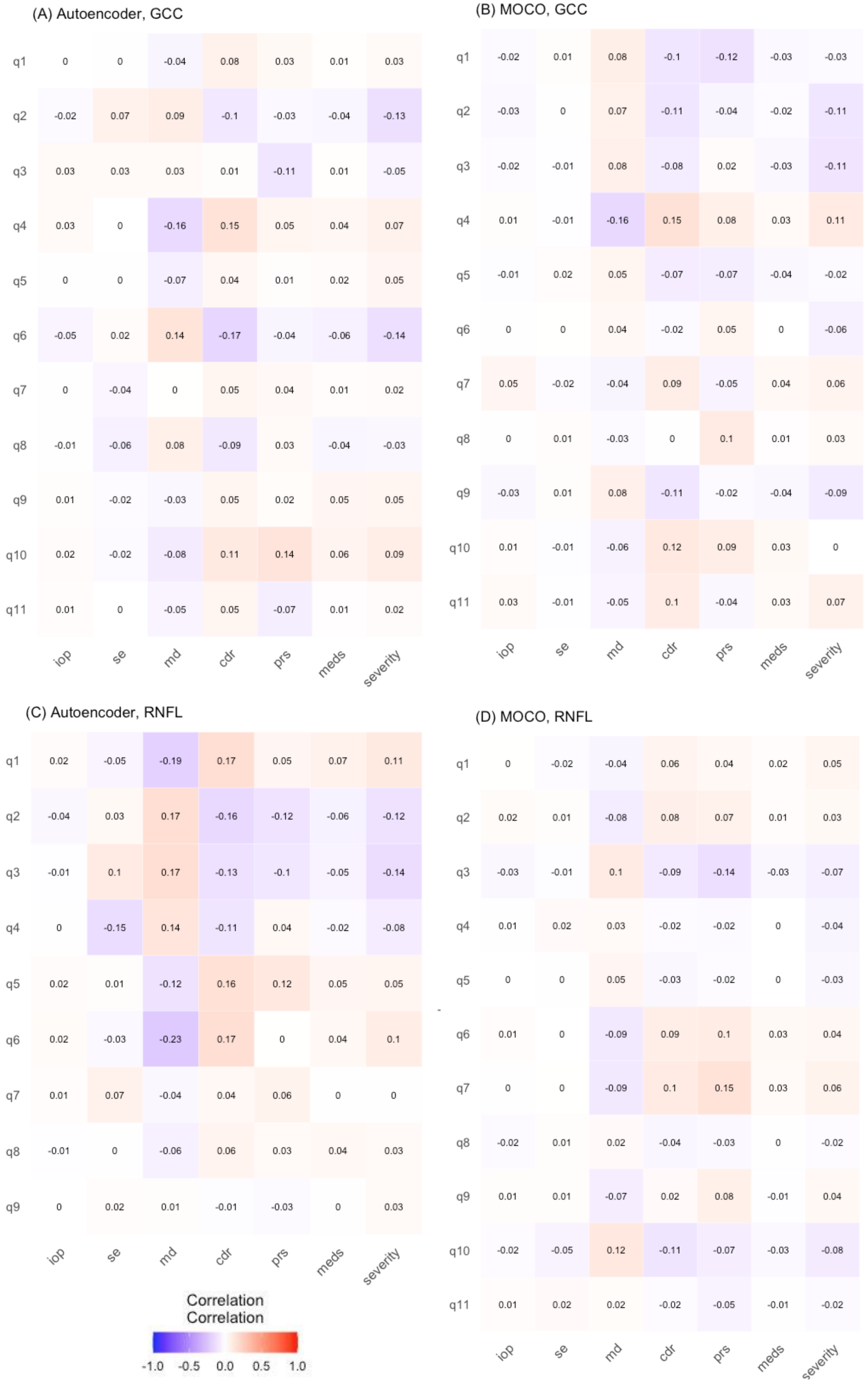
Our proposed methodology was effective for identifying clinically meaningful patterns in OCT images using RNFL and GCC thickness maps in patients with glaucoma as a model disease. We were also able to demonstrate the successful transfer of OCT phenotypes from one dataset to another, which may potentially increase the utility of these models for future translational and cross-disciplinary research including better understanding of glaucoma pathogenesis and genetic architecture. Finally, the introduction of a correlation-based assessment represents an exciting development for the performance validation of unsupervised learning methods.

ACKNOWLEDGEMENTS

This project was supported by the National Institute of Health (NIH) under grants 5K23EY032634-02 and R01EY030575.

Bibliography

1. Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.
2. Kai Tian, Shuigeng Zhou, and Jihong Guan. Deepcluster: A general clustering framework based on deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II 17*, pages 809–825. Springer, 2017.
3. Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.



(A) Autoencoder, GCC

(B) MOCO, GCC

(C) Autoencoder, RNFL

(D) MOCO, RNFL

Fig. 13. Correlation analysis for MEE phenotypes and clinical characteristics.

4. Fengfu Li, Hong Qiao, and Bo Zhang. Discriminatively boosted image clustering with fully convolutional auto-encoders. *Pattern Recognition*, 83:161–173, 2018.
5. Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017.
6. Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

7. Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.
8. Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.
9. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
10. Pengyue Zhang, Fusheng Wang, and Yefeng Zheng. Self supervised deep representation learning for fine-grained body part recognition. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 578–582. IEEE, 2017.
11. Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
12. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
13. Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.
14. R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
15. Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020.
16. Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 541–549. Springer, 2019.
17. Tobias Ross, David Zimmerer, Anant Vemuri, Fabian Isensee, Manuel Wiesenfarth, Sebastian Bodenstedt, Fabian Both, Philip Kessler, Martin Wagner, Beat Müller, et al. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International journal of computer assisted radiology and surgery*, 13:925–933, 2018.
18. Hannah Spitzer, Kai Kiwitz, Katrin Amunts, Stefan Harmeling, and Timo Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, pages 663–671. Springer, 2018.
19. Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P Harrison, Mohammadhadi Bagheri, and Ronald M Summers. Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9261–9270, 2018.
20. Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv preprint arXiv:1705.08260*, 2017.
21. Hongming Li and Yong Fan. Non-rigid image registration using self-supervised fully convolutional networks without training data. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1075–1078. IEEE, 2018.
22. Maximilian Blendowski, Hannes Nickisch, and Mattias P Heinrich. How to learn from unlabeled volume data: Self-supervised 3d context feature learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 649–657. Springer, 2019.
23. Jianbo Jiao, Richard Droste, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Self-supervised representation learning for ultrasound video. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 1847–1850. IEEE, 2020.
24. Nima Tajbakhsh, Yufei Hu, Junli Cao, Xingjian Yan, Yi Xiao, Yong Lu, Jianming Liang, Demetri Terzopoulos, and Xiaowei Ding. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 1251–1255. IEEE, 2019.
25. Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings*, pages 661–673. Springer, 2021.
26. Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 420–428. Springer, 2019.
27. Felipe A Medeiros, Alessandro A Jammal, and Atalie C Thompson. From machine to machine: an oct-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology*, 126(4):513–521, 2019.
28. Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, 2019.
29. Babak Alipanahi, Farhad Hormozdiari, Babak Behsaz, Justin Cosentino, Zachary R McCaw, Emanuel Schorsch, D Sculley, Elizabeth H Dorfman, Paul J Foster, Lily H Peng, et al. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *The American Journal of Human Genetics*, 108(7):1217–1230, 2021.
30. Yoshikatsu Hosoda, Masahiro Miyake, Kenji Yamashiro, Sotaro Ooto, Ayako Takahashi, Akio Oishi, Manabu Miyata, Akihito Uji, Yuki Muraoka, and Akitaka Tsujikawa. Deep phenotype unsupervised machine learning revealed the significance of pachychoroid features in etiology and visual prognosis of age-related macular degeneration. *Scientific Reports*, 10(1): 18423, 2020.
31. Yuki Mori, Masahiro Miyake, Yoshikatsu Hosoda, Akiko Miki, Ayako Takahashi, Yuki Muraoka, Manabu Miyata, Takehiro Sato, Hiroshi Tamura, Sotaro Ooto, et al. Genome-wide survival analysis for macular neovascularization development in central serous chorioretinopathy revealed shared genetic susceptibility with polypoidal choroidal vasculopathy. *Ophthalmology*, 129(9):1034–1042, 2022.
32. Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.
33. Katharina Grünberg, Oscar Jimenez-del Toro, Andras Jakab, Georg Langs, Tomàs Salas Fernandez, Marianne Winterstein, Marc-André Weber, and Markus Krenn. Annotating medical image data. *Cloud-Based Benchmarking of Medical Image Analysis*, pages 45–67, 2017.
34. Kanix Wang, Hallie Gaitsch, Hoifung Poon, Nancy J Cox, and Andrey Rzhetsky. Classification of common human diseases derived from shared genetic and environmental determinants. *Nature genetics*, 49(9):1319–1325, 2017.
35. Nabila Eladawi, Mohammed Elmogy, Mohammed Ghazal, et al. Classifcation of retinal diseases based on oct images. 2018.
36. Andrew Lang, Aaron Carass, Matthew Hauser, Elias S Sotirchos, Peter A Calabresi, Howard S Ying, and Jerry L Prince. Retinal layer segmentation of macular oct images using boundary classification. *Biomedical optics express*, 4(7):1133–1152, 2013.
37. Missforest. `https://pypi.org/project/MissForest/`. Accessed: October 20 2023.
38. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
39. Neil D Lawrence. A unifying probabilistic perspective for spectral dimensionality reduction: Insights and new models. 2012.
40. John A Lee, Michel Verleysen, et al. *Nonlinear dimensionality reduction*, volume 1. Springer, 2007.
41. Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
42. Harish S Bhat and Nitesh Kumar. On the derivation of the bayesian information criterion. *School of Natural Sciences, University of California*, 99, 2010.
43. Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.
44. Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.
45. Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
46. Sebastian M Waldstein, Philipp Seeböck, René Donner, Amir Sadeghipour, Hrvoje Bogunović, Aaron Osborne, and Ursula Schmidt-Erfurth. Unbiased identification of novel subclinical imaging biomarkers using unsupervised deep learning. *Scientific reports*, 10(1): 12954, 2020.
47. Valentin Koch, Olle Holmberg, Hannah Spitzer, Johannes Schiefelbein, Ben Asani, Michael Hafner, and Fabian J Theis. Noise transfer for unsupervised domain adaptation of retinal oct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 699–708. Springer, 2022.

Table S1. Hyperparameters for training the autoencoder model for feature extraction.

Hyper parameter	Value
Padding	Same
Activation	Relu
Optimization	Adams
Batch Size	256
Epochs	200
Input shape	128 × 128 ×1
Loss	Mean Square Error (MSE)

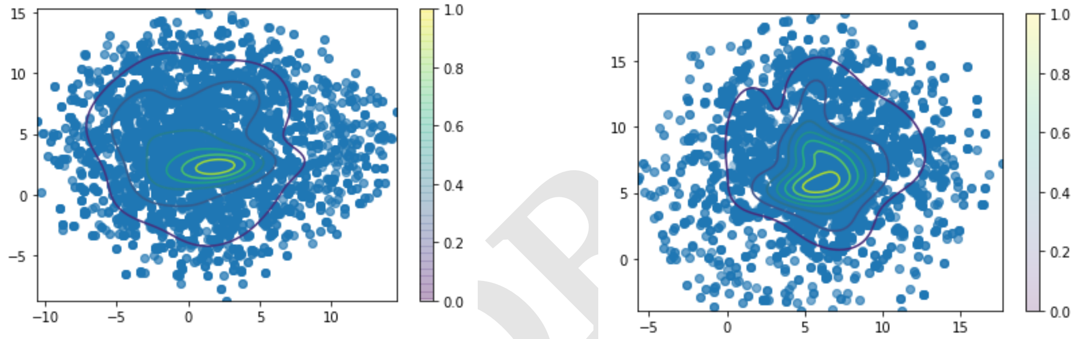
Supplementary Materials

Table S2. The parameters of Self-supervised learning based on momentum contrast.

Parameter	Description and Value
Backbone Encoder	ResNet50
Projection Head Output Dimension	128
Momentum Coefficient	0.999
Softmax Temperature	0.2
Bach size	256
Epoch	200
Augmentation	Random Crop

Table S3. UMAP Parameters

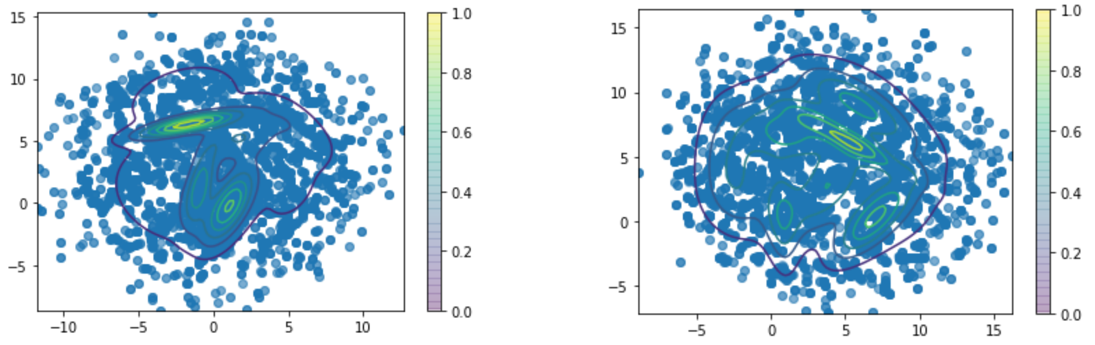
Parameter	Description or Value
Number of Reduced Elements	3
Number of Neighbors	5
Minimum Distance	0.0001
Random State	42



(a) GCC

(b) RNFL

Fig. S1. GMM Applied on the Embeddings of GCC and RNFL of MEE dataset for autoencoder model. The plots are in 2D.



(a) GCC

(b) RNFL

Fig. S2. GMM Applied on the Embeddings of GCC and RNFL of MEE dataset for encoder model. The plots are in 2D.