Title

Self-Supervised Feature Learning and Phenotyping for Assessing Age-Related Macular Degeneration Using Retinal Fundus Images

Permalink
https://escholarship.org/uc/item/2xx9n53v

Journal
Ophthalmology Retina, 6(2)

ISSN
2468-7219

Authors
Yellapragada, Baladitya
Hornauer, Sascha
Snyder, Kiersten
et al.

Publication Date
2022-02-01

DOI
10.1016/j.oret.2021.06.010

Peer reviewed

Published in final edited form as:
Ophthalmol Retina. 2022 February ; 6(2): 116–129. doi:10.1016/j.oret.2021.06.010.

Self-supervised feature learning and phenotyping for assessing age-related macular degeneration using retinal fundus images

Baladitya Yellapragada[1,2,3], **Sascha Hornauer**[2], **Kiersten Snyder**[3], **Stella Yu**[1,2], **Glenn Yiu**[3]

[1]Department of Vision Science, University of California, Berkeley, Berkeley, CA

[2]International Computer Science Institute, Berkeley, CA

[3]Department of Ophthalmology & Vision Science, University of California, Davis, Sacramento, CA

Abstract

Objective: Diseases such as age-related macular degeneration (AMD) are classified based on human rubrics that are prone to bias. Supervised neural networks trained using human-generated labels require labor-intensive annotations and are restricted to the specific trained tasks. Here, we trained a self-supervised deep learning network using unlabeled fundus images, enabling data-driven feature classification of AMD severity and discovery of ocular phenotypes.

Design: Development of a self-supervised training pipeline to enable grading of AMD severity using fundus photographs from the Age-Related Eye Disease Study (AREDS).

Subjects: 100,848 human-graded fundus images from 4,757 AREDS participants between 55-80 years of age.

Methods: We trained a deep neural network with self-supervised Non-Parametric Instance Discrimination (NPID) using AREDS fundus images without labels, then evaluated its performance in grading AMD severity using 2-step, 4-step, and 9-step classification schemes using a supervised classifier. We compared balanced and unbalanced accuracies of NPID against supervised-trained networks and ophthalmologists, explored network behavior using hierarchical learning of image subsets and spherical k-means clustering of feature vectors, then searched for ocular features that can be identified without labels.

Main Outcome Measures: Accuracy and kappa statistics

Results: NPID demonstrated versatility across different AMD classification schemes without re-training, and achieved balanced accuracies comparable to supervised-trained networks or human ophthalmologists in classifying advanced AMD (82% vs. 81-92% or 89%), referable AMD (87% vs. 90-92% or 96%), or on the 4-step AMD severity scale (65% vs. 63-75% or 67%), despite never directly using these labels during self-supervised feature learning. Drusen area drove network predictions on the 4-step scale, while depigmentation and geographic atrophy (GA) areas correlated with advanced AMD classes. Self-supervised learning revealed grader-mislabeled images and susceptibility of some classes within the more granular 9-step AMD scale to misclassification by both ophthalmologists and neural networks. Importantly, self-supervised

Corresponding Author: Glenn Yiu, 4860 Y St., Suite 2400, Sacramento, CA 95817, qyiu@ucdavis.edu.

learning enabled data-driven discovery of AMD features such as GA and other ocular phenotypes of the choroid (e.g. tessellated or blonde fundi), vitreous (e.g. asteroid hyalosis), and lens (e.g. nuclear cataracts) that were not pre-defined by human labels.

Conclusions: Self-supervised learning enables AMD severity grading comparable to ophthalmologists and supervised networks, reveals biases of human-defined AMD classification systems, and allows unbiased, data-driven discovery of AMD and non-AMD ocular phenotypes.

Keywords

Self-supervised Deep Learning; Deep Learning; Machine Learning; Artificial Intelligence; Age-related macular degeneration; AMD; AMD Classification; Feature Discovery

Introduction

Deep convolutional neural networks (CNNs) can be trained to perform visual tasks by learning patterns across hierarchically complex scales of representations [1] with earlier filters identifying low-level concepts such as color, edges, and curves, and later layers focused on higher-level features such as animals or animal parts [2]. Although CNNs are typically used for natural image tasks such as animal classification [3], aerial-view vehicle detection [4], and self-driving [5], these algorithms have also been adapted for medical image classification for clinical applications. In ophthalmology, deep learning algorithms can provide automated expert-level diagnostic tasks such as detection of diabetic retinopathy [6–11], age-related macular degeneration (AMD) [12–14], and glaucoma [15–17] using retinal fundus images. They can also extract information including age, sex, cardiovascular risk [18], and refractive error [19] that are not discernable by human experts.

However, supervised learning approaches are trained using expert-defined labels which classify disease type or severity into discrete classes based on human-derived rubrics that are prone to bias and may not accurately reflect the underlying disease pathophysiology. Because supervised networks can only identify phenotypes that are defined by human experts, they are also limited to identifying known image biomarkers. Moreover, training labels are labor intensive to generate, typically involving multiple expert graders who are susceptible to human error. Even trained ophthalmologists do not grade retinal images consistently, with significant variability in sensitivity for detecting retinal diseases [20].

Self-supervised and unsupervised learning organizes images based on features that are not predetermined by human graders. While unsupervised learning uses no labels and self-supervised learning generates labels for a proxy task, both methods are functionally similar in that neither require expert labels. These algorithms learn features from images without the constraints or arbitrary delineation of human labels during training, potentially enabling generalizability to classifying novel domains of data at the expense of reduced performance on known domains of data. Unsupervised and self-supervised neural networks have been developed using several methods, including instance-based learning [21], exemplar learning [22], deep clustering [23], and contrastive learning [24–27] As a contrastive learning approach, Non-Parametric Instance Discrimination (NPID) was previously designed for complex visual tasks [24]. NPID predicts a query image's class label by determining the most common label among its nearest neighbors within a multi-dimensional hypersphere of encoded feature vectors drawn from training images. This technique significantly outperforms other unsupervised networks for ImageNet, Places, and PASCAL Visual Object Classes classification tasks [24]. An updated version of NPID attempts to modulate the distance between the negative pairs based on presumed cross-level hierarchy of instances and groups [28].

In this study, we trained a self-supervised neural network through the NPID algorithm using unlabeled retinal fundus photographs from the Age-Related Eye Diseases Study (AREDS), then evaluated its ability to classify AMD across different human-derived severity scales using a supervised classifier. We then investigated the network's behavior to explore human label biases that may not conform to disease pathophysiology, and to enable unbiased discovery of ocular phenotypes. Because class boundaries are not explicitly established during the self-supervised training, (1) any set of labels can be used for evaluation without needing to train or retrain the classifier, and (2) visually-similar patterns outside of human-defined classes can be discovered.

We found that our CNN achieved similar accuracy to ophthalmologist graders [20] and supervised-trained CNNs, despite never learning the class definitions directly during training of the feature representations. Importantly, our examination of NPID behavior provides new insights into the visual features that drive test prediction, and enabled unbiased, data-driven discovery of AMD phenotypes not encompassed by human-assigned categories, as well as non-AMD features including camera artifacts, lens opacity, vitreous anomalies, and choroidal patterns. Our results show that self-supervised deep learning based on visual similarities rather than human-defined labels can bypass human bias and imprecision, enable accurate grading of disease severity comparable to supervised-trained neural networks or human experts, and discover novel pathologic or physiologic phenotypes that the algorithm was not specifically trained to detect.

MATERIALS & METHODS

Study data characteristics & partitioning

Sponsored by the National Eye Institute, the AREDS enrolled 4757 subjects aged 55 to 80 years in a prospective, randomized, placebo-controlled clinical trial to evaluate oral antioxidants as treatment for AMD. The AREDS design and results have been previously reported [29]. The study protocol was approved by a data and safety monitoring committee and by the institutional review board (IRB) for each participating center, adhered to the tenets of the Declaration of Helsinki, and was conducted prior to the advent of the Health Insurance Portability and Accountability Act (HIPAA).The AREDS sites received informed consent from subjects, which was not necessary for this this post-hoc analysis on the fundus data; digitized AREDS color fundus photographs and study data were obtained from the National Eye Institute's Online Database of Genotypes and Phenotypes website (dbGaP accession phs000001, v3.p1.c2) after approval for authorized access, and exemption by the IRB. The median age of participants was 68, 56% were women, and 96% were Caucasian [30,31]. Color fundus images from AREDS were previously graded by the University of Wisconsin fundus photograph reading center for anatomic features, including the size, area, and type of drusen, area of pigmentary abnormalities, area of geographic atrophy (GA), and presence of choroidal neovascularization (CNV) [30]. These gradings were used to develop a 9-step (more accurately a 9+3-step) AMD severity scale for each eye which predicts the 5-year progression risk to CNV or central GA [29], with steps 1-3 representing no AMD, 4-6 representing early AMD, 7-9 representing intermediate AMD, and 10-12 representing advanced AMD including central GA (step 10), CNV (step 11), or both (step 12) [29–32] (Supplemental Figure 1a). Both the 9+3-step scale and the simplified 4-step scale have been used to successfully train supervised CNNs to classify AREDS fundus images for AMD severity [13,20]. As NPID's feature space is more dependent on low-level visual variety to make its prediction space less susceptible to bias, performance is bolstered by not excluding any images, such as stereoscopic duplicates or repeated subject eyes from different visits. A total of 100,848 fundus images were available, with a long-tailed imbalance and overrepresentation of the no-AMD classes for both scales, and class 11 (CNV) in the 9+3-step scale (Supplemental Figure 1b–1c). Images were randomly partitioned into training, validation, and testing datasets in a 70:15:15 ratio, respectively, while ensuring that fundus images from the same subject did not appear across different datasets.

Data Preprocessing

Fundus images were down-sampled to 224x224 pixels along the short edge while maintaining the aspect ratio as similarly done in past literature [13]. Fundus images were also preprocessed with a Laplacian filter applied in each of the red-green-blue (RGB) color dimensions to better emulate the properties of more natural images of everyday scenes and objects (Supplemental Figure 2). Laplacian filtering is the difference of two Gaussian-filtered versions of the original image. In this study, it is the original fundus image (effectively, a Gaussian-filtered image with no blur) subtracted by the image Gaussian-filtered with a standard deviation (SD) of 9 pixels in each of the RGB color channels. Fundus photographs exhibit approximately the 1/f power distribution of natural images of everyday scenes and objects [33,34] but with more low-frequency than high-frequency information (Supplemental Figure 2a). The Laplacian-filtered fundus images more closely resembles that of natural statistics (Supplemental Figure 2b).

Network Pretraining

A CNN can transfer knowledge from one image dataset to another by using the same or similar filters [2]. Unlike natural images that contain a variety of shapes and colors that are spatially distributed throughout the image, fundus photographs are limited by shared fundus features such as the optic disc and retinal vessels, as well as the restricted colors of the retina and retinal lesions. This in turn limits the variability of the filters learned by the network. Thus, to transfer learning from a higher variety of discriminable features, we pretrained the network using the large visual database ImageNet (i.e., initialize the neurons across naturalistic filters), and then finetuned on the AREDS dataset without any weights frozen to further improve performance. A comparison of different sizes for the final layer feature vector for NPID, which depends on the complexity of the filters learned from the task, revealed an ideal size of 64 dimensions for our pretrained model to maximize the performance gained from transfer learning (Supplemental Figure 3).

NPID Training & Prediction

NPID discriminates unlabeled training images using instance-based classification of feature vectors in a spherical feature space. At its core, NPID uses a backbone network (ResNet-50) whose logit layer is replaced with a fully connected layer of a given size (64 dimensions), and an L2-normalization function on the output feature vectors. The vectors computed for the training images are stored and compared from the previous loop of the data to determine how to update the network. Details of the NPID Training and Pretraining Requirements are included in Appendix A, including performance without pretraining and the hyperparameters for the best NPID results.

Measurement of Network Performance

We evaluated trained network performance by measuring the overall testing accuracy on a novel group of images across both the 2-step classification and 4-step AMD severity scale. For classification, predictions are made through a weighted k-Nearest Neighbors (wkNN) voting function, a common performance evaluation scheme for self-supervised networks [35]. Though this method was traditionally used to benchmark the self-supervised pretraining that leads to better supervised fine-tuning, we instead adopted it as a supervised classification head that does not modify the underlying features learned. wkNN is more appropriate for evaluating NPID compared to other classification protocols for evaluating self-supervised networks because it requires no additional training (so there is no change in interpretation of the representations learned from NPID), its classification boundaries scale with the data, and the only hyperparameter we need to specify is the number of neighbors, k, to consider for voting. Further, NPID’s loss function is based on wkNN, so using this evaluation technique assesses the underlying representations learned from NPID directly.

For wkNN, we chose k=12, as it produced the highest balanced accuracies (Supplemental Figure 4). We chose the epoch that yielded the best balanced accuracy using wkNN classification voting scheme (see Appendix A). Then, we evaluated that epoch on a separate testing dataset using various metrics from the wkNN result including unbalanced accuracy, Cohen’s kappa, true positive rate, and false positive rate. Unbalanced accuracy is the average accuracy across all samples, whereas balanced accuracy is the average class accuracy [36,37]. While both accuracy metrics are relevant and positively highlight the performance of NPID, balanced accuracy is less biased to skewed class distributions by weighting underrepresented class scores as equally as overrepresented ones, and is more appropriate for comparing performance across different subsets of the same data as in our study. We also employed a second method to evaluate self-supervised features using Linear Support Vector Machines (Linear SVMs) [35].

Supervised Training & Prediction

To establish our own baseline, we perform supervised finetuning on ResNet-50 with the 9+3-step severity scale, after pretraining on ImageNet, using the same set of AREDS fundus photographs. The data augmentations and hyperparameters match that of our best implementation of NPID. To avoid retraining for each new scale, we mapped the logits from the 9+3-step scale to 4-step, 2-step advanced AMD, and 2-step referrable AMD classes to generalize coarse-grained performance. This baseline network is established to evaluate how our NPID-trained representations from fundus images without expert labels compare to those from a network supervised-trained with expert labels.

t-SNE visualization & Search Similarity

To assess neighborhoods of learned features, we evaluated search similarity and t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations. Search similarities show how a given query image's severity is predicted based on nearest neighbor references, and t-SNE visualizations show us how all the fundus images are distributed across neighborhoods of visual features chosen by the network. Specifically, t-SNE maps feature vectors from high-dimensional to low-dimensional coordinates while approximately preserving local topology. Here, we map the encoded 64D features onto 2D coordinates, wherein coordinates that are near each other in 2D are also near each other in the original feature space, meaning they are similarly encoded because they share visual features. Although t-SNE visualizations can distort some mapping from high dimensional to 2D feature spaces, our claims about NPID feature groupings were confirmed by visual review by a board-certified ophthalmologist (GY), and are thus based on the original images. The t-SNE visualization is used as a tool to discover these images faster for additional review. Thus, we can color each 2D coordinate by the known labels for each fundus image in the training set to observe which images are encoded near to each other and what visual groupings emerge from these locally similar encodings. This process is label agnostic, so evaluation across multiple domains of labels (e.g. 2-step AMD severity, 4-step AMD severity, drusen count, media opacity, etc.) is possible without retraining, unlike a supervised-trained network.

Hierarchical Learning

Because NPID appears more suitable for coarse-level than fine-level classification across dependent classes, we split up the 9+3-step dataset into each of the 4-step classes. We trained the NPID network on only no, early, intermediate, or advanced AMD images, then evaluated NPID's ability to discriminate between the three fine 9+3-step classes within each coarse 4-step class to identify which of the 9+3-step classes appear to show less visual discriminability than the grading rubric suggests.

Spherical K-Means Clustering

To locate specific, notable training images aside from exhaustive similarity searches of random query images, we employed spherical K-means to identify clusters of training images of interest. For conventional K-means clustering, the algorithm groups feature vectors into k distinct equally-sized gaussian-distributed groups based on the distances of the feature vectors to the approximated group centers [38,39]. Spherical K-means differs by calculating the distance along a sphere, instead of directly through Euclidean space, which is more suitable for NPID because it maps images into vectors on a sphere [40]. Mapping of K-means-defined labels onto the pre-existing t-SNE helps to identify regions that are notably defined by or distinct from the original labels for further analysis.

Results

Accuracy in grading AMD severity

We first evaluated NPID performance on a 2-step discrimination task for detecting advanced AMD (CNV and/or central GA), and found that wkNN applied to the self-supervised-learned features achieved an unbalanced accuracy (94%) that is comparable to the performance of our supervised-trained CNN (95.8%), a similar published supervised network (96.7%) or trained ophthalmologist (97.3%) [14]. The balanced accuracy, which is more applicable due to dataset imbalance, was also similar between the self-supervised-trained NPID (82%), our supervised-trained network (92%), the published supervised network (81%), and ophthalmologist (89%) (Figure 2a). Next, we compared the balanced accuracy of NPID with another supervised algorithm to distinguish “referable” AMD (intermediate or advanced) from no or early AMD, and found that our self-supervised-trained network performed only slightly worse (87%) than our supervised-trained network (90%), the published supervised network (92%), and ophthalmologist (96%) [20], despite never learning the class definitions directly (Figure 2b). For grading AMD severity using the 4-step scale, NPID achieved a 65% balanced accuracy, which was comparable to our supervised-trained network (75%), the published network (63%), and ophthalmologist (67%)(Figure 2c) [20]. In particular, the confusion matrix for NPID demonstrated superior performance for distinguishing early AMD (class 2) as compared to both the published supervised network and human expert (Figure 2d) [20].

When applied to a finer classification task, NPID only achieved a balanced accuracy of 25% on the 9+3-step scale, as compared to 40% using our supervised-trained network and 74% using the published supervised network [13] that utilized the same backbone network as our NPID approach. We achieved this balanced accuracy score using k=12 for wkNN, although we also tested k=5, 8, 23, and 50, and found that results were mostly consistent across different k-values (Supplemental Figure 4). Even though our most class-homogenous neighborhoods are defined by k=12 neighbors, they are still mostly coherent with k=50 neighbors, which was how NPID was originally evaluated on the ImageNet dataset [24]. With k=50, 28% shared the query image’s label while 68% were within 2 steps of the correct 9+3-step label (Figure 2e). Even for cases with incorrect 9+3-step class predictions, the 50 nearest neighbor images shared the query’s 4-step class label 56% of the time, which accounts for the higher accuracy of our network in the 4-step classification task. Thus, although self-supervised learning achieves lower supervised wkNN performance on the finer 9+3-step AMD severity scale compared to binary or 4-step AMD classifications, incorrect predictions deviate minimally from ground-truth labels. We confirmed our findings using linear SVM classifiers, which achieved a 26% balanced accuracy for 9+3-step classification consistent with the wkNN results.

Network behavior for grading AMD severity

To discern how the NPID network visually organizes images from different AMD classes, we employed t-SNE visualizations which mapped encoded 64-dimensional features onto 2-D coordinates. On the 4-step AMD severity scale, fundus images with no (blue), intermediate (yellow), and advanced (red) AMD formed distinct clusters, while early AMD (aqua / green) images are scattered throughout the plot (Figure 3a), which likely explains the lower performance in this class (Figure 2d). On the 9-step AMD severity scale (Figure 3b), the t-SNE plot appear similar to that of the 4-step scale, as each of the 4 major classes on the simplified scale are dominated by one or two of the finer classes within each subset (Supplemental Figure 1b), and may account for the poorer performance of our self-supervised-trained network on the 9+3-step task.

Examining the training images that contribute to NPID predictions helps explain the self-supervised-trained network's behavior in an interpretable way that supervised-trained networks cannot, as the specific training images that drive a supervised network's predictions cannot be easily recovered. In our study, comparison of query images with a selection of neighboring reference images demonstrates high phenotypic similarity across adjacent 9+3-step classes (Figure 3c), and explains class confusions that contribute to NPID performance loss on the finer-grained 9+3-step scale. Furthermore, hierarchical learning on image subsets of each of the 4-step simplified AMD classes showed that the early AMD subset (class 2 on the 4-step scale) exhibited the least fine-class separability across the 9+3-step scale, with many class 4 images that resembled no AMD and class 6 images that appear similar to intermediate AMD (Figure 3c, bottom rows), which helps explain the difficulty with distinguishing early AMD images by the NPID method, as well as by supervised-trained networks and human ophthalmologists (Figures 2d–2e).

To determine which AMD features contributed most to the self-supervised learning, we mapped AREDS reading center-designated labels including (1) drusen size, area, and type, (2) depigmentation or hyperpigmentation area, and (3) total or central GA area onto the t-SNE plots (Figure 4). We found that drusen area provided the most visually distinct clusters that matched the separation of the 4-step severity scale. GA area and depigmentation correlated well with advanced AMD classes as expected, while larger drusen size or soft drusen type corresponded to intermediate AMD classes. Our results show that t-SNE visualizations, similarity searches, and hierarchical learning based on NPID can unveil the susceptibility of more granular human-defined AMD severity schemes to misclassification by both ophthalmologists and neural networks, and provide insight into the anatomic features that may drive AMD severity predictions.

Data-driven AMD phenotype discovery

Current AMD severity scales suffer from human bias because they were developed in part to reflect clinical severity (i.e. impact on visual function) rather than disease pathophysiology. For example, only vision-threatening GA involving the central macula was ascribed as advanced AMD (class 10 on the 9+3-step scale), while non-central GA cases were scattered across other AMD classes. With the goal of extracting features of any GA, both central and non-central, we conducted hierarchical training on only the referable AMD subsets (classes 3 and 4 on the 4-step scale), which consist of the most prominent AMD features. We found that the intermediate and advanced AMD cases in this subset were mostly separable within the t-SNE-defined feature space, and that the intermediate AMD images that grouped with advanced AMD samples exhibited features of GA (Figures 5a–5c).

To more objectively delineate the feature pockets that define the GA phenotype, rather than human-defined demarcations between intermediate and advanced AMD classes, we performed spherical K-means clustering to segregate fundus image feature vectors into K clusters. Using K=6 to correspond to the 6 fine-grained classes within the referrable AMD subset (classes 7-12), we found three clusters (Clusters A, B, and C) among eyes with intermediate AMD that correspond to variable degrees of GA (Figures 5a–5b), including non-central GA (Figure 5c, top row), as well as cases with central GA that should have been labeled as class 10, but were possibly mislabeled by human graders (Figure 5c, bottom row).

Because the advanced AMD hierarchical subset predictably demonstrated the greatest separability between its three fine-grained 9+3-step classes (classes 10-12), we also performed spherical K-means clustering on this subset’s feature vectors using K=3 to correspond to these 3 classes (Figures 5d–5e). Here, we discovered one of the three clusters (cluster C) to contain 75% of class 10 (central GA) and class 12 (central GA + CNV) fundus images. Sampling from class 11 images within this unbiased cluster also revealed images with non-central GA, even though class 11 was agnostic to GA presence (Figures 5f). We did not locate class 11 images with obvious central GA. Thus, hierarchical learning and K-means clustering using NPID may enable unbiased, data-driven discovery of AMD phenotypes such as GA which are not specifically encoded by human-assigned AMD severity labels.

Non-AMD phenotype discovery

To identify other physiologic or pathologic phenotypes beyond AMD features, we performed K-means clustering on all training images using a K-value of 4, based on the presence of 4 coarse classes in the 4-step severity scale. We observed one cluster (Cluster A) which correspond to images with no AMD, and three other clusters (Clusters B, C, and D) which appear to straddle AMD classes, suggesting that these latter groups may be distinguished by features unrelated to AMD pathophysiology (Figures 6a–6b). A closer examination of cluster B images near the border between AMD and non-AMD classes revealed eyes with a prominent choroidal pattern known as a tessellated or tigroid fundus appearance (Figure 6c) – a feature associated with choroidal thinning and high myopia [41]. Cluster C images near this border contain fundi with a blonde appearance (Figure 6d), often found in patients with light-colored skin and eyes, or in patients with ocular or oculocutaneous albinism [42]. Images from cluster D in this area showed poorly-defined fundus appearances that were suspicious for media opacity (Figure 6e). To determine if this cluster may include eyes with greater degrees of lens opacity, we overlaid the main t-SNE plot with labels for nuclear sclerosis, cortical cataracts, or posterior sub-capsular opacity from corresponding slit lamp images obtained in AREDS, and found that eyes in cluster D corresponded to a higher degree of both nuclear and cortical cataracts (Supplemental Figure 5). Hence, fundus images contain other ophthalmologically-relevant information that are not constrained to the retina, and K-means clustering of retinal images can also identify eyes with tessellated or blonde fundi as well as visually-significant cataracts.

To explore other potential phenotypes not related to AMD, we also investigated images with no AMD that are grouped with those with more advanced stages of AMD (Figure 6f).

Among these images, we found examples of eyes with asteroid hyalosis – vitreous opacities consisting of calcium and lipid deposits, as well as camera artifacts such as lens flare or dirt (Figures 6g–6i), which may resemble AMD features to an nonexpert human or CNN that was not trained to identify these conditions. Our findings show that NPID-trained networks have the capacity for unbiased, data-driven discovery of both AMD features that were not encoded in the 4-step or 9+3-step human labels, as well as non-AMD phenotypes such as camera artifacts (lens flare or dirt), media opacity (nuclear cataracts or asteroid hyalosis), and choroidal appearance (tessellated or blonde fundus).

DISCUSSION

In this study, we successfully trained a self-supervised neural network using fundus photographs which could be used in combination with a simple classifier to predict AMD severity across different human-defined classification schema, reveal AMD features that drive network behavior, and identify novel pathologic and physiologic ocular phenotypes, all without the bias and constraints of human-assigned labels during the training process. NPID performance was comparable to a supervised-trained CNN using the same backbone network, previously-published supervised networks, and human experts in grading AMD severity on a 4-step scale (none, early, intermediate, and advanced AMD) [20], and in binary classification of advanced AMD (CNV or central GA) [14] and referable AMD (intermediate or advanced AMD) [20]. Our self-supervised-trained network also performed similarly to a supervised-trained network that was trained with both fundus images and genotype data on a custom 3-step classification of class 1, class 2-8, and class 9-12 on the 9+3-step severity scale (65% vs. 56-60) [43]. Our results suggest that even without human-generated labels during training, self-supervised learning without further feature refinement can be combined with a simple classifier to achieve predictive performance similar to expert human and supervised-trained neural networks.

Self-supervised learning using NPID has significant advantages over supervised learning. First, eliminating the need for labor-intensive annotation of training data vastly enhances scalability and removes human error or biases. Also, NPID predictions resemble ophthalmologists more closely than do supervised networks (Figure 2d). Like humans, the self-supervised-trained NPID network considers the AMD severity scale as a continuum and the relationship of adjacent classes. By contrast, supervised-trained algorithms generally assume independence across classes, are susceptible to noisy or mislabeled images, and may produce more egregious misclassifications. Because the NPID algorithm groups images by visual similarity rather than class labels, inaccurate predictions can be salvaged by other nearest neighbors during group voting.

Another advantage of NPID is its versatility across different labeling schemes (2-step, 4-step, or 9+3-step), whereas distinct supervised-trained networks must be trained or retrained for these different labeling splits, or for cross-study comparisons. Because self-supervised learning is label agnostic, the same network can be evaluated for different classification tasks, and its performance readily compared with other networks or human experts as we showed in our study. Our approach also benefits from the versatility to use other datasets, such as fundus images from the AREDS2 study, for external validation in future studies.

Also, NPID predictions using locally-defined wkNN voting are not dominated by overrepresented classes because local neighborhoods are populated with sufficient class homogeneity (especially with k=12 neighbors used here, compared to k=50). This is an advantage of self-supervised over supervised training approaches, as the latter trains neurons to drive overrepresented class predictions more than other classes. Although classification methodologies vary between different studies using AREDS fundus images, the training dataset used in our study exceeded the size of those used in training other supervised networks (70,349 [this study] vs 56,402 [14], 5,664 [20], or 28,135 [43]). These previous studies often excluded stereoscopic duplicates or same eye images from different visits to avoid associating non-relevant fundus features such as optic disc shape or vessel patterns with any given class. Because self-supervised learning is feature-driven, and not class-driven, our network could exploit the entire available AREDS dataset which improves its coherence across different features. For instance, the testing subset used in one prior study overrepresented eyes with intermediate (33%) and late AMD (33%), and underrepresented early AMD (3%), which deviates from the skewed distribution of these classes in the full dataset (50%, 20%, 15%, and 15% across the 4-step classes, respectively), and may account for the higher reported performance of some supervised networks which are susceptible to overrepresentation bias without appropriate measures to counterbalance such as class-aware sampling or weighted loss function [13,44,45]. A similar image subset selection with a higher frequency of late AMD images than the full dataset (33% vs. 6%) could also explain the higher unbalanced accuracy and kappa for another supervised network compared to NPID, despite a lower balanced accuracy, true positive rate, and false positive rate [14].

Our findings are remarkable because while the NPID-trained network was trained without labels, its performance was validated using human-assigned categories, analogous to testing students on a topic that was never taught to them. In contrast to supervised-trained networks that were originally trained with these labels, the self-supervised network had no a priori knowledge of the classification schema, many aspects of which are defined by humans using somewhat arbitrary rationale for taxonomy that may not reflect an actual distinction in disease pathophysiology, such as distinguishing central from non-central GA due to its impact on patients' visual function and quality of life. This likely explains why NPID performed better on binary or 4-step classification of AMD severity, which more likely presents true pathophysiologic distinction, than on the finer 9+3-step AMD severity scale, where subtle differences in phenotype such as drusen size or pigmentation are arbitrarily categorized into distinct, human-defined classes. For example, our t-SNE visualizations demonstrated a clear separation between no AMD and advanced AMD, but not between early-AMD classes (Figures 3a–3b), for which human grader performance is also the worst [20]. Also, the fine-grained class prediction result for each hierarchical learning setup trained on an individual 4-step class subset is consistent with the confusion matrices derived from training on the full dataset. One or two of the 9+3-step classes dominate each 4-step class due to the low visual variability among them. These findings support the need to reevaluate these finer class definitions using more unbiased, data-driven methodologies.

In our study, we probed the NPID network's behavior and found that AMD features such as drusen area drove predictions of AMD severity more than drusen size or type, or area of pigmentary changes. Using hierarchical learning and spherical K-means clustering, we also identified eyes with non-central GA among those with intermediate or advanced AMD based on proximity to eyes with central GA (class 10), even though this feature is not encoded in the human-labeled AMD severity scales. Our findings suggest that self-supervised learning can more objectively identify certain AMD phenotypes such as drusen area or GA presence which may better reflect disease pathophysiology, and enable the development of more unbiased, data-driven classification of AMD severity or subtypes that could better predict disease outcomes than human-assigned grades. Interestingly, K-means clustering also identified images with central GA that appeared mislabeled as intermediate AMD, further highlighting the ability of an self-supervised-trained network to discover miscategorized images in ways that label-driven supervised learning cannot.

Another interesting feature of self-supervised learning is the ability to identify non-retinal phenotypes from fundus images, including camera artifacts (lens dirt or flare), media opacity (cataracts or asteroid hyalosis), and choroidal patterns (tessellated or blonde fundus). While we identified these features by spherical K-means clustering using a K-value of 4, additional cluster resolution could unveil additional pathologic or physiologic phenotypes. Future studies using von-Mises mixture models for spherical K-means clustering, which do not assume identical cluster size, may enable smaller, localized clusters of phenotypic groupings to be identified. Thus, the application of NPID may not be limited to AMD grading, and its potential supersedes that of supervised-trained networks that are limited to the classification task for which it is trained.

Because fundus photographs exhibit very little visual and semantic variability compared to natural object images, we found that preprocessing by 2D Laplacian-filtering transformed the spatial frequency power spectra of fundus images to better resemble natural objective images (Supplemental Figure 2), and that pretraining with ImageNet before finetuning on AREDS increased network performance for the 9+3-step and 4-step tasks by 100% and 33% (Supplemental Figure 3), respectively, than compared to no ImageNet pretraining. This performance improvement implies that discriminating features relevant to the task were difficult to learn directly from the fundus photos, but improves with transfer learning, as seen on t-SNE comparisons with and without pretraining which demonstrate extra learned features that better correlate with intermediate AMD classes (data not shown).

In summary, we trained a self-supervised network with NPID using fundus photographs without human-generated class labels, and, using a supervised classification scheme, it produced balanced class accuracies for predicting AMD severity similar to human ophthalmologists and supervised-trained networks that require labor-intensive manual annotations and are susceptible to human error and biases. The NPID algorithm exhibits versatility across different labeling schemes without the need for retraining and is less susceptible to class imbalances, overrepresentation bias, and noisy or mislabeled images. Importantly, self-supervised learning provides unbiased, data-driven discovery of both AMD-related and other ocular phenotypes independent of human labels, which can provide insight into disease pathophysiology, and pave the way to more objective and robust classification schemes for complex, multifactorial eye diseases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

This study was supported by the CITRIS/Banatao Institute. G.Y. is supported by NIH K08 EY026101, NIH R01 EY032238, NIH R21 EY031108, BrightFocus Foundation, and Macula Society. S.Y. received research support from National Science Foundation, National Geospatial Agency, Defense Advanced Research Projects Agency, United States Geological Survey, United States Fish and Wildlife Service, and Lawrence Berkeley National Lab. S.H. received research support from the National Science Foundation and the the University of California, Berkeley. B.Y. received research support from the CITRIS/Banatao Institute, National Geospatial Agency, and the University of California, Berkeley. No funding organizations had any role in the design or conduct of this research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Conflicts of Interest:

G.Y. received research support from Clearside Biomedical, Genentech, and Iridex, and personal fees for consultancy from Alimera, Allergan, Carl Zeiss Meditec, Clearside Biomedical, Genentech, Intergalactic Therapeutics, Iridex, Regeneron, Topcon, and Verily. S.Y. received research support from Leidos, Etegent Technologies, Samsung, Glidewell Laboratories, Bosch Research, Amazon Research, Google Research, Facebook AI Research, and personal fees for consultancy from Intel Research, KLA Tencor, iCueMotion, and Cerebral Therapeutics.

References

1. Yann LeCun, Yoshua Bengio GH. Deep learning. Nature 2015.
2. Zhou B, Khosla A, Lapedriza A, et al. Object detectors emerge in deep scene CNNs. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.; 2015.
3. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 2015;115.
4. Yu H, Yang W, Xia GS, Liu G. A color-texture-structure descriptor for high-resolution satellite image classification. Remote Sens 2016;8.
5. Bojarski M, Yeres P, Choromanaska A, et al. Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv 2017.
6. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA - J Am Med Assoc 2016;316.
7. Gargeya R, Leng T. Automated Identification of Diabetic Retinopathy Using Deep Learning. Ophthalmology 2017;124.
8. Ting DSW, Cheung CYL, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA - J Am Med Assoc 2017;318.
9. Sayres R, Taly A, Rahimy E, et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. Ophthalmology 2019;126.
10. Abràmoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investig Ophthalmol Vis Sci 2016;57.
11. Krause J, Gulshan V, Rahimy E, et al. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. Ophthalmology 2018;125.
12. Burlina PM, Joshi N, Pekala M, et al. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmol 2017;135.
13. Grassmann F, Mengelkamp J, Brandl C, et al. A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography. Ophthalmology 2018;125.

14. Peng Y, Dharssi S, Chen Q, et al. DeepSeeNet: A Deep Learning Model for Automated Classification of Patient-based Age-related Macular Degeneration Severity from Color Fundus Photographs. Ophthalmology 2019;126.

15. Liu S, Graham SL, Schulz A, et al. A Deep Learning-Based Algorithm Identifies Glaucomatous Discs Using Monoscopic Fundus Photographs. Ophthalmol Glaucoma 2018;1.

16. Christopher M, Belghith A, Bowd C, et al. Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs. Sci Rep 2018;8.

17. Son J, Shin JY, Kim HD, et al. Development and Validation of Deep Learning Models for Screening Multiple Abnormal Findings in Retinal Fundus Images. Ophthalmology 2020;127.

18. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2.

19. Varadarajan AV, Poplin R, Blumer K, et al. Deep learning for predicting refractive error from retinal fundus images. Investig Ophthalmol Vis Sci 2018;59.

20. Burlina P, Pacheco KD, Joshi N, et al. Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis. Comput Biol Med 2017;82.

21. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. In: 35th International Conference on Machine Learning, ICML 2018. Vol 5.; 2018.

22. Dosovitskiy A, Fischer P, Springenberg JT, et al. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Trans Pattern Anal Mach Intell 2016;38.

23. Caron M, Bojanowski P, Joulin A, Douze M. Deep clustering for unsupervised learning of visual features. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).Vol 11218 LNCS.; 2018.

24. Wu Z, Xiong Y, Yu SX, Lin D. Unsupervised Feature Learning via Non-parametric Instance Discrimination. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.; 2018.

25. Van Den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv 2018.

26. He K, Fan H, Wu Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.; 2020.

27. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. arXiv 2020.

28. Wang X, Liu Z, Yu SX. Unsupervised feature learning by cross-level discrimination between instances and groups. arXiv 2020.

29. Davis MD, Gangnon RE, Lee LY, et al. The age-related eye disease study severity scale for age-related macular degeneration: AREDS report no. 17. Arch Ophthalmol 2005;123.

30. AREDS Research Group. The age-related eye disease study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: The age-related eye disease study report number 6. Am J Ophthalmol 2001;132:668–681. [PubMed: 11704028]

31. Ferris FL, Davis MD, Clemons TE, et al. A simplified severity scale for age-related macular degeneration: AREDS report no. 18. Arch Ophthalmol 2005;123.

32. AREDS Research Group. The age-related eye disease study (AREDS) system for classifying cataracts from photographs: AREDS report no. 4**Members of the Age-Related Eye Disease Study Research Group are listed at the end of the article. Am J Ophthalmol 2001;131.

33. Ruderman DL. The statistics of natural images. Netw Comput Neural Syst 1994;5.

34. Torralba A, Oliva A. Statistics of natural image categories. Netw Comput Neural Syst 2003;14.

35. Goyal P, Mahajan D, Gupta A, Misra I. Scaling and benchmarking self-supervised visual representation learning. In: Proceedings of the IEEE International Conference on Computer Vision. Vol 2019-October.; 2019.

36. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: Proceedings - International Conference on Pattern Recognition.; 2010.

37. Kelleher JD, Namee B Mac, D'Arcy A. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. 2015.

38. MacQueen J Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability.Vol 1.; 1967.

39. Lloyd SP. Least Squares Quantization in PCM. IEEE Trans Inf Theory 1982;28.

40. Dhillon IS, Modha DS. Concept decompositions for large sparse text data using clustering. Mach Learn 2001;42.

41. Zhou Y, Song M, Zhou M, et al. Choroidal and Retinal Thickness of Highly Myopic Eyes with Early Stage of Myopic Chorioretinopathy: Tessellation. J Ophthalmol 2018;2018.

42. Federico JR, Krishnamurthy K. Albinism. StatPearls Publishing; 2021.

43. Yan Q, Weeks DE, Xin H, et al. Deep-learning-based prediction of late age-related macular degeneration progression. Nat Mach Intell 2020;2.

44. Shen L, Lin Z, Huang Q. Relay backpropagation for effective learning of deep convolutional neural networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol 9911 LNCS.; 2016.

45. Zhang R, Isola P, Efros AA. Colorful image colorization. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).Vol 9907 LNCS.; 2016.

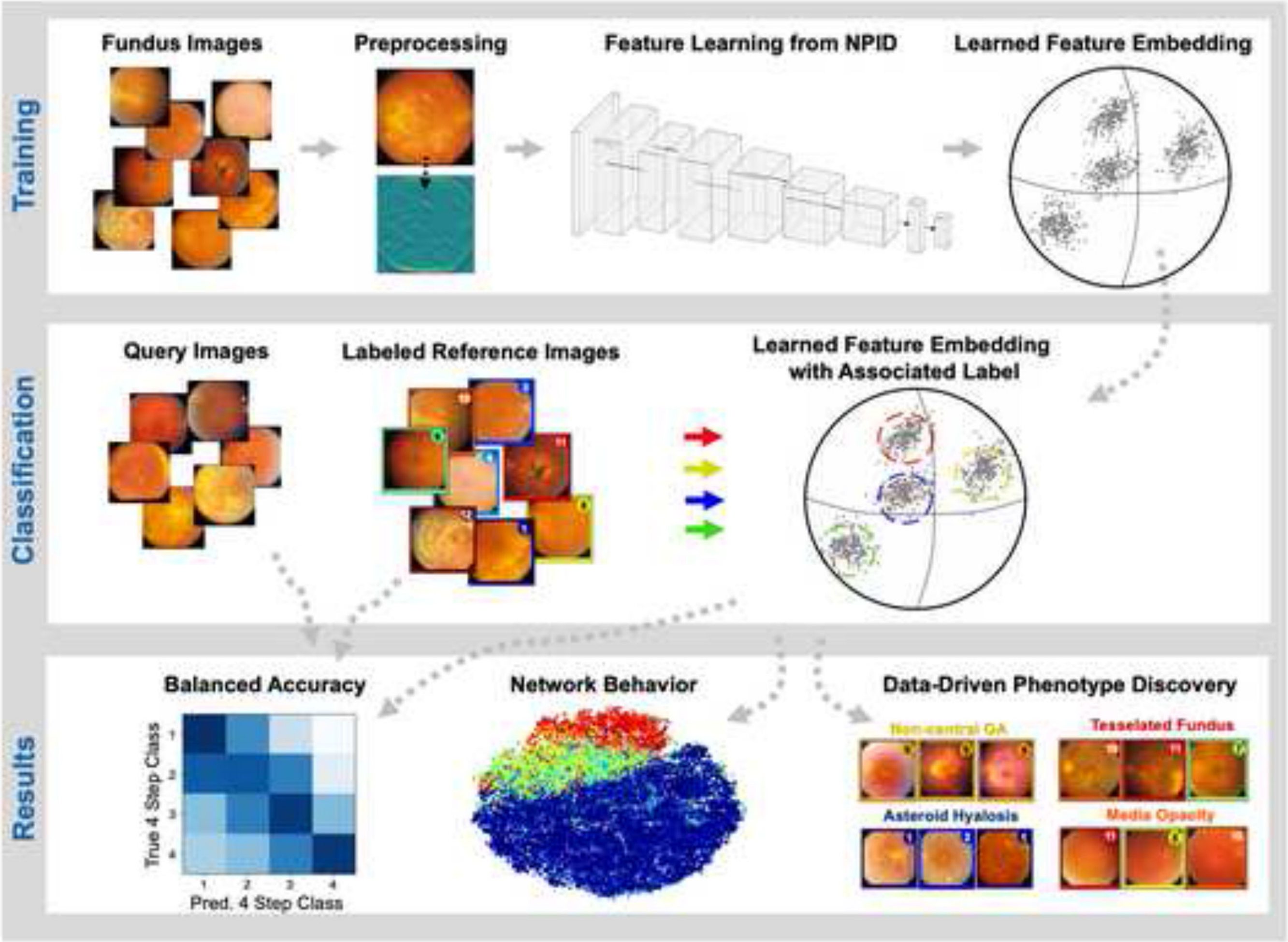
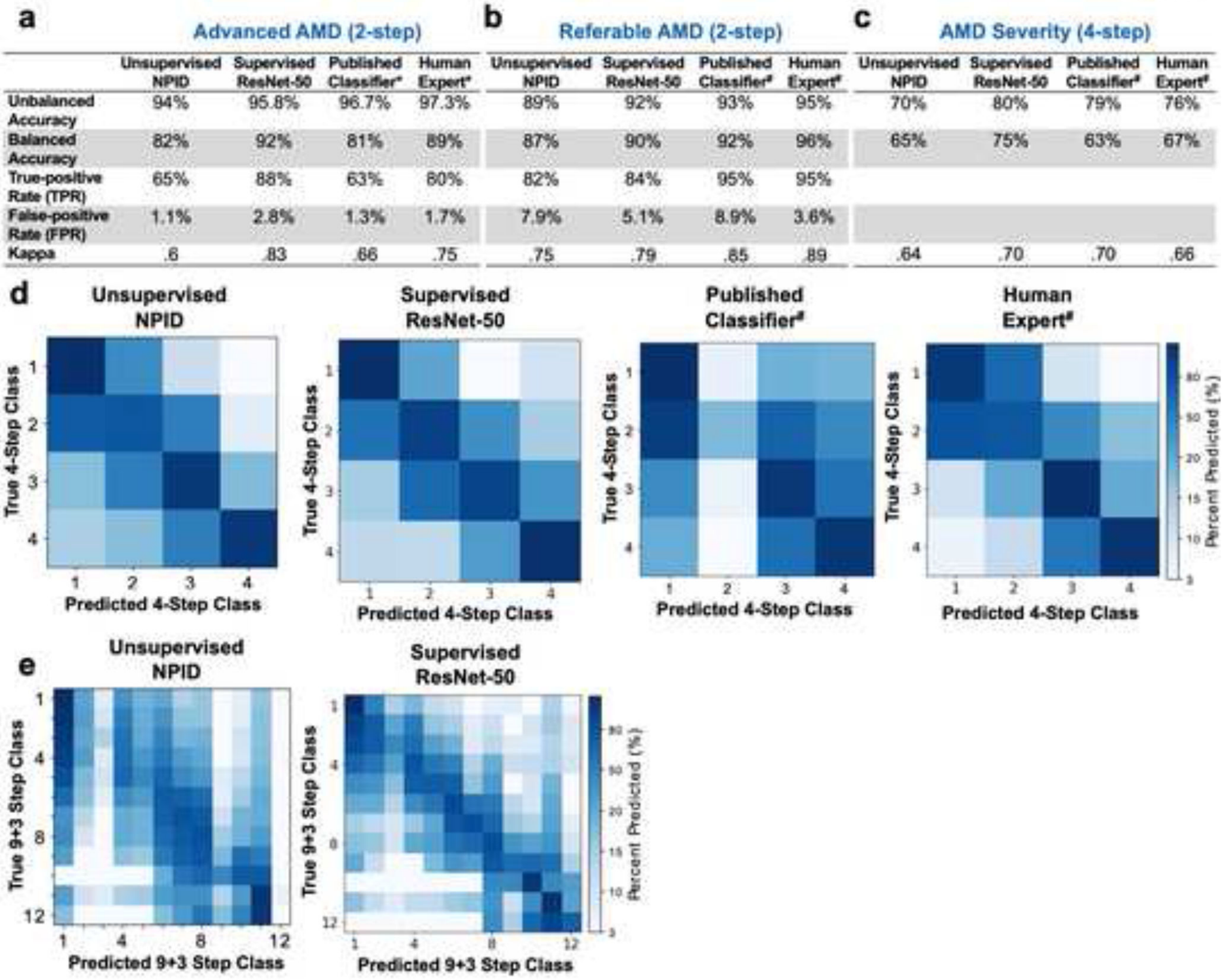


Figure 1: Schematic of NPID training & testing.

Schematic diagram of the process by which Non-Parametric Instance Discrimination (NPID) trains a self-supervised neural network to map preprocessed fundus images to embedded feature vectors. The feature vectors and associated AMD labels are used as a reference for queried severity discovery through neighborhood similarity matching. The NPID network can then be analyzed to measure balanced accuracy in AMD severity grading, explore visual features that drive network behavior, and discover novel AMD-related features and other ocular phenotypes in an unbiased, data-driven manner.



	a Advanced AMD (2-step)				b Referable AMD (2-step)				c AMD Severity (4-step)			
	Unsupervised NPID	Supervised ResNet-50	Published Classifier*	Human Expert*	Unsupervised NPID	Supervised ResNet-50	Published Classifier#	Human Expert#	Unsupervised NPID	Supervised ResNet-50	Published Classifier#	Human Expert#
Unbalanced Accuracy	94%	95.8%	96.7%	97.3%	89%	92%	93%	95%	70%	80%	79%	76%
Balanced Accuracy	82%	92%	81%	89%	87%	90%	92%	96%	65%	75%	63%	67%
True-positive Rate (TPR)	65%	88%	63%	80%	82%	84%	95%	95%				
False-positive Rate (FPR)	1.1%	2.8%	1.3%	1.7%	7.9%	5.1%	8.9%	3.6%				
Kappa	.6	.83	.66	.75	.75	.79	.85	.89	.64	.70	.70	.66

Figure 2: Comparison of NPID-trained performance with supervised-trained networks and human experts.

(a-c) Comparisons of the self-supervised-trained NPID network performance with a supervised-trained ResNet-50 network, as well as published supervised baselines and human ophthalmologists as reported by *Peng, et al. [14] and #Burlina, et al. [20] for binary classification of advanced AMD (a) or referable AMD (b), as well as the 4-step AMD severity scale (c). **(d)** Comparison of confusion matrices of our self-supervised-trained network with our supervised-trained network, published supervised baselines, and human expert gradings reported in #Burlina, et al. [20] for the 4-step AMD severity scale task. **(e)** Confusion matrices of the NPID network and our supervised-trained network on the 9+3-step AMD severity classification task.

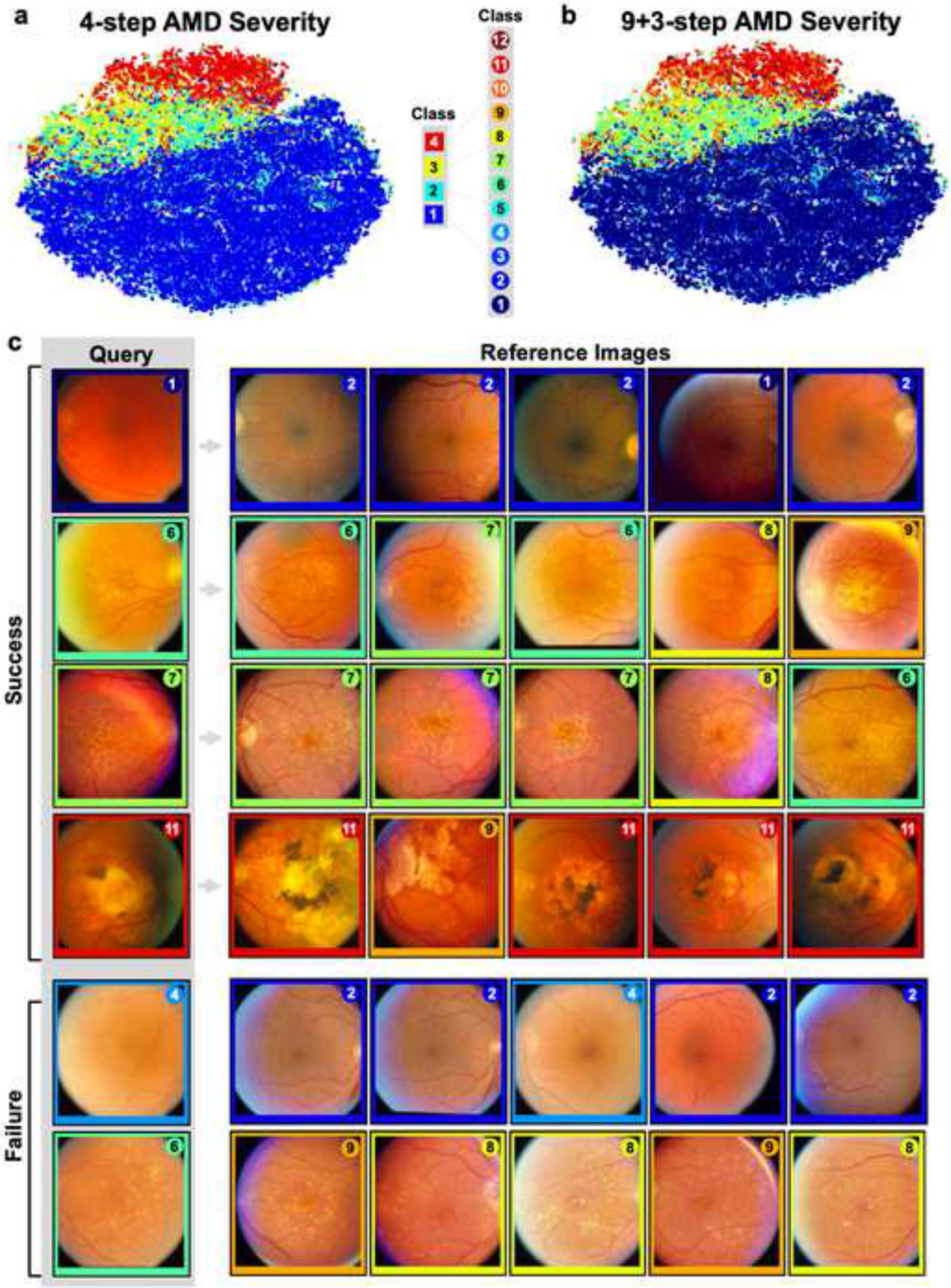


Figure 3. Self-supervised NPID clusters fundus images based on visual similarity

t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations of NPID feature vectors colored by **(a)** 4-step and **(b)** 9+3-step AMD severity labels, where each colored spot represents a single fundus image with AMD severity class as described in the legend and Supplemental Figure 1. **(c)** Representative search similarity images for successful and failed cases for the 9+3-step AMD severity scale task. The leftmost column corresponds to the query fundus image, while the next 5 images on each row correspond to the top 5 neighbors as defined by network features. The colored borders and numeric labels for each image define the true class label defined by the reading center for AREDS, and correspond to the color scheme in Supplemental Figure 1.

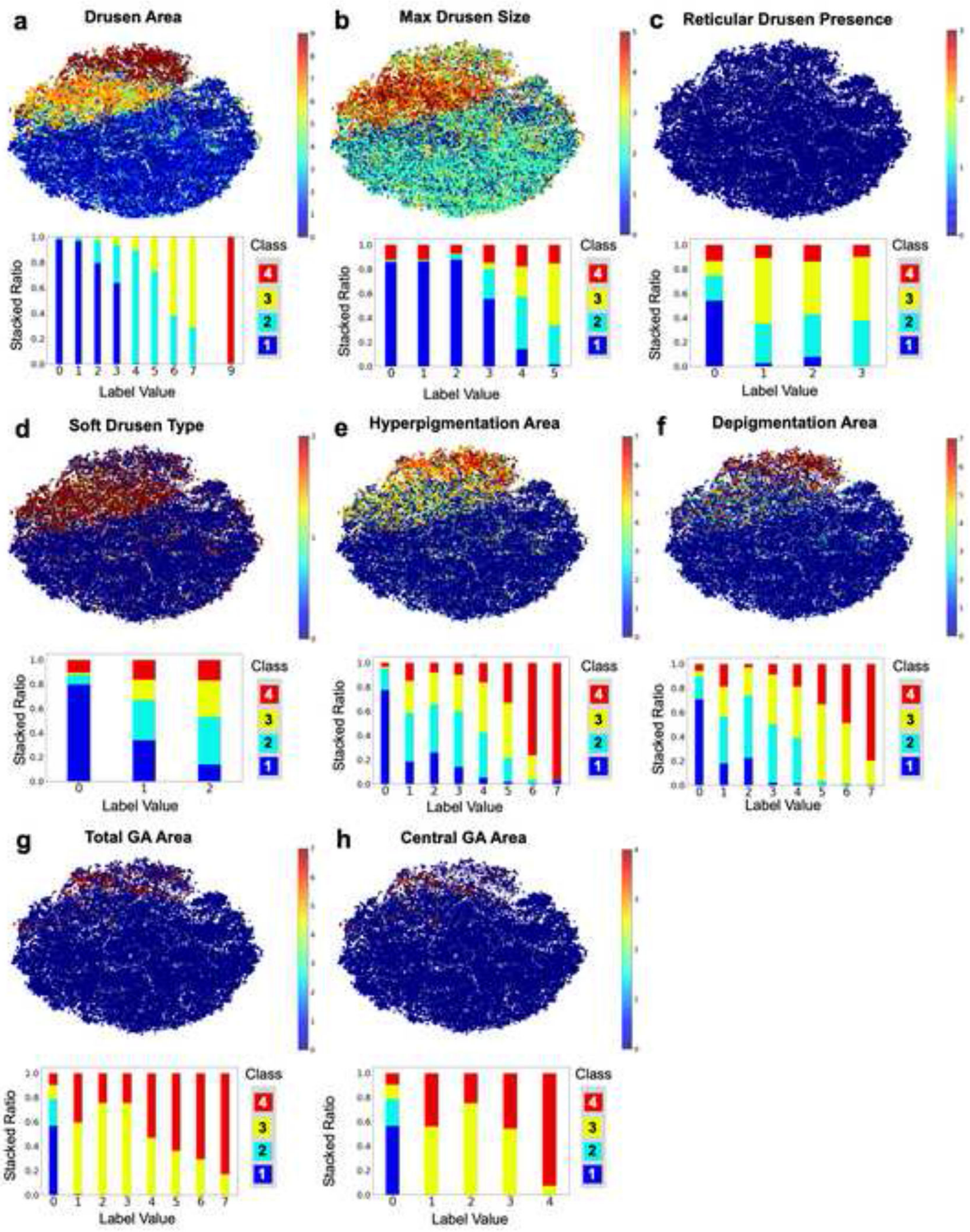


Figure 4. AMD-related fundus features that drive NPID-trained network predictions.
t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations of NPID feature vectors colored by AREDS reading center labels for AMD-related fundus features, with corresponding stacked bar plots showing ratio of each label across the 4-step AMD severity classes. Labels include **(a)** drusen area, **(b)** maximum drusen size, **(c)** reticular drusen presence, **(d)** soft drusen type, **(e)** hyperpigmentation area, **(f)** depigmentation area, **(g)** total geographic atrophy (GA) area, and **(h)** central GA area. Category definitions for each fundus feature are shown in Supplemental Table 1.

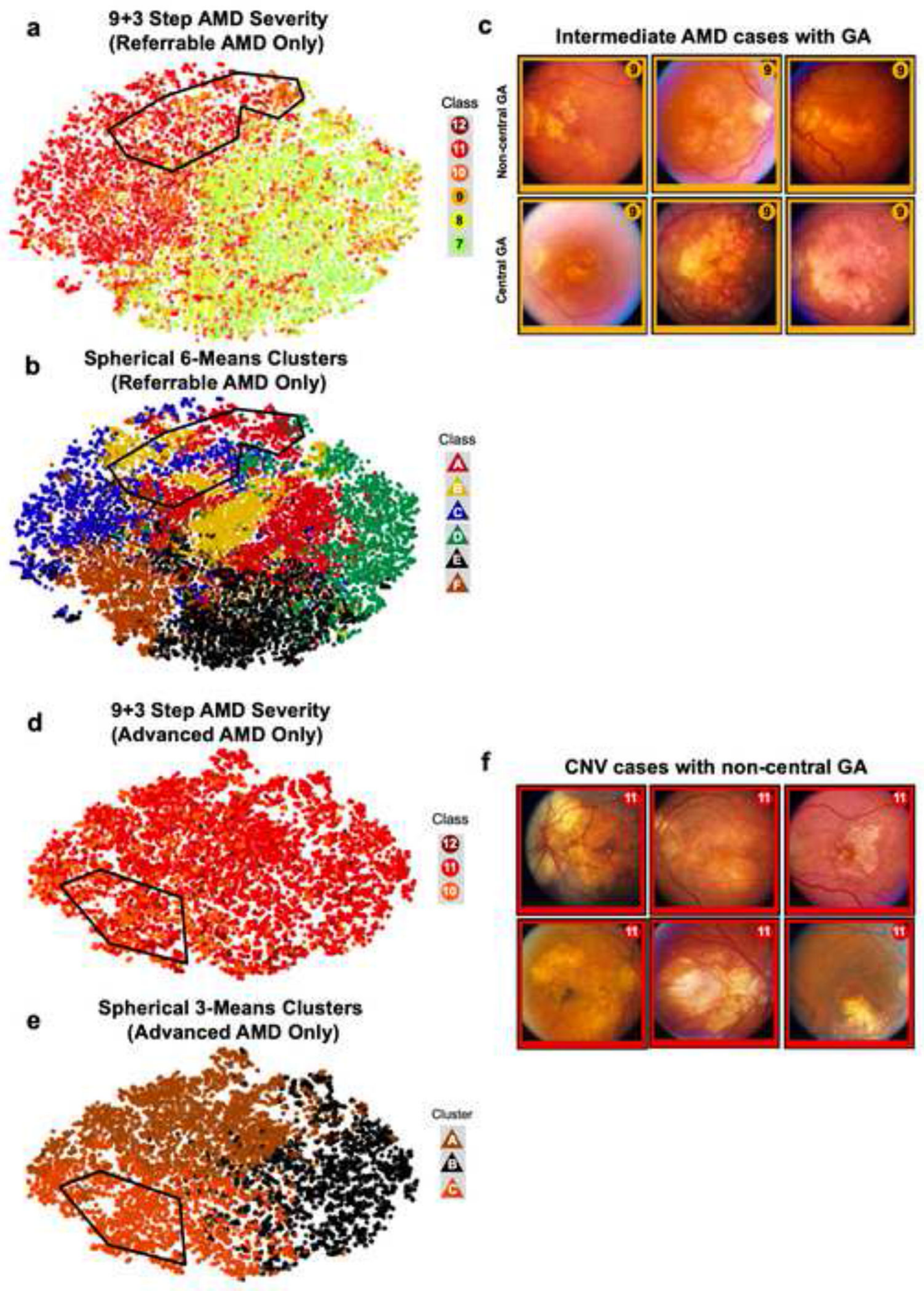


Figure 5. Data-driven discovery of central and non-central geographic atrophy.

t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations of NPID feature vectors colored by **(a)** 9+3-step AMD severity labels and **(b)** spherical K-means cluster labels with K=6, based on hierarchical learning using only fundus images with referable AMD (intermediate or advanced AMD). A selection (outlined area) of intermediate AMD cases (classes 7-9) adjacent to advanced AMD cases (classes 10-12) from clusters A-C show **(c)** fundus images with non-central GA (top row) and central GA (bottom row). t-SNE visualizations of NPID feature vectors colored by **(d)** with 9+3-step AMD severity labels and **(e)** spherical K-means cluster labels with K=3, based on hierarchical learning using only fundus images with advanced AMD (classes 10-12). A selection (outlined area) of CNV cases (class 11) adjacent to images with central GA with or without CNV (classes 10 and 12) from cluster C show **(f)** non-central GA.

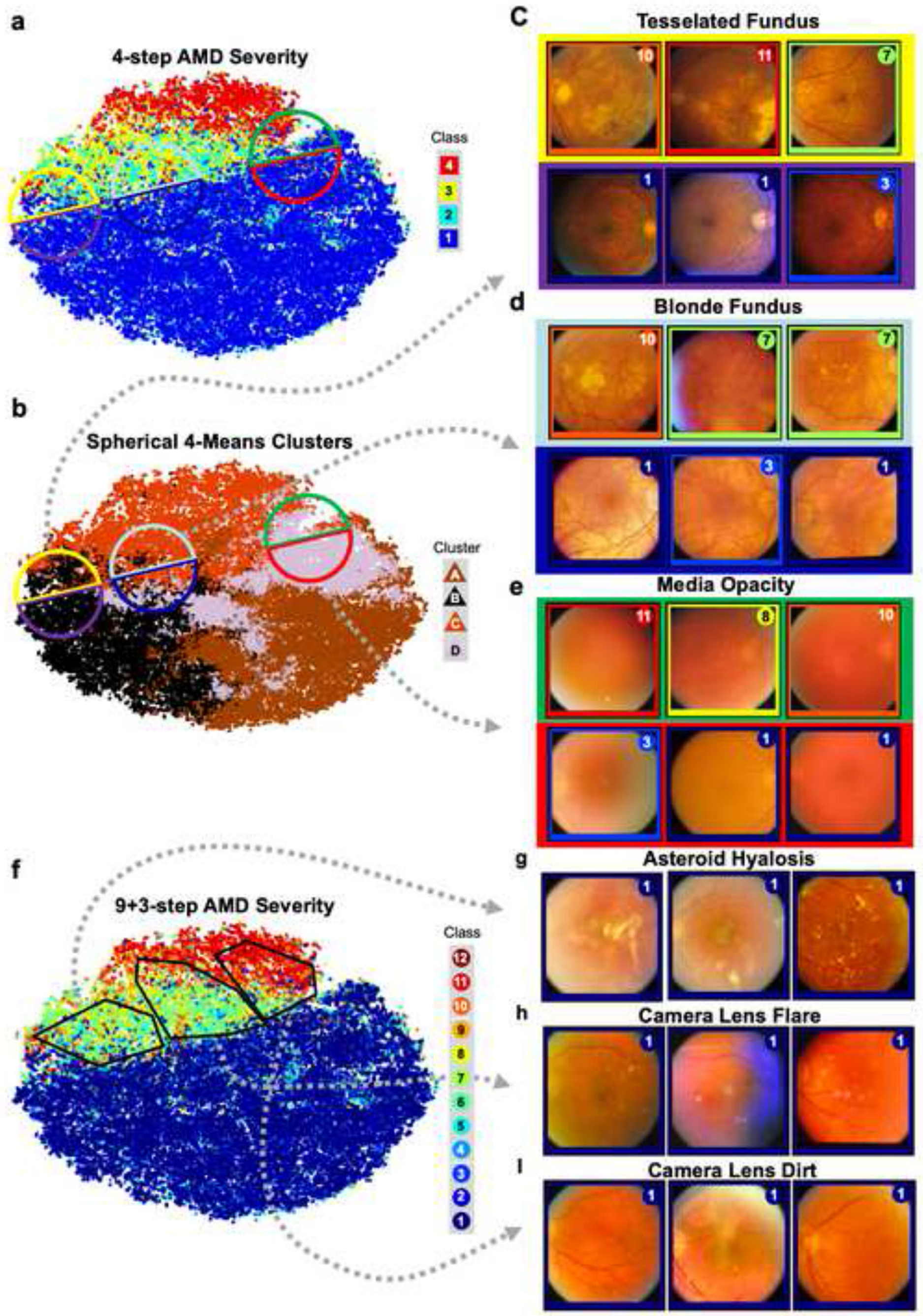


Figure 6. Data-driven discovery of ophthalmic features.

t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations of NPID feature vectors colorerd by **(a)** 4-step AMD severity labels and **(b)** spherical k-means (K=4) cluster labels. Fundus images that straddle no AMD vs. early, intermediate, or advanced AMD within K-means cluster B (yellow-purple circle), cluster C (teal-blue circle), and cluster D (green-red circle), corresponded to fundus images with **(c)** tessellated fundus, **(d)** blonde fundus, and **(e)** media opacity. **(f)** t-SNE visualization of 9+3-step AMD severity labels with a selection (outlined areas) of fundus images with no AMD (class 1) located within clusters of early, intermediate, or late AMD classes corresponded to fundus images with **(g)** asteroid hyalosis, **(h)** camera lens flare, and **(i)** camera lens dirt.