



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

Baladitya Yellapragada,^{1,2,3} Sascha Hornauer, PhD,² Kiersten Snyder,³ Stella Yu, PhD,^{1,2} Glenn Yiu, MD, PhD³

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 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 evaluate fundus photographs from the Age-Related Eye Disease Study (AREDS).

Participants: One hundred thousand eight hundred forty-eight human-graded fundus images from 4757 AREDS participants between 55 and 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 with those of 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 more granular AMD scales to misclassification by both ophthalmologists and neural networks. Importantly, self-supervised 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 predefined by human labels.

Conclusions: Self-supervised learning enables AMD severity grading comparable with that of 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. *Ophthalmology Retina* 2021;■:1–14 © 2021 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



Supplemental material available at www.opthalmologyretina.org.

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 error¹⁹ that is not discernable by human experts.

However, supervised learning approaches are trained using expert-defined labels that classify disease type or severity into discrete classes based on human-derived rubrics that are prone to bias and may not reflect the underlying disease pathophysiologic features accurately. Because supervised networks can identify only 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. Although 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, nonparametric instance discrimination (NPID) was designed previously for complex visual tasks.²⁴ Nonparametric instance discrimination predicts a query image's class label by determining the most common label among its nearest neighbors within a multidimensional hypersphere of encoded feature vectors drawn from training images. This technique significantly outperforms other unsupervised networks for ImageNet, Places (MIT Computer Science and Artificial Intelligence Laboratory), and PASCAL Visual Object Classes (Oxford) 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 pathophysiologic features and to enable unbiased discovery of ocular phenotypes. Because class boundaries are not established explicitly 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 graders²⁰ and supervised-trained CNNs, despite never learning the class definitions directly during training of the feature representations. Importantly, our examination of NPID behavior provided 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 showed 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 with that of supervised-trained neural networks or human experts, and discover novel pathologic or physiologic phenotypes that the algorithm was not specifically trained to detect.

Methods

Study Data Characteristics and Partitioning

Sponsored by the National Eye Institute, the AREDS enrolled 4757 participants 55 to 80 years of age in a prospective, randomized, placebo-controlled clinical trial to evaluate oral antioxidants as treatment for AMD. The AREDS design and results have been reported previously.²⁹ The study protocol was approved by a data and safety monitoring committee and the institutional review board for each participating center, adhered to the tenets of the Declaration of Helsinki, and was conducted before the advent of the Health Insurance Portability and Accountability Act. The AREDS sites obtained informed consent from participants, which was not necessary for 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 (accession no., phs000001, v3.p1.c2) after approval for authorized access and exemption by the institutional review board at the University of California, Davis. The median age of participants was 68 years, 56% were women, and 96% were White.^{30,31} Color fundus images from AREDS were graded previously 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- plus 3-step) AMD severity scale for each eye that predicts the 5-year progression risk to CNV or central GA,²⁹ with steps 1 through 3 representing no AMD, steps 4 through 6 representing early AMD, steps 7 through 9 representing intermediate AMD, and steps 10 through 12 representing advanced AMD including central GA (step 10), CNV (step 11), or both (step 12; [Supplemental Fig 1A](#), available at www.opthalmologyretina.org).^{29–32} Both the 9- plus 3-step scale and simplified 4-step scale have been used to train supervised CNNs successfully to classify AREDS fundus images for AMD severity.^{13,20} Because 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 participant 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- plus 3-step scale ([Supplemental Fig 1B, C](#)). Images were randomly into partitioned training, validation, and testing datasets in a 70:15:15 ratio, respectively, while ensuring that fundus images from the same participant did not appear across different datasets.

Data Preprocessing

Fundus images were down-sampled to 224 × 224 pixels along the short edge while maintaining the aspect ratio, as was done similarly in

a previous study.¹³ Fundus images were also preprocessed with a Laplacian filter applied in each of the red-green-blue color dimensions to better emulate the properties of more natural images of everyday scenes and objects (Supplemental Fig 2, available at www.ophtalmologyretina.org). 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 of 9 pixels in each of the red-green-blue 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 Fig 2A). The Laplacian-filtered fundus images more closely resemble those of natural statistics (Supplemental Fig 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 (ImageNet Project) (i.e., initialize the neurons across naturalistic filters) and then fine-tuned on the AREDS dataset without any weights frozen to improve performance further. 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 Fig 3, available at www.ophtalmologyretina.org).

Nonparametric Instance Discrimination Training and Prediction

Nonparametric instance discrimination discriminates unlabeled training images using instance-based classification of feature vectors in a spherical feature space. At its core, NPID uses a backbone Residual Network 50 (ResNet-50, Microsoft Research), 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 (Fig 1). 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 the Supplemental Appendix (available at www.ophtalmologyretina.org), 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 Although this method has traditionally been 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. The wkNN function is more appropriate for evaluating NPID compared with other classification protocols for

evaluating self-supervised networks because it requires no additional training (so no change in interpretation of the representations learned from NPID occurs), 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 $k = 12$, because it produced the highest balanced accuracies (Supplemental Fig 4, available at www.ophtalmologyretina.org). We chose the epoch that yielded the best balanced accuracy using wkNN classification voting scheme (Supplemental Appendix). Then, we evaluated that epoch on a separate testing dataset using various metrics from the wkNN result, including unbalanced accuracy, Cohen's κ , 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} Although 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 used a second method to evaluate self-supervised features using linear support vector machines.³⁵

Supervised Training and Prediction

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

t-Distributed Stochastic Neighbor Embedding Visualization and 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 topologic features. Herein, we mapped the encoded 64-dimension features onto 2-dimensional (2D) coordinates, wherein coordinates that are near each other in 2D are also near each other in the original feature space, meaning that they are encoded similarly because they share visual features (Fig 1). 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 (G.Y.) and thus are 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

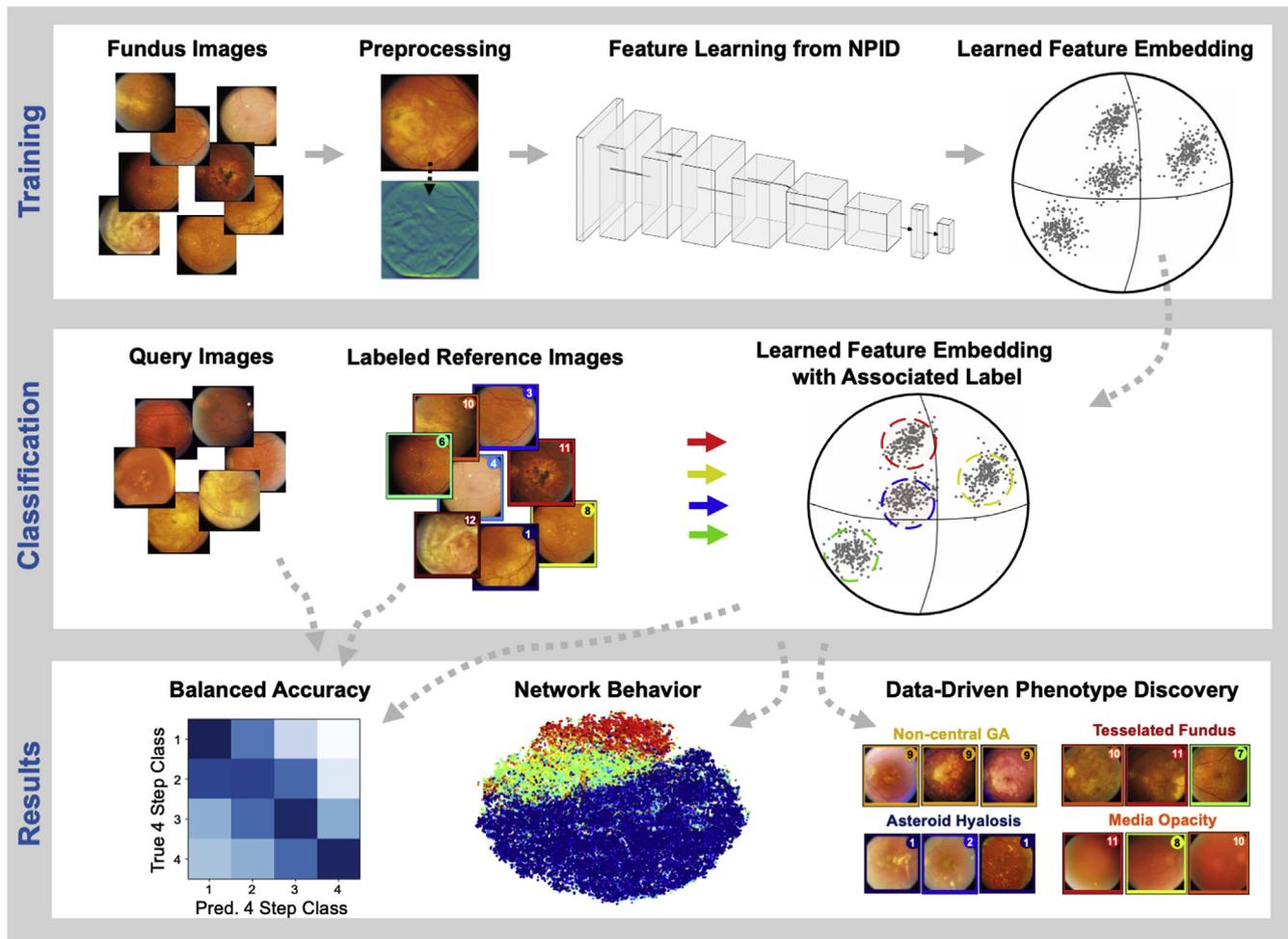


Figure 1. Schematic diagram of the process by which nonparametric instance discrimination (NPID) trains a self-supervised neural network to map preprocessed fundus images to embedded feature vectors. The feature vectors and associated age-related macular degeneration (AMD) labels are used as a reference 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. GA = geographic atrophy.

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 seems more suitable for coarse-level than fine-level classification across dependent classes, we split up the 9- plus 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 3 fine 9- plus 3-step classes within each coarse 4-step class to identify which of the 9- plus 3-step classes seem 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 used 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 defined notably by or distinct from the original labels for further analysis (Fig 1).

Results

Accuracy in Grading Age-Related Macular Degeneration Severity

We first evaluated NPID performance on a 2-step discrimination task for detecting advanced AMD (CNV, central GA, or both) and found that wkNN applied to the self-supervised-learned features achieved an unbalanced accuracy (94%) that is comparable with 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 because of dataset imbalance, was also similar between the self-supervised-trained NPID (82%), our supervised-trained

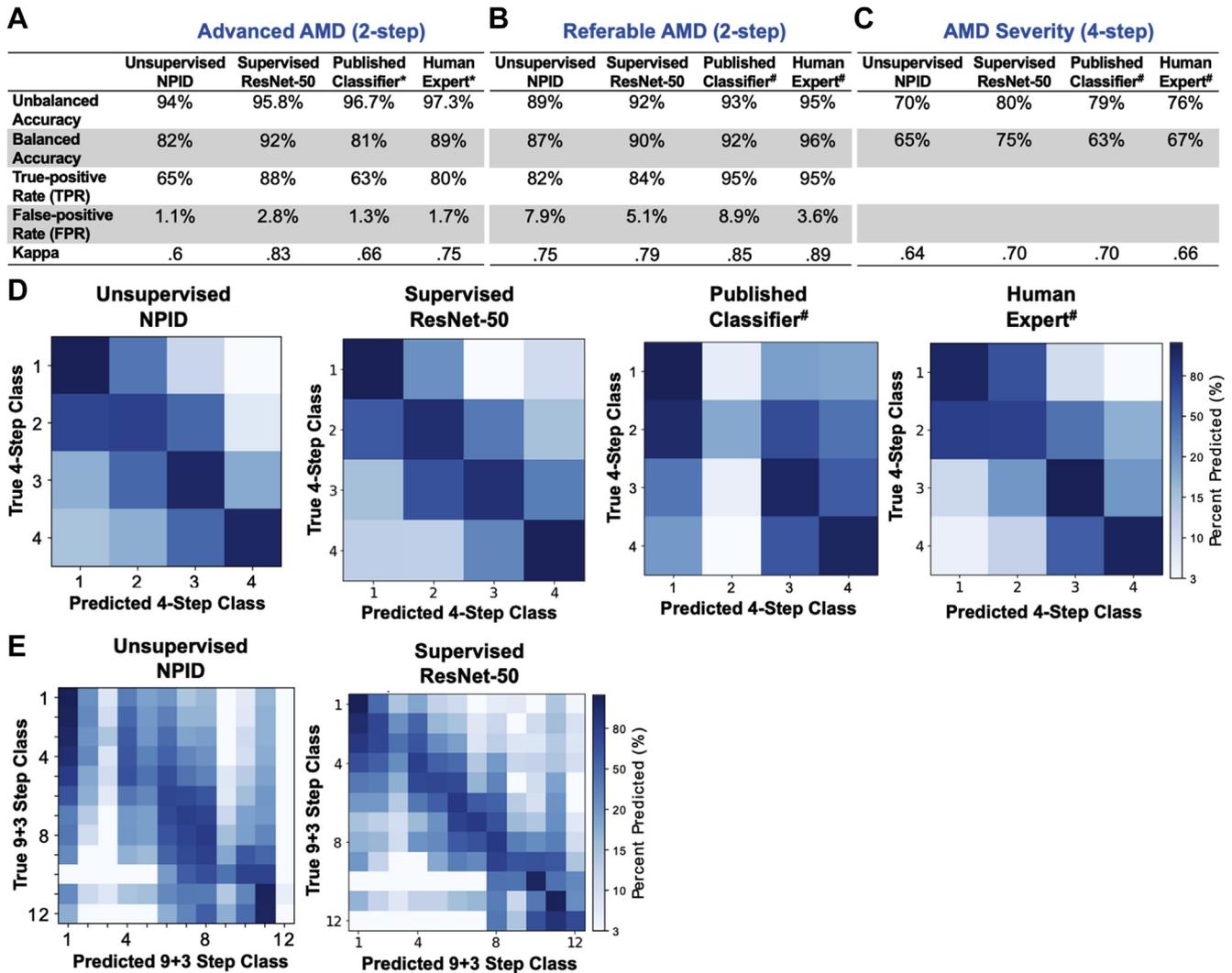


Figure 2. A-C, Comparisons of the self-supervised—trained nonparametric instance discrimination (NPID) network performance with that of a supervised-trained ResNet-50 network, as well as published supervised baselines and that of human ophthalmologists as reported by *Peng et al¹⁴ and *Burlina et al²⁰ for binary classification of (A) advanced age-related macular degeneration (AMD) or (B) referable AMD, as well as (C) the 4-step AMD severity scale. 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 al²⁰ for the 4-step AMD severity scale task. E, Confusion matrices of the NPID network and our supervised-trained network on the 9- plus 3-step AMD severity classification task.

network (92%), the published supervised network (81%), and an ophthalmologist (89%; Fig 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 an ophthalmologist (96%),²⁰ despite never learning the class definitions directly (Fig 2B). For grading AMD severity using the 4-step scale, NPID achieved a 65% balanced accuracy, which was comparable with that of our supervised-trained network (75%), the published network (63%), and an ophthalmologist (67%; Fig 2C).²⁰ In particular, the confusion matrix for NPID demonstrated superior performance for distinguishing early AMD (class 2) as compared with both the published supervised network and human expert (Fig 2D).²⁰

When applied to a finer classification task, NPID achieved a balanced accuracy of only 25% on the 9- plus 3-step scale, as compared with 40% using our supervised-trained network and 74% using the published supervised network¹³ that used the same backbone network as our NPID approach. We achieved this balanced accuracy score using $k = 12$ for wkNN, although we also tested $k = 5$, $k = 8$, $k = 23$, and $k = 50$ and found that results were mostly consistent across different k values (Supplemental Fig 4). Although our most class-homogenous neighborhoods were defined by $k = 12$ neighbors, they remained mostly coherent with $k = 50$ neighbors, which was how NPID was evaluated originally 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- plus 3-step label (Fig 2E). Even for cases with incorrect 9- plus 3-step class predictions, the 50 nearest neighbor images shared the query's 4-step class label 56% of the time, which

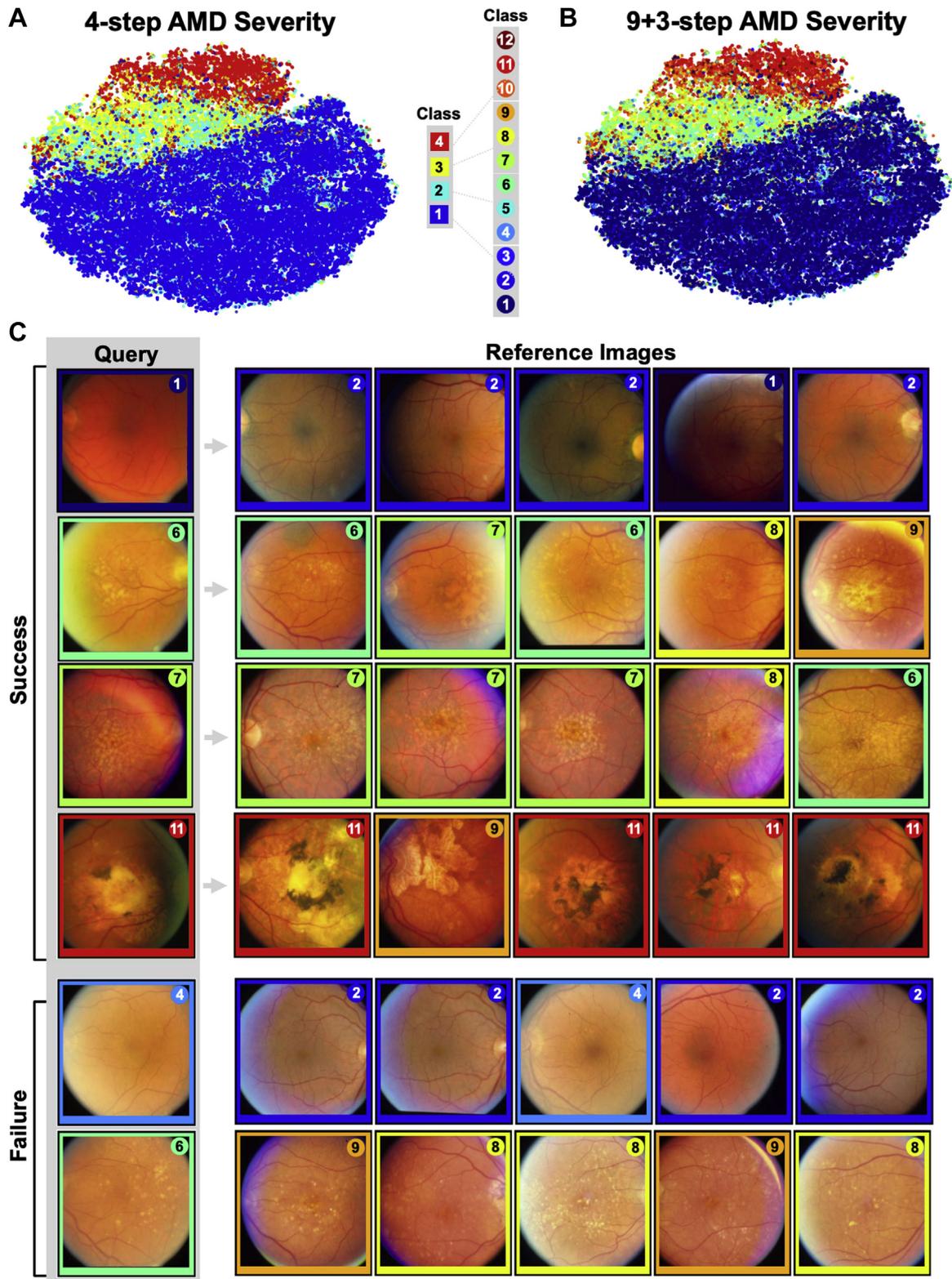


Figure 3. A, B, t-Distributed stochastic neighbor embedding visualizations of nonparametric instance discrimination (NPID) feature vectors colored by (A) 4-step and (B) 9- plus 3-step age-related macular degeneration (AMD) severity labels, where each colored spot represents a single fundus image with 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- plus 3-step AMD severity scale task. The leftmost column corresponds to the query fundus image, whereas 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 the Age-Related Eye Disease Study and correspond to the color scheme in [Supplemental Figure 1](#).

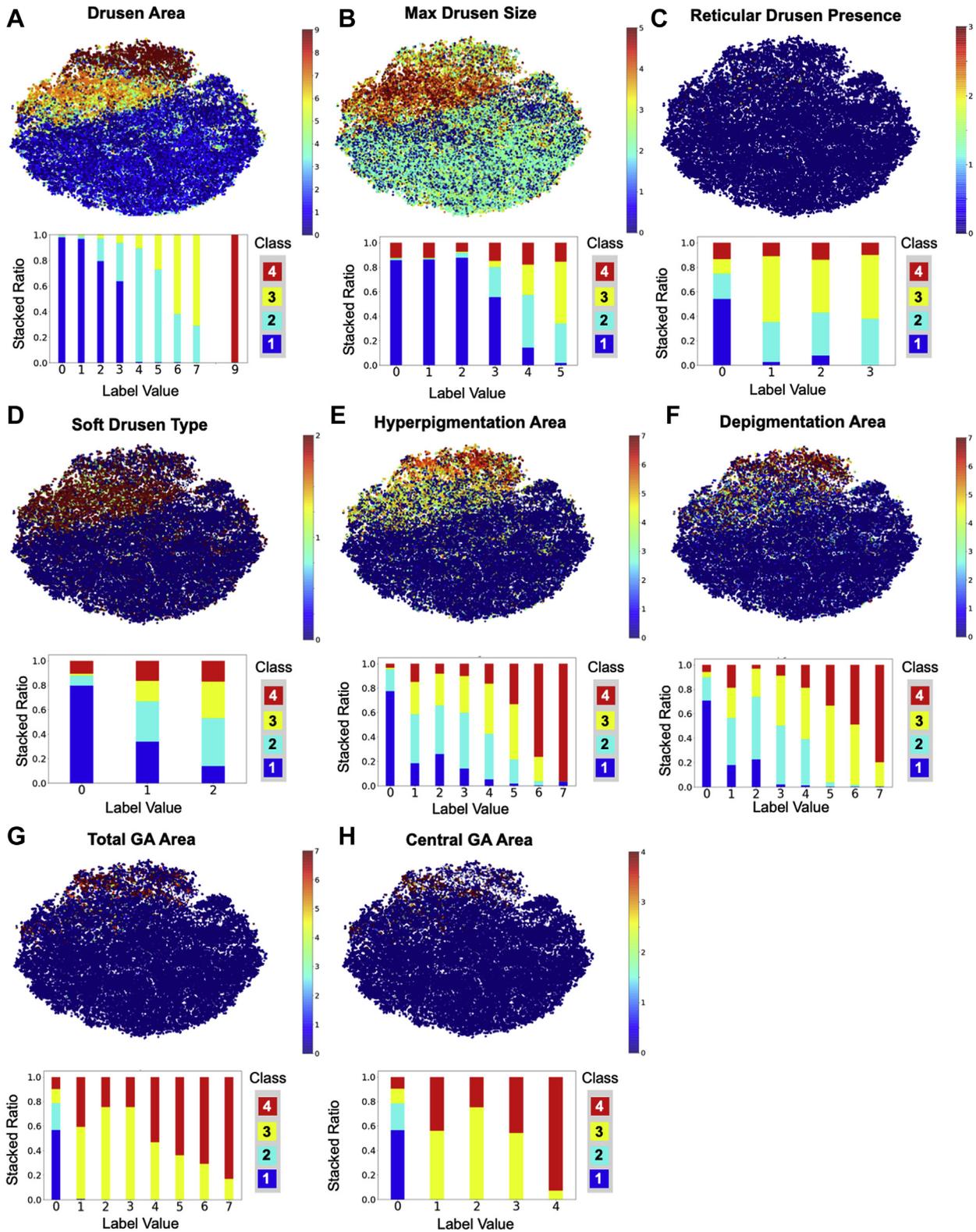


Figure 4. A–H, *t*-Distributed stochastic neighbor embedding visualizations of nonparametric instance discrimination (NPID) feature vectors colored by Age-Related Eye Disease Study reading center labels for age-related macular degeneration (AMD)-related fundus features, with corresponding stacked bar plots showing the 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](#) (available at www.opthalmologyretina.org).

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- plus 3-step AMD severity scale compared with binary or 4-step AMD classifications, incorrect predictions deviate minimally from ground-truth labels. We confirmed our findings using linear support vector machine classifiers, which achieved a 26% balanced accuracy for 9-plus 3-step classification consistent with the wkNN results.

Network Behavior for Grading Age-Related Macular Degeneration Severity

To discern how the NPID network visually organizes images from different AMD classes, we used 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, whereas early AMD (aqua or green) images were scattered throughout the plot (Fig 3A), which likely explains the lower performance in this class (Fig 2D). On the 9-step AMD severity scale (Fig 3B), the t-SNE plot seems similar to that of the 4-step scale, because each of the 4 major classes on the simplified scale are dominated by 1 or 2 of the finer classes within each subset (Supplemental Fig 1B) and may account for the poorer performance of our self-supervised-trained network on the 9-plus 3-step task.

Examining the training images that contribute to NPID predictions helps to explain the self-supervised-trained network's behavior in an interpretable way that supervised-trained networks cannot, because the specific training images that drive a supervised network's predictions cannot be recovered easily. In our study, comparison of query images with a selection of neighboring reference images demonstrated high phenotypic similarity across adjacent 9- plus 3-step classes (Fig 3C) and explained class confusions that contributed to NPID performance loss on the finer-grained 9-plus 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- plus 3-step scale, with many class 4 images that resembled no AMD and class 6 images that seemed similar to intermediate AMD (Fig 3C, bottom rows), which helps to explain the difficulty with distinguishing early AMD images by the NPID method, as well as by supervised-trained networks and human ophthalmologists (Fig 2D, E).

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 (Fig 4). We found that drusen area provided the most visually distinct clusters that matched the separation of the 4-step severity scale. Geographic atrophy area and depigmentation correlated well with advanced AMD classes as expected, whereas 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 can provide insight into the anatomic features that may drive AMD severity predictions.

Data-Driven Age-Related Macular Degeneration Phenotype Discovery

Current AMD severity scales demonstrate human bias because they were developed in part to reflect clinical severity (i.e., impact on visual function) rather than disease pathophysiologic features. For example, only vision-threatening GA involving the central macula was ascribed as advanced AMD (class 10 on the 9- plus 3-step scale), whereas noncentral GA cases were scattered across other AMD classes. With the goal of extracting features of any GA, both central and noncentral, 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 (Fig 5A–C).

To delineate the feature pockets that define the GA phenotype more objectively, 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 referable AMD subset (classes 7–12), we found 3 clusters (clusters A, B, and C) among eyes with intermediate AMD that corresponded to variable degrees of GA (Fig 5A, B), including noncentral GA (Fig 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 (Fig 5C, bottom row).

Because the advanced AMD hierarchical subset predictably demonstrated the greatest separability among its 3 fine-grained 9-plus 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 (Fig 5D, E). Herein, we discovered 1 of the 3 clusters (cluster C) to contain 75% of class 10 (central GA) and class 12 (central GA plus CNV) fundus images. Sampling from class 11 images within this unbiased cluster also revealed images with noncentral GA, although class 11 was agnostic to GA presence (Fig 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-Age-Related Macular Degeneration 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 1 cluster (cluster A) that corresponded to images with no AMD and 3 other clusters (clusters B, C, and D) that seemed to straddle AMD classes, suggesting that these latter groups may be distinguished by features unrelated to AMD pathophysiologic features (Fig 6A, B). 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 (Fig 6C), a feature associated with choroidal thinning and high myopia.⁴¹ Cluster C images near this border contain fundi with a blonde appearance (Fig 6D), often found in patients with light-colored

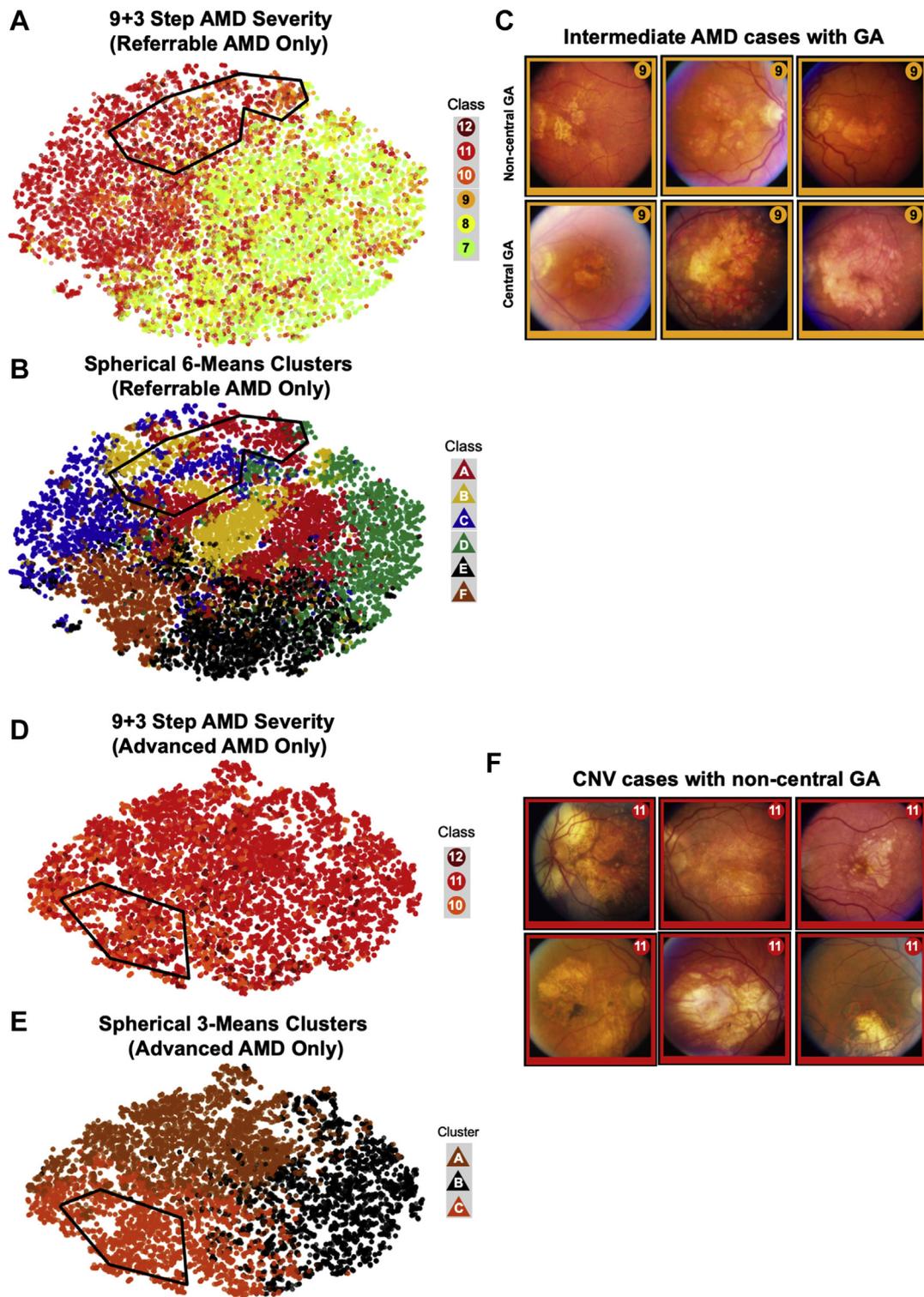


Figure 5. A, B, t-Distributed stochastic neighbor embedding visualizations of nonparametric instance discrimination (NPID) feature vectors colored by (A) 9- plus 3-step age-related macular degeneration (AMD) severity labels and (B) spherical K-means cluster labels with $K = 6$, based on hierarchical learning using only fundus images with referable AMD (intermediate or advanced AMD). C, A selection (outlined area) of intermediate AMD cases (classes 7–9) adjacent to advanced AMD cases (classes 10–12) from clusters (A–C) showing fundus images with noncentral geographic atrophy (GA; top row) and central GA (bottom row). D, E, t-Distributed stochastic neighbor embedding visualizations of NPID feature vectors colored with (D) 9- plus 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). F, A selection (outlined area) of choroidal neovascularization (CNV) cases (class 11) adjacent to images with central GA with or without CNV (classes 10 and 12) from cluster C show noncentral GA.

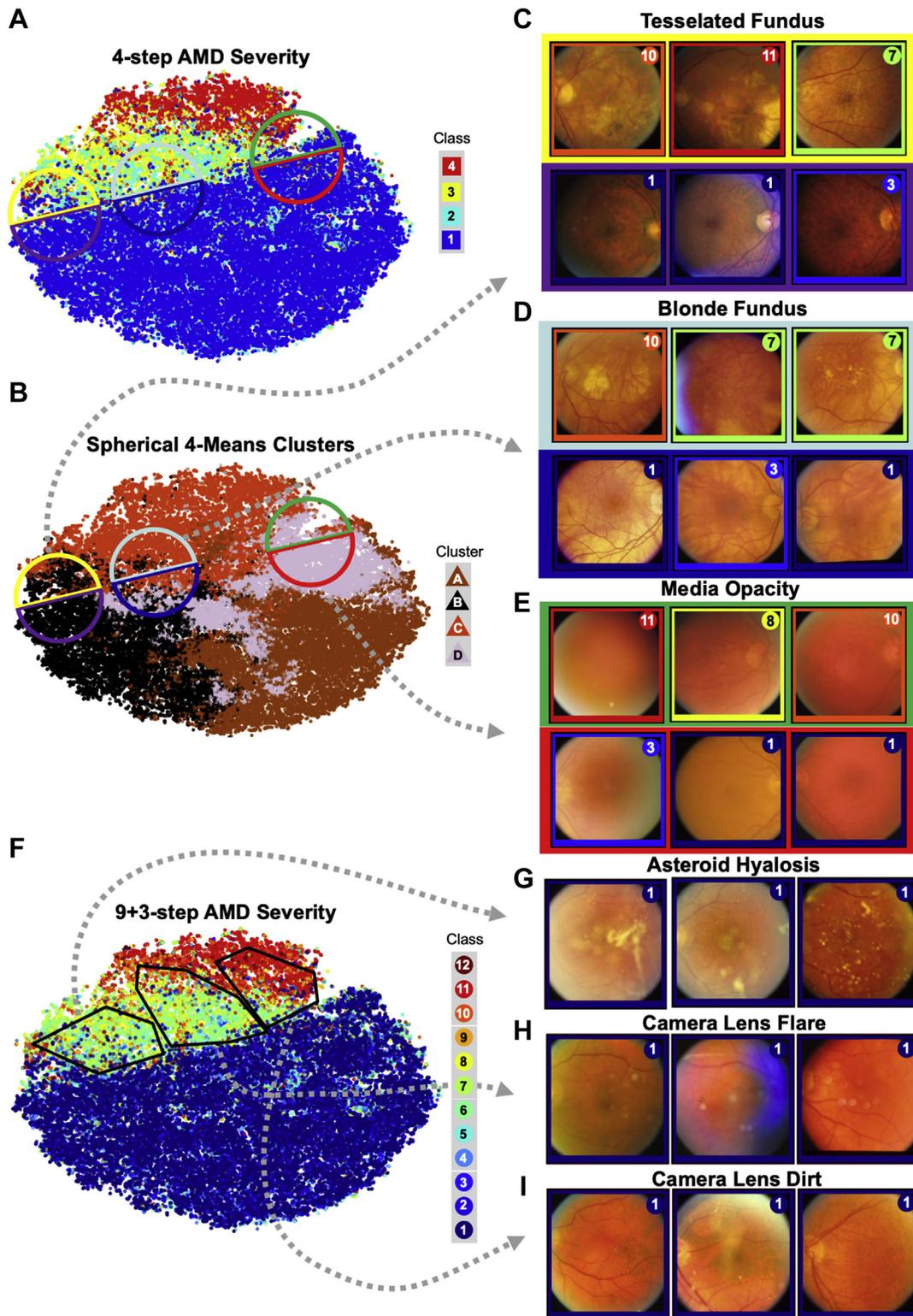


Figure 6. A, B, t-Distributed stochastic neighbor embedding visualizations of nonparametric instance discrimination (NPID) feature vectors colored by (A) 4-step age-related macular degeneration (AMD) severity labels and (B) spherical K-means ($K = 4$) cluster labels. C–E, Fundus images that straddle no AMD versus 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, respectively. F–I, t-Distributed stochastic neighbor embedding visualization of 9- plus 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.

skin and eyes or 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 (Fig 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 subcapsular opacity from corresponding slit-lamp images obtained in the AREDS and found that eyes in cluster D corresponded to a higher degree of both nuclear and cortical cataracts (Supplemental Fig 5, available at www.opthalmologyretina.org). Hence, fundus images contain other ophthalmologically relevant information that is 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 (Fig 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 (Fig 6G–I), which may resemble AMD features to a 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- plus 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 that could be used in combination with a simple classifier to predict AMD severity across different human-defined classification schema, revealed AMD features that drive network behavior, and identified novel pathologic and physiologic ocular phenotypes, all without the bias and constraints of human-assigned labels during the training process. Nonparametric instance discrimination performance was comparable with 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, classes 2 through 8, and classes 9 through 12 on the 9- plus 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 (Fig 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- plus 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 can be compared readily with that of 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 AREDS 2 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 herein, compared with $k = 50$). This is an advantage of self-supervised over supervised training approaches, because 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 28,135⁴³). These previous studies often excluded stereoscopic duplicates or same-eye images from different visits to avoid associating nonrelevant 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 1 prior study overrepresented eyes with intermediate (33%) and late (33%) AMD 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 that 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 κ value for another supervised network compared with NPID, despite a lower balanced accuracy, true-positive rate, and false-positive rate.¹⁴

Our findings are remarkable because although 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 pathophysiologic features, such as distinguishing central from noncentral GA because of 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- plus 3-step AMD severity scale, where subtle differences in phenotype such as drusen size or pigmentation are categorized arbitrarily 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 (Fig 3A, B), 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 2 of the 9- plus 3-step classes dominate each 4-step class because of the low visual variability among them. These findings support the need to re-evaluate 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 noncentral GA among those with intermediate or advanced AMD based on proximity to eyes with central GA (class 10), although this feature is not encoded in the human-labeled AMD severity scales. Our findings suggest that self-supervised learning can identify certain AMD phenotypes more objectively, such as drusen area or GA presence, which may reflect disease pathophysiology better and may enable the development of more unbiased, data-driven classification of AMD severity or subtypes that could predict disease outcomes better than human-assigned grades. Interestingly, K-means clustering also identified images with central GA that seemed to be mislabeled as intermediate AMD, further highlighting the ability of a self-supervised—trained network to discover mis-categorized images in ways that label-driven supervised learning cannot.

Another interesting feature of self-supervised learning is the ability to identify nonretinal phenotypes from fundus images, including camera artifacts (lens dirt or flare), media opacity (cataracts or asteroid hyalosis), and choroidal patterns (tessellated or blonde fundus). Although 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 they are trained.

Because fundus photographs exhibit very little visual and semantic variability compared with natural object images, we found that preprocessing by 2D Laplacian-filtering transformed the spatial frequency power spectra of fundus images to resemble natural objective images better (Supplemental Fig 2) and that pretraining with ImageNet before fine-tuning on AREDS increased network performance for the 9- plus 3-step and 4-step tasks by 100% and 33% (Supplemental Fig 3), respectively, compared with no ImageNet pretraining. This performance improvement implies that discriminating features relevant to the task were difficult to learn directly from the fundus photographs but improved with transfer learning, as seen on t-SNE comparisons with and without pretraining, which demonstrated extra learned features that correlate better 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 that of 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, over-representation 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 pathophysiologic features and pave the way to more objective and robust classification schemes for complex, multifactorial eye diseases.

Footnotes and Disclosures

Originally received: February 6, 2021.

Final revision: June 24, 2021.

Accepted: June 25, 2021.

Available online: ■■■■

Manuscript no. 2021-56.

¹ Department of Vision Science, University of California, Berkeley, Berkeley, California.

² International Computer Science Institute, Berkeley, California.

³ Department of Ophthalmology & Vision Science, University of California, Davis, Sacramento, California.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

S.Y.: Consultant - Intel Research, KLA Tencor, iCueMotion, Cerebral Therapeutics; Financial support - Leidos, Etegent Technologies, Samsung,

Glidewell Laboratories, Bosch Research, Amazon Research, Google Research, Facebook AI Research

G.Y.: Consultant - Alimera, Allergan, Carl Zeiss Meditec, Clearside Biomedical, Genentech, Intergalactic Therapeutics, Iridex, Regeneron, Topcon, Verily; Financial support - Clearside Biomedical, Genentech, Iridex

Supported by the CITRIS/Banatao Institute (B.Y.); the National Institutes of Health, Bethesda, Maryland (grant nos.: K08 EY026101 [G.Y.] and R21 EY031108 [G.Y.]); BrightFocus Foundation (G.Y.); Macula Society (G.Y.); the National Science Foundation (S.Y., S.H.); the National Geospatial Agency (S.Y., B.Y.); the Defense Advanced Research Projects Agency (S.Y.); the United States Geological Survey (S.Y.); the United States Fish and Wildlife Service (S.Y.); Lawrence Berkeley National Lab (S.Y.); the University of California, Berkeley, Berkeley, California (S.H., B.Y.). 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.

HUMAN SUBJECTS: Human subjects were included in this study. The institutional review board at the University of California, Davis deemed this study exempt as only de-identified image data from human subjects were used. All research adhered to the tenets of the Declaration of Helsinki. The Age-Related Eye Disease Study sites obtained informed consent from participants, which was not necessary for this post hoc analysis on the fundus data; digitized Age-Related Eye Disease Study color fundus photographs and study data were obtained from the National Eye Institute's Online Database of Genotypes and Phenotypes website (dbGaP accession

no., phs000001, v3.p1.c2) after approval for authorized access and exemption by the institutional review board at the University of California, Davis.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Yellapragada, Yu, Yiu

Analysis and interpretation: Yellapragada, Hornauer, Yiu

Data collection: Yellapragada, Snyder, Yiu

Obtained funding: Yiu

Overall responsibility: Yellapragada, Hornauer, Snyder, Yu, Yiu

Abbreviations and Acronyms:

AMD = age-related macular degeneration; **AREDS** = Age-Related Eye Disease Study; **CNN** = convolutional neural network; **CNV** = choroidal neovascularization; **GA** = geographic atrophy; **NPID** = nonparametric instance discrimination; **t-SNE** = t-distributed stochastic neighbor embedding; **2D** = 2-dimensional; **wkNN** = weighted *k*-nearest neighbors.

Keywords:

Age-related macular degeneration, AMD, Artificial intelligence, Deep learning, Machine learning.

Correspondence:

Glenn Yiu, MD, PhD, Department of Ophthalmology & Vision Science, University of California, Davis, 4860 Y Street, Suite 2400, Sacramento, CA 95817. E-mail: gyiu@ucdavis.edu.

References

1. LeCun Y, Yoshua Bengio GH. Deep learning. *Nature*. 2015;521(7553):436–444. <https://doi.org/10.1038/nature14539>.
2. Zhou B, Khosla A, Lapedriza A. Object detectors emerge in deep scene CNNs In: 3rd International Conference on Learning Representations ICLR 2015—Conference Track Proceedings 2015. Available at: <https://arxiv.org/abs/1412.6856>.
3. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115: 211–252. <https://doi.org/10.1007/s11263-015-0816-y>.
4. Yu H, Yang W, Xia GS, Liu G. A color-texture-structure descriptor for high-resolution satellite image classification. *Remote Sens*. 2016;8(3). <https://doi.org/10.3390/rs8030259>.
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. <https://arxiv.org/pdf/1704.07911.pdf>.
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*. 2016;316(22):2402–2410. <https://doi.org/10.1001/jama.2016.17216>.
7. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124(7): 962–969. <https://doi.org/10.1016/j.ophtha.2017.02.008>.
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*. 2017;318(22):2211–2223. <https://doi.org/10.1001/jama.2017.18152>.
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(26): 552–564. <https://doi.org/10.1016/j.ophtha.2018.11.016>.
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. *Invest Ophthalmol Vis Sci*. 2016;57:5200–5206. <https://doi.org/10.1167/iovs.16-19964>.
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(8): 1264–1272. <https://doi.org/10.1016/j.ophtha.2018.01.034>.
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(11):1170–1176. <https://doi.org/10.1001/jamaophthalmol.2017.3782>.
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(9): 1410–1420. <https://doi.org/10.1016/j.ophtha.2018.02.037>.
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(4):565–575. <https://doi.org/10.1016/j.ophtha.2018.11.015>.
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(1): 15–22. <https://doi.org/10.1016/j.ogla.2018.04.002>.
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(1). <https://doi.org/10.1038/s41598-018-35044-9>.

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(1):95–96. <https://doi.org/10.1016/j.ophtha.2019.05.029>.
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(3):158–164. <https://doi.org/10.1038/s41551-018-0195-0>.
19. Varadarajan AV, Poplin R, Blumer K, et al. Deep learning for predicting refractive error from retinal fundus images. *Invest Ophthalmol Vis Sci*. 2018;59(7):2861–2868. <https://doi.org/10.1167/iovs.18-23887>.
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;82:80–86. <https://doi.org/10.1016/j.compbio.2017.01.018>.
21. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. In: Jennifer Dy, Andreas Krause, eds. 35th International Conference on Machine Learning. Proceedings of Machine Learning Research; 2018:2127–2136.
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(9):1734–1747. <https://doi.org/10.1109/TPAMI.2015.2496141>.
23. Caron M, Bojanowski P, Joulin A, Douze M. Deep clustering for unsupervised learning of visual features. *Proceedings of ECCV Conf*. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol; 11218. Springer Science+Business Media; 2018. https://openaccess.thecvf.com/content_ECCV_2018/papers/Mathilde_Caron_Deep_Clustering_for_ECCV_2018_paper.pdf.
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. IEEE; 2018. https://openaccess.thecvf.com/content_cvpr_2018/CameraReady/0801.pdf.
25. Van Den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. *arXiv*. 2018. <https://arxiv.org/pdf/1807.03748.pdf>.
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. IEEE; 2020. https://openaccess.thecvf.com/content_CVPR_2020/papers/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.pdf.
27. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. *arXiv*. 2020. <https://arxiv.org/pdf/2002.05709.pdf>.
28. Wang X, Liu Z, Yu SX. Unsupervised feature learning by cross-level discrimination between instances and groups. *arXiv*. 2020. <https://arxiv.org/pdf/2008.03813.pdf>.
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(11):1484–1498. <https://doi.org/10.1001/archophth.123.11.1484>.
30. Age-Related Eye Disease Study 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.
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(11):1570–1574. <https://doi.org/10.1001/archophth.123.11.1570>.
32. AREDS Research Group. The Age-Related Eye Disease Study (AREDS) system for classifying cataracts from photographs: AREDS report no. 4. *Am J Ophthalmol*. 2001;131(2):167–175. [https://doi.org/10.1016/s0002-9394\(00\)00732-7](https://doi.org/10.1016/s0002-9394(00)00732-7).
33. Ruderman DL. The statistics of natural images. *Netw Comput Neural Syst*. 1994;5:517–548. https://doi.org/10.1088/0954-898X_5_4_006.
34. Torralba A, Oliva A. Statistics of natural image categories. *Netw Comput Neural Syst*. 2003;14:391–412. https://doi.org/10.1088/0954-898X_14_3_302.
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. Vol2019 IEEE; October 2019. <https://arxiv.org/pdf/1905.01235.pdf>.
36. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. *The balanced accuracy and its posterior distribution*. In: Proceedings—International Conference on Pattern Recognition. IEEE; 2010:3121–3124.
37. Kelleher JD, Mac Namee B, D’Arcy A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. Cambridge, MA: The MIT Press; 2015.
38. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. In: Lucien Le Cam, Jerzy Neyman, eds. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; 1. University of California Press; 1967:281–297.
39. Lloyd SP. Least squares quantization in PCM. *IEEE Trans Inf Theory*. 1982;28(2):129–137. <https://doi.org/10.1109/TIT.1982.1056489>.
40. Dhillon IS, Modha DS. Concept decompositions for large sparse text data using clustering. *Mach Learn*. 2001;42:143–175. <https://doi.org/10.1023/A:1007612920971>.
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. <https://doi.org/10.1155/2018/2181602>.
42. Federico JR, Krishnamurthy K. *Albinism*. Treasure Island, FL: 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(2):141–150. <https://doi.org/10.1038/s42256-020-0154-9>.
44. Shen L, Lin Z, Huang Q. Relay backpropagation for effective learning of deep convolutional neural networks. *Computer Vision – ECCV 2016*. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol; 9911. Springer; 2016:467–482. <http://dblp.uni-trier.de/db/conf/eccv/eccv2016-7.html#ShenLH16>.
45. Zhang R, Isola P, Efros AA. Colorful image colorization. *Computer Vision - 14th European Conference, ECCV 2016*, Proceedings. In: Matas Jiri, Sebe Nicu, Welling Max, Leibe Bastian, eds. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol; 9907. Springer; 2016