

# Deep Ensemble Learning for Retinal Image Classification

Edward Ho<sup>1</sup>, Edward Wang<sup>1</sup>, Saerom Youn<sup>1</sup>, Asanth Sivajohan<sup>1</sup>, Kevin Lane<sup>1</sup>, Jin Chun<sup>1</sup>, and Cindy M. L. Hutnik<sup>1,2</sup>

<sup>1</sup> Schulich School of Medicine & Dentistry, University of Western Ontario, London, Ontario, Canada

<sup>2</sup> Departments of Ophthalmology and Pathology, University of Western Ontario, London, Ontario, Canada

**Correspondence:** Edward Ho, Schulich School of Medicine & Dentistry, 1151 Richmond St., London, ON N6A 5C1, Canada. e-mail: [eho87@uwo.ca](mailto:eho87@uwo.ca)

**Received:** February 1, 2022

**Accepted:** September 2, 2022

**Published:** October 28, 2022

**Keywords:** disease screening; funduscopy; deep learning; artificial intelligence (AI)

**Citation:** Ho E, Wang E, Youn S, Sivajohan A, Lane K, Chun J, Hutnik CML. Deep ensemble learning for retinal image classification. *Transl Vis Sci Technol.* 2022;11(10):39. <https://doi.org/10.1167/tvst.11.10.39>

**Purpose:** Vision impairment affects 2.2 billion people worldwide, half of which is preventable with early detection and treatment. Currently, automatic screening of ocular pathologies using convolutional neural networks (CNNs) on retinal fundus photographs is limited to a few pathologies. Simultaneous detection of multiple ophthalmic pathologies would increase clinical usability and uptake.

**Methods:** Two thousand five hundred sixty images were used from the Retinal Fundus Multi-Disease Image Dataset (RFMiD). Models were trained ( $n = 1920$ ) and validated ( $n = 640$ ). Five selected CNN architectures were trained to predict the presence of any pathology and categorize the 28 pathologies. All models were trained to minimize asymmetric loss, a modified form of binary cross-entropy. Individual model predictions were averaged to obtain a final ensemble model and assessed for mean area under the receiver-operator characteristic curve (AUROC) for disease screening (healthy versus pathologic image) and classification (AUROC for each class).

**Results:** The ensemble network achieved a disease screening (healthy versus pathologic) AUROC score of 0.9613. The highest single network score was 0.9586 using the SE-ResNeXt architecture. For individual disease classification, the average AUROC score for each class was 0.9295.

**Conclusions:** Retinal fundus images analyzed by an ensemble of CNNs trained to minimize asymmetric loss were effective in detection and classification of ocular pathologies than individual models. External validation is needed to translate machine learning models to diverse clinical contexts.

**Translational Relevance:** This study demonstrates the potential benefit of ensemble-based deep learning methods on improving automatic screening and diagnosis of multiple ocular pathologies from funduscopy imaging.

## Introduction

In 2019, the World Health Organization reported that 2.2 billion people worldwide have a visual impairment or blindness, half of which were either preventable or were not yet addressed.<sup>1</sup> The limited number of eye health professionals, especially in certain populations and geographic areas, is a barrier to more widespread in-person eye screening. One way to address this gap in coverage is through detection of eye pathologies using artificial intelligence (AI). This allows for efficient remote screening followed by prompt patient referral to the appropriate eye health professional and treatment if necessary.

Fundus images have been used for mass screening and detection of many eye pathologies because they are noninvasive and cost-effective.<sup>2</sup> Data-driven deep learning has developed rapidly and its application to funduscopy image analysis can broadly be grouped into classification, segmentation, and synthesis. More recently, the use of deep convolutional neural networks (CNNs) has been on the rise due to its ability to accurately classify images.<sup>2,3</sup>

The application of AI in classification began with targeting single ocular pathologies, like diabetic retinopathy.<sup>4,5</sup> The reality is that the most common risk factors for eye disease result in patients presenting with multiple simultaneous pathologies. As a result, investigations have evolved into detecting

multiple ophthalmic pathologies due to such classifications being more common and practical in real clinical settings. However, these studies remain limited due to challenges such as costly datasets, lack of labeling, severe class imbalance, and reduced image quality.

Presented herein is an introduction and description of a method to automatically perform disease risk prediction and to classify 28 different ophthalmic pathologies based on retinal fundus photography. An ensemble of convolutional neural networks trained with a modified classification loss function was used to overcome the barriers of class imbalance. The aim was to demonstrate how the addition of diverse classifiers and a replacement for training loss is a simple method to improve model performance which could potentially aid in automatic screening of ocular pathologies using retinal imaging.

## Methods

### Image Dataset

Retinal images were sourced from a Retinal Fundus Multi-Disease Image Dataset (RFMiD)<sup>6</sup>; please see the data description paper for full details. Retinal fundus images were acquired using one of the three digital fundus cameras (Kowa VX – 10 $\alpha$ , TOPCON 3D OCT-2000, and TOPCON TRC-NW300) from a trained retinal specialist. These images were obtained from patients visiting an eye clinic due to concerns about their eye health during the period of 2009 to 2020. Prior to funduscopy, all pupils were dilated with one drop of tropicamide at 0.5% concentration. A total of 3200 retinal images were available with 1920 (60%) available for training, 640 (20%) for an evaluation set, and 640 (20%) for the online test set.

Each image was annotated with the presence of 45 different ocular diseases or pathological findings by 2 independent ophthalmologists based on the image and the corresponding clinical records including visual fields. Discrepancies were resolved through consensus via a discussion with a third independent reviewer. Both high and low-quality images were included. For the entire dataset, only classes with more than 10 images were classified independently, and all others are merged into an “other” class. Additionally, each image is labeled as normal or abnormal depending on the presence or absence of any disease findings. This resulted in 29 classes for each image, 28 representing different pathological findings and 1 representing the presence of any abnormalities. Pachade et al.’s data description paper outlines the preselected clinical

findings used to label images to specific disease categories. The pathologies and their distributions within the training and evaluation sets are shown in Table 1.

### Preprocessing

For images acquired with TOPCON cameras, a square region in the center of the image was cropped as the field of view is roughly centered. For the Kowa images, the borders of the crops were found by finding the rectangular area containing all pixels above a constant threshold value.

### Image Classifiers

Because retinal images can have any number of pathological findings, this is a multiclass, multilabel classification problem. Therefore, an image classifier was used to model the relationship between the set of input images and a vector  $y_n$  representing the presence or absence of each of  $n$  findings. As such, should an image have no pathological findings, the value of  $y_n$  would be 0 for all  $n$ . For this task, CNNs were used, with specific utilization of a number of different current standard architectures, all of which have been used in the medical and retinal imaging literature. The architectures used were Inception V3,<sup>3</sup> SE-ResNeXt<sup>7</sup> (50-layers), DenseNet<sup>8</sup> (121-layers), and EfficientNet<sup>9</sup> (B4 and B5 scaled version).

### Loss Function

The problem of having both common and infrequent pathologies was addressed with asymmetric loss.<sup>10</sup> The binary cross-entropy loss was modified such that separate hyperparameters exist for positive and negative cases of each pathology. Specifically, if the binary cross-entropy loss is formulated as:

$$L = -yL_+ - (1 - y)L_-$$

where  $L_+$  and  $L_-$  represent the positive and negative loss parts and are defined as:

$$L_+ = \log(p)$$

$$L_- = \log(1 - p)$$

Then the asymmetric loss is a modification of the component loss parts such that:

$$L_+ = (1 - p)^{\gamma_+} \log(p)$$

$$L_- = (p_m)^{\gamma_-} \log(1 - p_m)$$

**Table 1.** Pathologies and Their Distributions Within the Training and Evaluation Set of the RFMiD Dataset

Pathology Label	Pathology Description	Count in Training Set	Count in Evaluation Set
DR	Diabetic retinopathy	376	132
MH	Media haze	317	102
ODC	Optic disc cupping	282	72
TSLN	Tessellation	186	65
DN	Drusen	138	46
MYA	Myopia	101	34
ARMD	Age-related macular degeneration	100	38
BRVO	Branch retinal vein occlusion	73	23
ODP	Optic disc pallor	65	26
ODE	Optic disc edema	58	21
LS	Laser scars	47	17
RS	Retinitis	43	14
CSR	Central serous retinopathy	37	11
Other	Other	34	21
CRS	Chorioretinitis	32	11
CRVO	Central retinal vein occlusion	28	8
RPEC	Retinal pigment epithelium changes	22	6
AION	Anterior ischemic optic neuropathy	17	5
AH	Asteroid hyalosis	16	4
MS	Macular scars	15	5
EDN	Exudation	15	5
ERM	Epiretinal membrane	14	7
RT	Retinal traction detachment	14	6
PT	Parafoveal telangiectasia	11	2
MHL	Macular hole	11	3
TV	Tortuous vessels	6	2
RP	Retinitis pigmentosa	6	2
ST	Optociliary shunt	5	4

where  $\gamma_+$  and  $\gamma_-$  are the positive and negative focusing parameters, respectively; and  $p_m$  is the shifted probability defined by:

$$p_m = \max(p - m, 0)$$

With the shifted probability in the negative loss part, this allows easy negative examples (such that the prediction probability is less than the probability margin  $m$ ) to not factor into the loss. By setting  $\gamma_- > \gamma_+$ , we can increase the contribution of positive samples to the loss. The focusing parameters and probability margins are tunable hyperparameters.

## Model Ensemble

Using multiple different classifiers and incorporating their results in an ensemble has been shown to improve the performance of image classifiers.<sup>11</sup> In this work, ensembling was performed in two ways:

first, the ensembling was performed across training folds. Five-fold cross-validation was used for model validation, leading to five independent models. Second, for evaluation on the test set, the output of each model was averaged. The above process was repeated for five different architectures, and the outputs of each five-fold ensemble were also averaged, resulting in our final ensemble.

## Model Training

As mentioned previously, 5 different architectures were used and trained with the hyperparameters and image input size listed on Table 2. For the SE-ResNeXt model, a dropout layer was added before the last linear layer. All other architectures were unmodified.

All networks were initialized by the default PyTorch pre-training on ImageNet, and trained using the Adam optimizer<sup>12</sup> with the asymmetric loss described above. Random horizontal flip and random rotation (between

**Table 2.** Hyperparameters and Image Input Sizes for the Five Different Architectures

Architecture	Image Size	Batch Size	Learning Rate	Training Epoch
SE-ResNeXt	512	8	$1 \times 10^4$	20
DenseNet-121	299	16	$2 \times 10^4$	10
Inception V3	299	32	$2 \times 10^4$	10
EfficientNet-B4	380	12	$2 \times 10^4$	10
EfficientNet-B5	456	8	$2 \times 10^4$	10

–30 and 30 degrees) augmentations were used for all networks except Inception V3, which rotated between –90 and 90 degrees and incorporated a random brightness transformation.

Each model was trained and evaluated using five-fold cross-validation. In order to best balance the classes represented in each fold, iterative stratification was performed as previously described.<sup>13,14</sup> Final results were obtained by using the ensemble of networks as described above and performing predictions on a hidden test set.

### Model Metrics

Model performance was assessed using the area under the receiver operator characteristic (AUROC), which was calculated using the probability outputs for each class from the final layer of our neural networks. To assess the model’s performance in disease screening, the AUROC score was calculated based on the prediction of healthy versus pathologic image. For assessment of disease classification performance, the AUROC was calculated based on the model prediction for each pathology class. The average AUROC for all classes was also calculated. Additionally, the sensitivity and specificity for disease screening and for each individual pathology was also measured using a threshold probability of greater than 0.5.

### Prediction Visualization

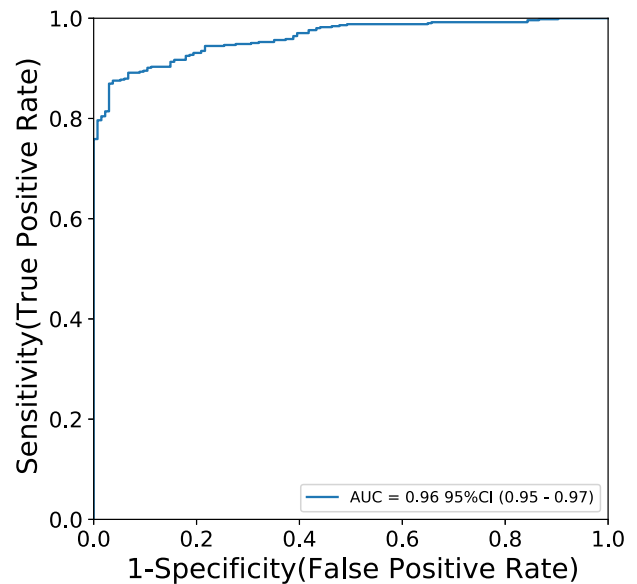
The Grad-CAM<sup>15</sup> saliency mapping tool was used to visualize the pixel areas contributory to the classification of images into specific classes.

## Results

The receiver operator characteristic (ROC) curve for the disease screening (healthy versus pathologic image) task is shown in Figure 1. The AUROC score achieved by each individual network of the ensemble and the final ensemble are shown in Table 3. The AUROC score

for the full ensemble was found to be 0.9613, with the highest individual network AUROC score being 0.9587 achieved by the SE-ResNeXt architecture. This difference is not statistically significant when using the DeLong test<sup>16</sup> for AUROC comparison ( $P = 0.3380$ ).

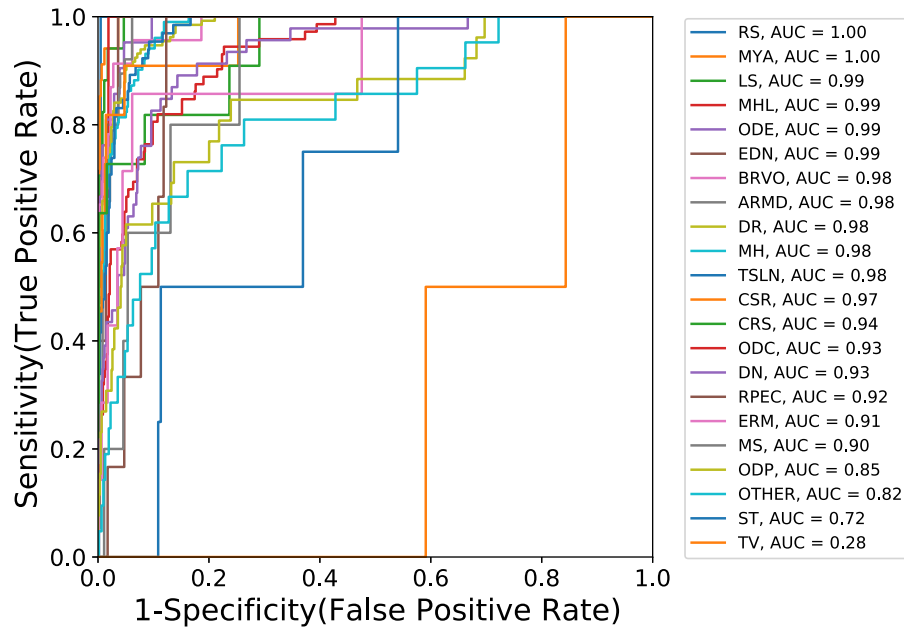
The ROC curves for disease classification of each pathology are shown in Figure 2. The average AUROC score was 0.9295, with values ranging from 0.28 to 1.00.



**Figure 1.** Receiver operator characteristic (ROC) curve for disease screening task performed by the network ensemble.

**Table 3.** AUROC Scores for Disease Screening (Healthy Versus Pathologic Image) and Average Score for Each Disease Classification

Architecture	Disease Screening AUROC	Disease Classification Average AUROC
Inception V3	0.9569	0.9091
SE-ResNeXt	0.9587	0.9066
DenseNet-121	0.9519	0.9298
EfficientNet-B4	0.9477	0.9030
EfficientNet-B5	0.9540	0.9163
Ensemble	0.9613	0.9295



**Figure 2.** Receiver operator characteristic (ROC) curves for each pathology class classified by the network ensemble.

Apart from identification of tortuous vessels (AUROC = 0.28) and retinal shunts (AUROC = 0.82), all classes had AUROC greater than 0.85. When compared to the ensemble network, the DenseNet architecture achieved a comparable average AUROC score of 0.9298.

The sensitivity and specificity results for disease screening (disease risk) as well as each individual pathology is presented in Table 4. The model was able to achieve a sensitivity of 0.9705 and a specificity of 0.5896 on disease screening. Performance for individual pathology varied greatly, with sensitivity values ranging from 0 to 0.9704, and specificity ranging from 0.5896 to 1.0. Confusion matrices for disease screening and individual pathologies are presented in Supplementary Table S1.

### Results of Activation Mapping

Grad-CAM activation maps are shown in Table 5. Visualization of activation maps demonstrates that predictions for the different pathologies does correspond with expected fundoscopic features. For example, pathologies involving the optic disc do focus on the optic disc itself. Similarly, pathologies which are associated with specific entities (for example: drusen with lipid deposits and laser scars) tend to focus on these findings when present. Pathologies that involve more generalized changes, such as diabetic retinopathy, and tessellations have activation maps that focus on several different areas. A survey of false positive examples demonstrates that the model is still limited

in the amount of detail it can gain from these foci, as many false positives look in the right place and yet arrive at the wrong classification.

## Discussion

Screening for multiple ocular diseases in retinal images using an ensemble neural network approach was shown to be feasible in this study. The proposed approach achieved an AUROC score of 0.9613 on the RFMiD dataset for detection of any ocular pathology, which was found to be higher than the single network best score of 0.9587, although the difference is not statistically significant. For individual classification of pathologies, the average AUROC for each class was 0.9298. Whereas most pathology classification had AUROC > 0.85, as illustrated by Figure 2, it is apparent from the sensitivity and specificity values that performance is still impacted by the small number of training and validation images for some classes, as is the case for tortuous vessels, shunts, and other pathologies where performance is at either extreme.

Classifying multiple ocular pathologies using retinal imaging has been studied before, although they were limited due to lack of labeled data. Li et al.<sup>17</sup> developed a deep learning approach that classified 12 different pathologies using a SE-ResNeXt network comparable to the one used in the present ensemble approach. Although their performance and testing dataset size is greater than the current study, they did not cover the breadth of pathologies studied in

**Table 4.** Sensitivity and Specificity Values for Predictions on the Validation Set for Each Class

Pathology Description	Sensitivity	Specificity
Disease risk	0.9704	0.5896
Diabetic retinopathy	0.8712	0.9488
Media haze	0.9020	0.9275
Optic disc cupping	0.7778	0.9278
Tessellation	0.7846	0.9704
Drusen	0.5435	0.9529
Myopia	0.9412	0.9835
Age-related macular degeneration	0.6842	0.9801
Branch retinal vein occlusion	0.6087	0.9984
Optic disc pallor	0.2692	0.9902
Optic disc edema	0.7619	0.9871
Laser scars	0.7059	0.9984
Retinitis	1.000	0.9936
Central serous retinopathy	0.4545	0.9936
Other	0.0476	0.9984
Chorioretinitis	0.2727	0.9984
Central retinal vein occlusion	0.8750	0.9968
Retinal pigment epithelium changes	0.000	0.9968
Anterior ischemic optic neuropathy	0.4000	1.000
Asteroid hyalosis	0.500	1.000
Macular scars	0.000	1.000
Exudation	0.4000	0.9843
Epiretinal membrane	0.000	1.000
Retinal traction detachment	0.8333	0.9984
Parafoveal telangiectasia	0.000	1.000
Macular hole	0.000	1.000
Tortuous vessels	0.000	0.9984
Retinitis pigmentosa	0.000	1.000
Optociliary shunt	0.000	1.000

Values are measured using a threshold of greater than 0.5 for the probability output of the neural network.

the present work. Quellec et al.<sup>18</sup> utilized a few-shot learning approach to classify rare pathologies with an average AUROC of 0.938. Despite the larger dataset (164,660 examinations) and different demographics (collected from the OPHDIAT<sup>19</sup> screening network in France), they struggled with similar pathologies, shunts (AUROC = 0.7586), and emboli (AUROC = 0.7946). Ting et al.<sup>19</sup> studied diabetic retinopathy and related diseases in retinal images from multi-ethnic populations with diabetes, and found comparable performance in detecting diabetic retinopathy and age-related macular degeneration. However, their study focused mainly on different severities of diabetic retinopathy and macular degeneration as opposed to a wide variety of pathological signs. Last, in a recent paper, Cen et al.<sup>20</sup> developed a deep learning platform that was able to classify 39 different conditions with an AUROC of 0.9984 from 249,620 images. Additionally, the model

was evaluated on an external dataset, and achieved similar results (AUROC = 0.9990). Although many of the pathologies studied by Cen et al. and the present authors overlapped, Cen et al. did not investigate media haze, anterior ischemic optic neuropathy, parafoveal telangiectasia, or optociliary shunt. Optociliary shunt performed poorly in this present work, with an AUROC of 0.72. This performance is likely explained by the poor availability of this pathological finding in the current dataset, and is one of the pathologies where the trained models failed to identify any positive samples in the evaluation set (sensitivity = 0).

A key strength of this study is the large breadth of pathologies labeled in the current dataset. Although many algorithms have been developed to detect single, common diseases, such as diabetic retinopathy and macular degeneration, few studies attempt to classify many different pathologies at once, mainly due to lack

**Table 5.** Activation Map Visualizations for Representative True Positive and False Positive Predictions for Different Pathology Classifications

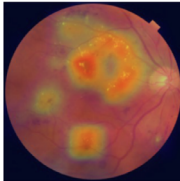
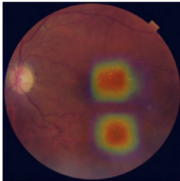
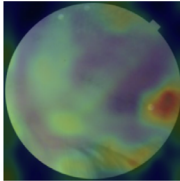
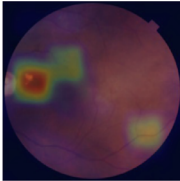
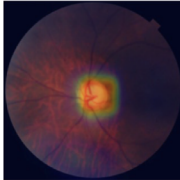
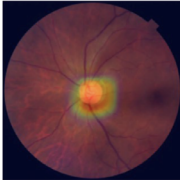
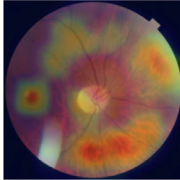
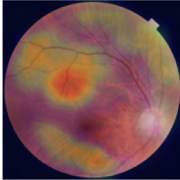
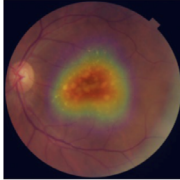
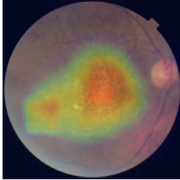
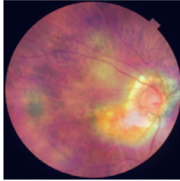
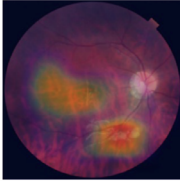
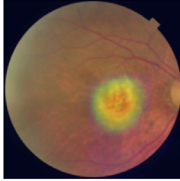
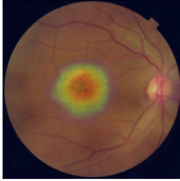
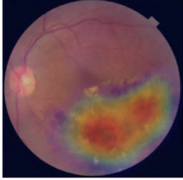
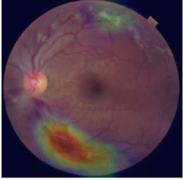
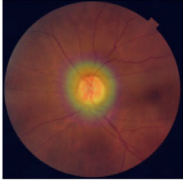
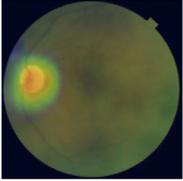
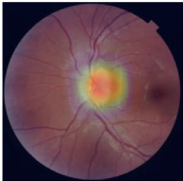
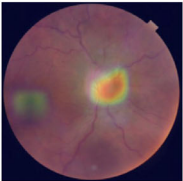
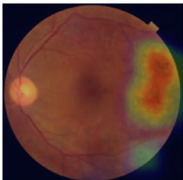
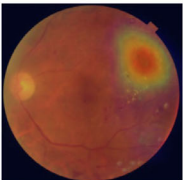
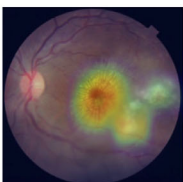
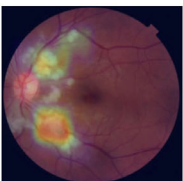
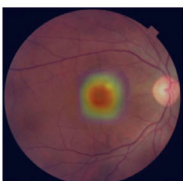
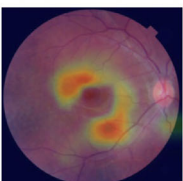
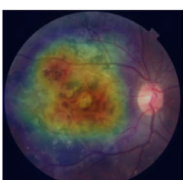
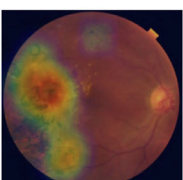
Pathology Description	True Positive Example	False Positive Example
Diabetic retinopathy		
Media haze		
Optic disc cupping		
Tessellation		
Drusen		
Myopia		
Age-related macular degeneration		

Table 5. Continued

Branch retinal vein occlusion		
Optic disc pallor		
Optic disc edema		
Laser scars		
Retinitis		
Central serous retinopathy		
Chorioretinitis		

Only classes with double digit number of samples in the validation are included, and the “other” class is also excluded.



of labeled data. Whereas detecting single pathologies is a meaningful academic exercise, the real life value of an automated screening tool necessitates simultaneous assessments of multiple pathologies to overcome barriers to implementation. The simplicity of our approach allows applications in low resource settings.

Central to the performance of our approach is the diversity of neural network architectures used, each of which address challenges in deep learning with different approaches. Inception V3<sup>3</sup> is a CNN architecture that uses “Inception” style building blocks, which consists of concatenating image filters of different sizes at the same level in order to approximate an optimal sparse network structure. This addressed several technical issues with CNNs at the time, and also followed the intuition that image features should be analyzed at different scales simultaneously. The architecture has been used in various computer vision tasks in retinal imaging, including detection of diabetic retinopathy<sup>4,21</sup> and multi-disease detection.<sup>18</sup>

SE-ResNeXt<sup>7</sup> is an architecture that uses residual learning,<sup>22</sup> which implements shortcut connections between network layers to mitigate the difficulties of training very deep neural networks. Additionally, this architecture uses squeeze-excitation blocks, which explicitly models interdependencies between image channels to improve performance on computer vision tasks. In this study, the 50 layer version (SE-ResNeXt-50) was used.

The DenseNet<sup>8</sup> architecture similarly aims to counteract the problem of training very deep neural networks; however, instead of the shortcut connections used in residual learning, it connects each layer with each other network, allowing for fewer parameters and utilization of image features from all complexity levels for classification. In this work, the 121 layer version of the architecture (DenseNet-121) is used.

EfficientNet<sup>9</sup> is an architecture that was developed to address the challenge of scaling up CNNs efficiently with respect to the model performance and the number of parameters. The base version of the architecture (EfficientNet-B0) was developed using neural architecture search methods that optimized for accuracy and floating point operations per second (FLOPS). From this baseline, optimal values of the network depth (the number of layers), width (the number of channels), and resolution (the size of the input image) were found such that each larger version of the network would take the power of the baseline values. The proposed approach uses the B4 and B5 scaled versions of the architecture.

Additionally, the class imbalance of the present study was addressed successfully with asymmetric loss.<sup>10</sup> For multiclass multilabel classification, typically a binary cross-entropy loss is used to train neural

network classifiers. However, in problems with significant class imbalance, the loss may be skewed by the contributions from more represented classes, which are easier for the network to learn. In order to focus more on ambiguous samples, the loss functions can be weighted based on the confidence of the network output. Additionally, as there is a high representation of negative samples for many pathologies, the contribution of negative samples to the loss was reduced with the asymmetric loss, even when the output probability is high.

In this work, an ensemble of different architectures was explored given its known improvements to model performance, as described in previous literature.<sup>11</sup> In our experiments, there were observable numerical improvements to AUROC scores for screening and classification, however, they did not reach statistical significance compared to trials without ensembling. These differences are not significant. Nonetheless, ensembling did produce a network that was able to combine the benefits from models that performed better at disease screening (in our example, the SE-ResNeXt architecture) with those that performed better at individual disease classification (DenseNet). Despite this, the lack of statistical significance calls into question the widely known and accepted benefit of ensembling. Further studies can be conducted to determine whether these benefits persist in neural networks built for ophthalmologic image analysis. Furthermore, testing the efficacy of ensembling for different clinical tasks can better characterize its potential for gaining stronger performance on simpler models capable of running on limited computational resources through knowledge distillation<sup>23</sup> and similar techniques. In recent years, mobile screening tools in the form of smartphones<sup>24</sup> and other edge computing solutions<sup>25</sup> have demonstrated that neural networks and other computational tools can be deployed even without extensive hardware to benefit vision care. Detection of multiple ocular pathologies on these platforms have yet to be tried.

One of the major limitations of this study is the relatively small dataset. The RFMiD data has only 3200 images, 640 of which were not labeled. This may decrease generalizability. Additionally, whereas a large range of pathologies were studied, many pathologies were very limited in sample size. Although the ROC curves and high average AUROC for disease classification suggest that our approach was able to learn features tied to these pathologies, the sensitivity and specificity results indicate that sample size remains a limitation to achieving performance suitable for clinical implementation. Despite this, our study included pathologies with smaller sample sizes to test the limits

of multi-class detection models that can be built with datasets with a wide variety of pathologies.

Another limitation was lack of access to the baseline patient and clinical characteristics of the sample images made available to us from the RFMiD dataset. Our approach may lack external validity if the ocular characteristics of external data differ significantly from the normative data used to train the CNNs. Because it is the only publicly available annotated dataset of its kind,<sup>6</sup> future studies aimed at external validation of our model may be reassuring. What is presented here is a valuable proof of concept that a deep learning system trained to classify multiple ocular pathologies with acceptable performance for screening is possible. With more training data, the performance and generalizability of a neural network ensemble will improve. Additionally, a significant limitation is the lack of pixel-level data. Whereas the Grad-CAM analysis gives an overall impression of what visual features are aiding in its decision making, precise medical rationale cannot be elucidated without training with images that have been labeled at the pixel level to clearly identify the pathological features associated with a classification.

Future studies should be performed to evaluate the internal and external validity of the ensemble network approach for detection of multiple ocular pathologies using fundus photography. In particular, it is important to ensure that any disease screening application is not biased by limited or unrepresentative training data. This should be investigated by studying the generalizability of networks trained on specific population data. Because a major limitation of this study is access to annotated data, other avenues of few-shot learning, such as feature reduction techniques<sup>18</sup> or generative adversarial networks,<sup>25</sup> to create more training samples may be explored.

Future versions of this algorithm can enhance its accuracy by incorporating relevant clinical information. These algorithms can combine the decision score generated from fundoscopic analysis along with the clinical variable to arrive at a more accurate prediction. For instance, the presence of symptoms, such as eye pain, headaches, and nausea, can be strong diagnostic indicators for acute angle-closure glaucoma.<sup>26</sup> Demographic features, such as age, gender, and ethnicity, are also useful aspects to consider as ocular diseases are more present in certain populations than others.<sup>27</sup>

There are several barriers to real world implementation. As described above, the limited sensitivity and specificity metrics limit the efficacy of such a model as a screening tool. Additionally, a product offering this type of automated feature should be designed for continuous development and monitoring of the model's performance. Thus, both labeling and diagnostic capacity should be looped to

allow for periodic auditing and tracking of changes to performance metrics. The product should then be piloted in a controlled setting to observe and analyze effects on clinical outcomes and practice management (e.g. number of patients screened, types of pathologies detected, stage of disease when detected, vision outcomes, and cost savings) that outweigh the monetary and non-monetary costs associated with the new technology. Once clinical outcomes and costs are better characterized, a discussion among stakeholders regarding risk assessment for data privacy, algorithm failure, auditing, and medico legal responsibility can take place for safe, controlled implementation.

To conclude, we demonstrated that an ensemble neural network approach with a modified classification loss can be used to perform automated screening for 28 different ocular pathologies using fundus photography. Further work needs to be performed to evaluate the performance of these algorithms in clinical practice with real world data from diverse populations.

## Acknowledgments

Disclosure: **E. Ho**, None; **E. Wang**, None; **S. Youn**, None; **A. Sivajohan**, None; **K. Lane**, None; **J. Chun**, None; **C.M.L. Hutnik**, None

## References

1. World Health Organization. World report on vision. Published online 2019. Available at: <https://www.who.int/publications/i/item/9789241516570> port on vision (who.int).
2. Edupuganti VG, Chawla A, Kale A. Automatic Optic Disk and Cup Segmentation of Fundus Images Using Deep Learning. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. 2018:2227–2231. Available at: <https://ieeexplore.ieee.org/document/8451753> IEEE Conference Publication | IEEE Xplore.
3. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2016:2818–2826. Available at: <https://ieeexplore.ieee.org/document/7780677> | IEEE Xplore.
4. 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.

5. 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(13):5200–5206.
6. Pachade S, Porwal P, Thulkar D, et al. Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for Multi-Disease Detection Research. *Data*. 2021;6(2):14.
7. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. *arXiv:170901507 [cs]*. Published online May 16, 2019. Accessed March 3, 2021, <http://arxiv.org/abs/1709.01507>.
8. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2017:2261–2269.
9. Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: *International Conference on Machine Learning*. PMLR; 2019:6105–6114. Accessed March 3, 2021, <http://proceedings.mlr.press/v97/tan19a.html>.
10. Ben-Baruch E, Ridnik T, Zamir N, et al. Asymmetric Loss For Multi-Label Classification. *arXiv:200914119 [cs]*. Published online November 17, 2020. Accessed March 3, 2021, <http://arxiv.org/abs/2009.14119>.
11. Ju C, Bibaut A, van der Laan M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*. 2018;45(15):2800–2818.
12. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv:14126980 [cs]*. Published online January 29, 2017. Accessed March 3, 2021, <http://arxiv.org/abs/1412.6980>.
13. Sechidis K, Tsoumakas G, Vlahavas I. On the Stratification of Multi-label Data. In: Gunopulos D, Hofmann T, Malerba D, Vazirgiannis M, eds. *Machine Learning and Knowledge Discovery in Databases*. Vol. 6913. Lecture Notes in Computer Science. Berlin Heidelberg: Springer; 2011:145–158.
14. Szymański P, Kajdanowicz T. A Network Perspective on Stratification of Multi-Label Data. In: *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*. PMLR; 2017:22–35. Accessed March 3, 2021, <http://proceedings.mlr.press/v74/szysma%20C5%84ski17a.html>.
15. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 2017:617–626. Available at: <https://ieeexplore.ieee.org/document/8237336> M: Visual Explanations from Deep Networks via Gradient-Based Localization | IEEE Conference Publication | IEEE Xplore.
16. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845.
17. Li B, Chen H, Zhang B, et al. Development and evaluation of a deep learning model for the detection of multiple fundus diseases based on colour fundus photography. *Br J Ophthalmol*. 2022;106(8):1079–1086.
18. Quellec G, Lamard M, Conze PH, Massin P, Cochener B. Automatic detection of rare pathologies in fundus photographs using few-shot learning. *Medical Image Analysis*. 2020;61:101660.
19. 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.
20. Cen LP, Ji J, Lin JW, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat Commun*. 2021;12(1):4828, doi:10.1038/s41467-021-25138-w.
21. Gao Z, Li J, Guo J, Chen Y, Yi Z, Zhong J. Diagnosis of Diabetic Retinopathy Using Deep Neural Networks. *IEEE Access*. 2019;7:3360–3370.
22. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *arXiv:151203385 [cs]*. Published online December 10, 2015. Accessed June 22, 2020, <http://arxiv.org/abs/1512.03385>.
23. Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. *arXiv:150302531 [cs, stat]*. Published online March 9, 2015. Accessed August 29, 2021, <http://arxiv.org/abs/1503.02531>.
24. Kim TN, Myers F, Reber C, et al. A Smartphone-Based Tool for Rapid, Portable, and Automated Wide-Field Retinal Imaging. *Transl Vis Sci Technol*. 2018;7(5):21.
25. Yoo TK, Choi JY, Kim HK. Feasibility study to improve deep learning in OCT diagnosis of rare retinal diseases with few-shot classification. *Med Biol Eng Comput*. 2021;59(2):401–415.
26. Dietze J, Blair K, Havens SJ. Glaucoma. In: *StatPearls*. Treasure Island, FL: StatPearls Publishing; 2021. Accessed January 10, 2022, <http://www.ncbi.nlm.nih.gov/books/NBK538217/>.
27. Bourne RRA. Ethnicity and ocular imaging. *Eye (Lond)*. 2011;25(3):297–300.